CHAPTER 1

A primer on Markov chain Monte Carlo

Peter J. Green, University of Bristol

1.1 Introduction

Markov chain Monte Carlo is probably about 50 years old, and has been both developed and extensively used in physics for the last four decades. However, the most spectacular increase in its impact and influence in statistics and probability has come since the late '80's.

It has now come to be an all-pervading technique in statistical computation, in particular for Bayesian inference, and especially in complex stochastic systems. A huge research effort is being expended, in devising new generic techniques, in extending the application of existing techniques, and in investigating the mathematical properties of the methods.

The target audience for the *Sémstat* lectures is European post-doctoral researchers in probability and statistics, and the present chapter is both the written version of these lectures and a primer for others seeking to get started in some aspect of MCMC research. By 'MCMC research' I mean *both* research into the mathematical properties of MCMC algorithms, and research that aims to develop new classes of algorithm for new and challenging problems; in both cases, I am thinking primarily but not quite exclusively of ultimate application in Bayesian statistics. Thus the chapter is not primarily intended for those who wish to make use of standard MCMC methods as implemented in a package, and to make sense of the output; however, it should be of some use to those wishing to apply standard methods to some new application by means of their own code. The focus is on understanding the principles underlying the methods, and the main ideas in evaluating their performance. With that objective, I will begin with some very basic examples, covered in detail, which are aimed at those who are complete novices. Those with a basic understanding of

Bayesian analysis and the Gibbs sampler may not need this motivation, and can skip Section 1.2.

The selection of material is necessarily a personal one — the subject is by now too big for the 4 or 5 hours allocated to the lectures, and indeed I would not claim expertise over all of the potential coverage of a lecture series of this kind. To save space, some sections have been reduced to just a few key references.

I have decided not to try to cover any very substantial applications, although plenty of reference is made to such work. I do make use of a running example — on point processes with change points, exemplified by a Bayesian analysis of some data on cyclones — that is intended to provide continuity as I cover the main topics.

1.2 Getting started: Bayesian inference and the Gibbs sampler

1.2.1 Bayes theorem and inference

The recent great impetus to research in MCMC has been the widespread realisation of its important application in Bayesian inference, following the work of Besag and York (1989) and Gelfand and Smith (1990), building on the 'Gibbs sampler' (popularly ascribed to Geman and Geman (1984)). The book of Gilks, Richardson and Spiegelhalter (1996), comprising articles contributed by 32 authors, provides an excellent introduction and overview to the theory, implementation and application of Bayesian MCMC.

Let us start with the simplest basic set-up, a model relating data Y and parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$. We need two probabilistic models: a data model specifying the likelihood: $p(Y|\boldsymbol{\theta})$, and a prior model, specifying the prior distribution $p(\boldsymbol{\theta})$.

In the Bayesian approach, inference is based on the *joint posterior*

$$\begin{aligned} p(\boldsymbol{\theta}|Y) &= \frac{p(\boldsymbol{\theta})p(Y|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(Y|\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto p(\boldsymbol{\theta})p(Y|\boldsymbol{\theta}) \\ \text{i.e.} \qquad \textit{Posterior} &\propto \textit{Prior} \times \textit{Likelihood} \end{aligned}$$

For a proper account of Bayesian theory, the reader is referred to Bernardo and Smith (1994) or O'Hagan (1994).

1.2.2 Cyclones example: point processes and change points

We are going to illustrate the ideas of MCMC with a running example: the observations are a point process of *events* at times $y_1, y_2, \ldots, y_N$ in an observation interval $[0, L)$. For simplicity, we suppose the events occur *at random* — that is, as a Poisson process — but at a possibly non-uniform rate: say rate $x(t)$ per unit time, at time t. The objective is to make inference

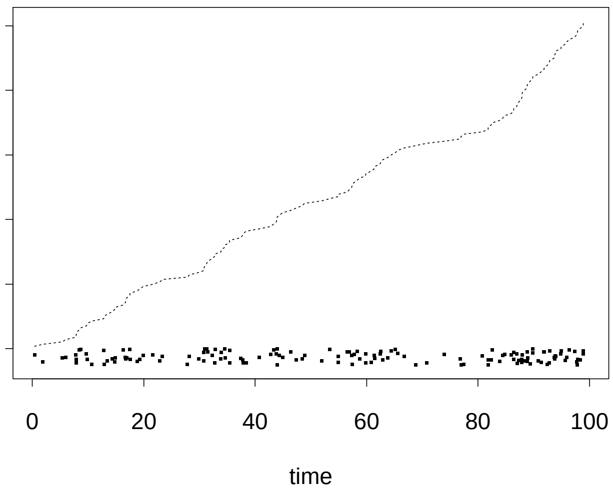

Figure 1.1 *Cyclones data, as a jittered dot plot, and their cumulative counting process.*

about $x(t)$. We will work up through a series of models, ultimately allowing an unknown number of change points, unknown hyperparameters, and a parametric periodic component.

The models and the respective algorithms and inferences will be illustrated by an analysis of a data set of the times of cyclones hitting the Bay of Bengal; there were 141 cyclones over a period of 101 years (Mooley, 1981). The data are plotted, both as a jittered dot plot, and by means of their cumulative counting process, in Figure 1.1.

Model 1: constant rate

First suppose that $x(t) \equiv x$ for all t.

Then the times of the events are immaterial: we observe N events in a time interval of length L; the obvious estimate of x is

$$\widehat{x} = \frac{N}{L}.$$

This is the *maximum likelihood estimator* of x under the assumption (implied by the 'randomness' assumption above), that N has a Poisson distribution, with mean xL:

$$p(N|x) = e^{-xL}\frac{(xL)^N}{N!}.$$

Model 2: constant rate, the Bayesian way

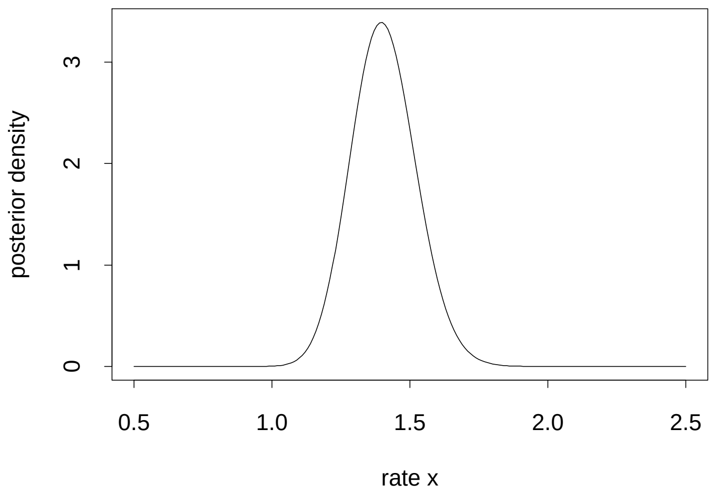

Figure 1.2 *Cyclones data: posterior for x in model 2.*

To take a Bayesian approach to this example, suppose that we have prior information about x (from previous studies, for example). Let us suppose we can model this by saying

$$x \sim \Gamma(\alpha, \beta),$$

a Gamma distribution (with mean α/β and variance α/β^2).

Then since

$$p(x|N) \propto p(x)p(N|x),$$

we find that

$$\begin{aligned} p(x|N) &\propto \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} e^{-xL}\frac{(xL)^N}{N!} \\ &\propto x^{\alpha+N-1}\exp(-(\beta+L)x) \end{aligned}$$

or in other words

$$x|N \sim \Gamma(\alpha+N, \beta+L).$$

So x has a Gamma distribution with mean $(\alpha+N)/(\beta+L)$, or approximately N/L if N and L are large compared with α and β. Thus *with a lot of data*, the Bayesian posterior mean is close to the maximum likelihood estimator. The posterior distribution of x for model 2 fitted to the cyclones data is shown in Figure 1.2; we used $\alpha=\beta=1$ here.

There is no need for MCMC in this model: you can calculate the posterior exactly, and recognise it as a standard distribution. It would not have worked out like this for any other prior; this choice is called *conjugate*.

1.2.3 The Gibbs sampler for a Normal random sample

Before we elaborate the cyclones example to a point where exact calculation is no longer practicable, let alone formally introduce Markov chain Monte Carlo methods, let us consider an even simpler, and completely familiar, example, but following an elementary Bayesian approach.

Our data are a random sample of size n from $N(\mu,\sigma^2)$. We place independent priors on μ and σ:

$$\begin{aligned} \mu &\sim N(\xi,\kappa^{-1}) \\ \sigma^{-2} &\sim \Gamma(\alpha,\beta), \end{aligned}$$

and it is easy to see that the resulting joint posterior is

$$\begin{aligned} p(\mu,\sigma^{-2}|Y) &\propto (\sigma^{-2})^{\alpha+n/2-1} \\ &\quad \times \exp\left\{-\frac{\beta}{\sigma^2}-\frac{\kappa(\mu-\xi)^2}{2}-\frac{\sum(Y_i-\mu)^2}{2\sigma^2}\right\}. \end{aligned} \tag{1.1}$$

This is somewhat awkward to handle; the parameters are dependent *a posteriori*, although they were independent *a priori*. However, the *full conditionals* — the conditional distributions of each parameter given the other parameter(s) and the data — are easily found:

$$\begin{aligned} \mu|\sigma,Y &\sim N\left(\frac{\sigma^{-2}\sum Y_i+\kappa\xi}{\sigma^{-2}n+\kappa},\frac{1}{\sigma^{-2}n+\kappa}\right) \\ \sigma^{-2}|\mu,Y &\sim \Gamma(\alpha+n/2,\beta+\textstyle\sum(Y_i-\mu)^2/2). \end{aligned}$$

What happens if we generate a sample of (μ,σ) pairs by alternately drawing μ and σ^{-2} from these distributions? The beginning of this process is illustrated in Figure 1.3, using the (improper) uninformative prior setting $\xi=\kappa=\alpha=\beta=0$.

This is a simple example of a *Gibbs sampler*. The alternating updates of one variable conditioned on the other induces Markov dependence: the successively sampled pairs form a Markov chain (on the uncountable state space $\mathcal{R}\times\mathcal{R}^+$), and it is readily shown that the joint posterior (1.1) is the (unique) invariant distribution of the chain. Standard theorems, quoted in Section 1.3.1 below, imply that the chain converges to this invariant distri-

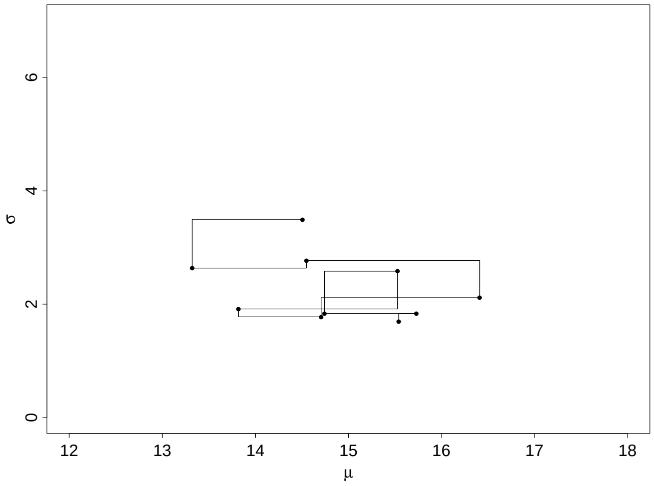

Figure 1.3 *First 10 samples from a Gibbs sampler of* (μ, σ) *from Normal random sample with* $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. *Uninformative prior.*

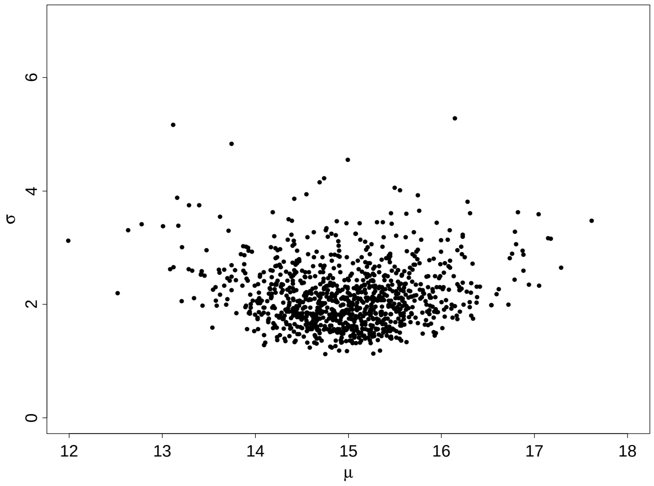

Figure 1.4 *Posterior sample of* (μ, σ) *from Normal random sample with* $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. *Uninformative prior.*

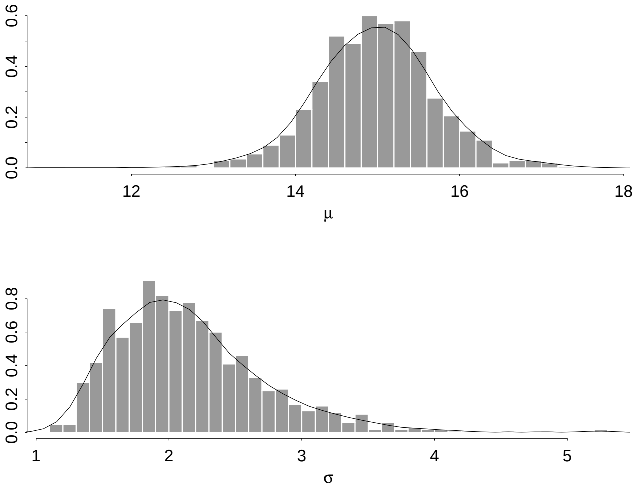

Figure 1.5 *Posterior distributions of* μ *and* σ *from Normal random sample with* $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. *Uninformative prior.*

bution in several useful senses, so that we can treat the realised values as a sample from the posterior. A sample of 1000 pairs is shown in Figure 1.4, and the shape of the joint distribution can now be discerned. Examples of possible outputs of interest are the marginal distributions shown in Figure 1.5.

However, we need not be confined to pictorial displays of marginal posteriors. One of the great liberating influences of MCMC in Bayesian inference has been the flexibility of inference afforded by sample-based computation. For example, consider prediction: we can calculate $P\{Y_{n+1} > 19\}$ by averaging $1 - \Phi(\{19 - \mu\}/\sigma)$:

$$\frac{1}{N}\sum_{t=1}^{N}\left[1 - \Phi(\{19 - \mu^{(t)}\}/\sigma^{(t)})\right] \approx 0.045$$

for the sample of Figure 1.4. Incidentally, it is interesting that this is more than twice the value (0.0175) that a frequentist would obtain by plugging the maximum likelihood estimates into $1 - \Phi(\{19 - \mu\}/\sigma)$. (Of course, this, like any other inference based on this model, is influenced by the prior setting used.)

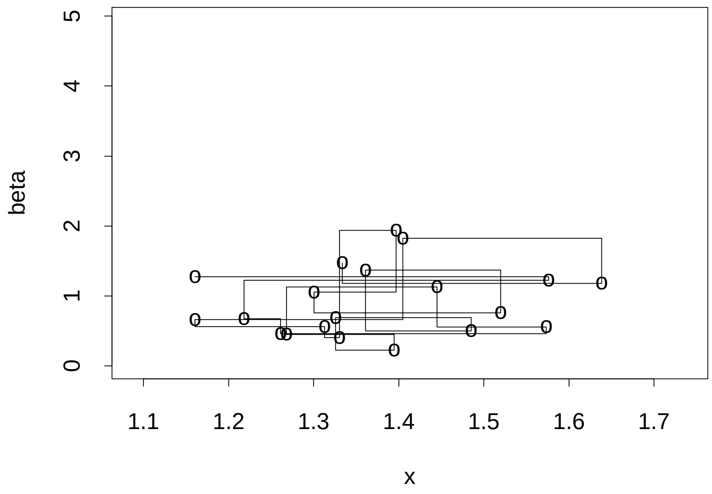

Figure 1.6 *First few moves of the Gibbs sampler for the cyclones data, model 3.*

1.2.4 Cyclones example, continued

For a more interesting and substantial application, let us return to the cyclones example, and consider some elaborations of the basic model 2.

Model 3: constant rate, with hyperparameter

Suppose you are reluctant to specify your prior fully: you are happy to say

$$x \sim \Gamma(\alpha, \beta)$$

and can specify α but not β, and want to state only

$$\beta \sim \Gamma(e, f)$$

for fixed e and f. (This formulation actually makes rather more sense in our next formulation, model 4).

Now $p(x|N, \alpha, e, f)$ is no longer available: it does not have an explicit form. But $p(x|N, \alpha, \beta, e, f)$ and $p(\beta|x, N, \alpha, e, f)$ are simple:

$$x|N, \alpha, \beta, e, f \sim \Gamma(\alpha + N, \beta + L)$$

as before, and

$$\beta|x, N, \alpha, e, f \sim \Gamma(e + \alpha, f + x).$$

So we can use the Gibbs sampler, and sample from these distributions in

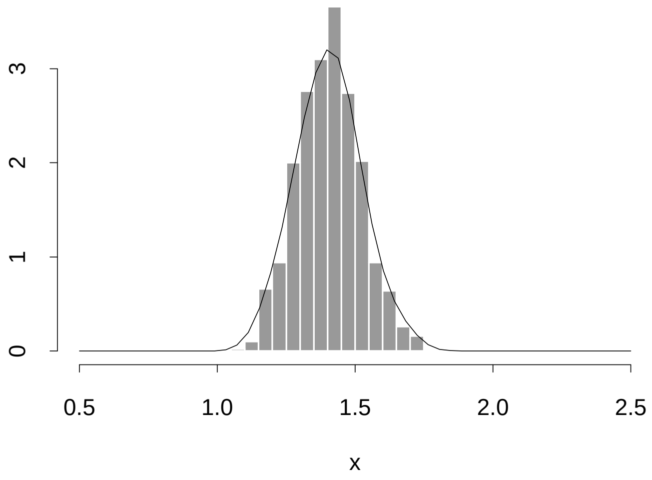

Figure 1.7 *Marginal distribution for x for the cyclones data, model 3.*

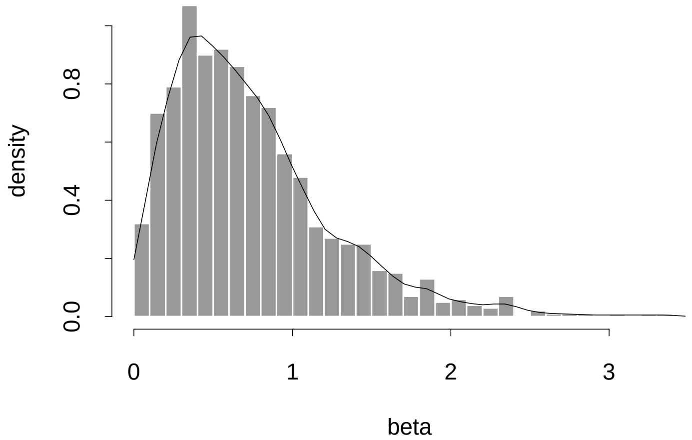

Figure 1.8 *Marginal distribution for β for the cyclones data, model 3.*

turn, updating x and β alternately. This creates a Markov chain with states (x, β), the unknown parameters in this model.

Figure 1.6 shows the first few moves of a Gibbs sampler applied to model 3 on the cyclones data; we took $e = 1$ and $f = N/L = 1.396$, and kept $\alpha = 1$. The marginal distributions for x and β, as accumulated from the first 1000 sweeps of this Gibbs sampler are displayed in Figures 1.7 and 1.8.

Model 4: constant rate, with change point

Now let us allow $x(t)$ to vary, but in a particular way.

Suppose $x(t)$ is piecewise constant, that is, a step function. This might be a suitable model if we postulate one or more *change points*; the process is completely random, but the rate switches between levels, perhaps as part of an underlying process, perhaps due to the recording mechanism.

Let us first take one change point, at *known* time $T \in (0, L)$, so that

$$x(t) = \begin{cases} x_0 & \text{if } 0 \le t < T \\ x_1 & \text{if } T \le t < L \end{cases} .$$

Suppose that x_0 and x_1 are *a priori* independently drawn from Gamma distributions, as before:

$$x_j \sim \Gamma(\alpha, \beta).$$

Then if N_0 and N_1 are the numbers of events before and after T, the above method extends to sampling in turn from

$$x_0 | \cdots \sim \Gamma(\alpha + N_0, \beta + T),$$

$$x_1 | \cdots \sim \Gamma(\alpha + N_1, \beta + (L - T)),$$

and

$$\beta | \cdots \sim \Gamma(e + 2\alpha, f + x_0 + x_1),$$

forming a Markov chain with a three-dimensional state space $\{(x_0, x_1, \beta)\}$. Note that for the sake of clarity and compactness we write '$| \cdots$' to mean 'given all other variables' — including the data.

The hierarchical model using random β makes more sense now: the effect is to 'borrow strength' in estimation from both halves of the data together: x_0 and x_1 are conditionally independent given β, but are *un*conditionally *dependent*. In inference their values will be shrunk together.

Model 5: multiple change points

If there are k change points $T_1, T_2, \ldots, T_k$ with

$$x(t) = \begin{cases} x_0 & \text{if } 0 \le t < T_1 \\ x_1 & \text{if } T_1 \le t < T_2 \\ \cdots & \cdots \\ x_k & \text{if } T_k \le t < L \end{cases} ,$$

then everything is extended in a very similar way, giving a Markov chain with states $(x_0, x_1, \ldots, x_k, \beta)$.

1.2.5 Other approaches to Bayesian computation

Do we have to resort to Gibbs sampling for this application, and examples like it? Under the posterior distribution in a Bayesian formulation, the parameters $\boldsymbol{\theta}$ are generally *dependent*, so we have to compute with a multivariate distribution, often in a high number of dimensions, with arbitrarily complex patterns of dependence. Here, "compute with" could mean almost anything; examples would be to calculate a marginal (posterior) density or make a probabilistic prediction. See Bernardo and Smith (1994) and O'Hagan (1994).

There are various possible approaches to Bayesian computation:

- Exact analytic integration: this is usually only available when we make use of conjugate priors, which is in itself often an unreasonable restriction, and in any case is usually restricted to very simple formulations.

- Asymptotic analytic approximations (e.g. Laplace; see, for example, Kass *et al.*, 1988): these are somewhat awkward to set up, and can be unreliable.

- Conventional numerical methods: these require expertise and careful design to set up, and are only efficient in a low number of dimensions.

- Ordinary ("static") simulation: this is always available in principle, since any posterior distribution can be factorised as

$$p(\boldsymbol{\theta}|Y) = p(\theta_1, \theta_2, \ldots, \theta_p|Y)$$

$$= p(\theta_1|Y)p(\theta_2|\theta_1, Y) \ldots p(\theta_p|\theta_1, \ldots, \theta_{p-1}, Y)$$

  but the univariate distributions on the right hand side are rarely all available for simulation purposes (even after re-ordering).

Markov chain Monte Carlo (MCMC, also sometimes known as iterative or dynamic simulation) works even where static simulation does not, essentially because

- All simulation methods rely on the Law of Large Numbers, and this remains true (in the guise of the Ergodic theorem) when you have a Markov chain instead of an independent, identically distributed sequence.

- If you can tolerate Markov dependence, then you can update the parameters $\theta_1, \theta_2, \ldots, \theta_p$ one-by-one (or in small groups).

The result of combining these two simple points is very far-reaching indeed!

1.3 MCMC — the general idea and the main limit theorems

Having motivated the idea of MCMC by use of the Gibbs sampler in two very basic problems, we are now in a position to discuss the subject from a rather more general perspective.

Our object of interest is the *target distribution* π of a random quantity $\boldsymbol{x} \in \mathcal{X}$. In Bayesian statistics, $\boldsymbol{x}$ are the unknowns (parameters, latent variables, missing values, future data) in a statistical experiment, and π is the posterior distribution of these variables given the data Y:

$$\pi(A) = p(\boldsymbol{x} \in A|Y)$$

Henceforth in this chapter, we shall use $\boldsymbol{x}, \pi$ in this generic way, and reserve the $p(\cdot|\cdot)$ notation for discussion of specific models. One of many advantages of the generic notation is that it helps us not lose sight of other, non-Bayesian, applications of MCMC. (Although by far the greatest impact of MCMC in statistics has been in Bayesian analysis, because of the ubiquitous need there for integration, it has also found application in other contexts where variables are integrated out, for example in latent variables models, contingency tables and in models with complicated conditional likelihoods.)

The objective now is to construct a time-homogeneous discrete time Markov chain whose state space is $\mathcal{X}$ (the parameter space in Bayesian statistics), and whose limiting distribution is the specified target. That is, we want a transition kernel P such that

$$P\{\boldsymbol{x}^{(t)} \in A|\boldsymbol{x}^{(0)}\} \to \pi(A) \qquad \text{as } t \to \infty, \forall\, \boldsymbol{x}^{(0)}.$$

Having constructed such a Markov chain, in the sense of devising a transition kernel with this limiting property, we then construct it in another sense — we form a realisation of the chain $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$ and treat this *as if it was a random sample* from π.

Of course, in fact we should not be so naive as to ignore completely the fact that this is *not* a simple random sample! However, in practice we will routinely make displays (histograms, density estimates) of the empirical distribution as estimating the target, estimate moments of the target from those of the sample

$$E_\pi(g) = \int g(\boldsymbol{x})\pi(d\boldsymbol{x}) \approx \frac{1}{N}\sum_{t=1}^{N} g(\boldsymbol{x}^{(t)}) \tag{1.2}$$

(for suitable functions g), and compute probabilities under the target distribution by empirical frequencies

$$\pi(A) \approx \frac{1}{N}\sum_{t=1}^{N} I[\boldsymbol{x}^{(t)} \in A]. \tag{1.3}$$

All such computations are justified by the limit theory of Markov chains;

in order to handle the countless real applications where the space $\mathcal{X}$ is not discrete, we need these limit theorems for chains in a general state space.

1.3.1 *The basic limit theorems*

Our treatment of the limit theory for Markov chains given here is not at all complete, but will at least review the main concepts and results that are important to MCMC. A fuller treatment with the same objective can be found in Tierney (1994) and Tierney (1996), and the complete story is in Meyn and Tweedie (1993). This treatment borrows heavily from these sources.

The most important theorem in practice concerns convergence of sample means, and justifies (1.2) and (1.3) above. It requires the concepts of invariance and irreducibility. A probability distribution π is *invariant* for a transition kernel P if $\int P(\boldsymbol{x}, A)\pi(d\boldsymbol{x}) = \pi(A)$. The kernel P is *irreducible* if there exists a probability distribution, ψ say, on $\mathcal{X}$ such that $\psi(A) > 0 \Rightarrow P(\tau_A < \infty | \boldsymbol{x}^{(0)} = \boldsymbol{x}) = 1$ for all π-almost all $\boldsymbol{x} \in \mathcal{X}$, where τ_A is the hitting time $\min\{t : \boldsymbol{x}^{(t)} \in A\}$. Any such ψ is called an irreducibility distribution for P.

If $\{\boldsymbol{x}^{(t)}\}$ is an irreducible Markov chain with transition kernel P and invariant distribution π, and g is a real valued function with $\int |g(\boldsymbol{x})|\pi(d\boldsymbol{x}) < \infty$, then

$$\frac{1}{N}\sum_{t=1}^{N} g(\boldsymbol{x}^{(t)}) \to \int g(\boldsymbol{x})\pi(d\boldsymbol{x}) \tag{1.4}$$

almost surely, for π-almost all $\boldsymbol{x}^{(0)}$.

Sometimes, it is useful to say a little more — that the distribution of $\boldsymbol{x}^{(t)}$ converges to π. As in the simple discrete case, this requires the additional assumption that the chain is not periodic.

An m-cycle for an irreducible chain with kernel P is a collection of subsets $\{E_0, E_1, \ldots, E_{m-1}\}$ such that $P(x, E_{i+1 \bmod m}) = 1$ for all $x \in E_i$ and all i; the period d is the largest m for which an m-cycle exists, and the chain is aperiodic if $d = 1$.

If the chain is aperiodic, the t-step transition kernel converges:

$$||P^t(\boldsymbol{x}^{(0)}, \cdot) - \pi(\cdot)|| \to 0 \tag{1.5}$$

as $t \to \infty$, for π-almost all $\boldsymbol{x}^{(0)}$. Here, the norm is the total variation distance between two probability measures, defined by $||\nu_1 - \nu_2|| = 2\sup_A |\nu_1(A) - \nu_2(A)|$.

1.3.2 *Harris recurrence*

The assumptions of invariance and irreducibility are usually rather easy to check for a given transition kernel, so the results of the previous subsection are then available. However, when used to justify a simulation computation, they are subject to a crucial *caveat*. Both of these limit theorems apply only to π-almost all starting values $\boldsymbol{x}^{(0)}$. For routine purposes, this restriction is of little concern, but in simulation, we really need to know that we were not unlucky enough to be running our chain from an initial state in the probability-zero exceptional set!

We say that an irreducible kernel P is *Harris recurrent* if, for any irreducibility distribution ψ and any A such that $\psi(A) > 0$, we have $P\{\boldsymbol{x}^{(t)} \in A \text{ i.o.} | \boldsymbol{x}^{(0)} = \boldsymbol{x}\} = 1$ for all $\boldsymbol{x}$ (where 'i.o.' means 'infinitely often').

If the chain is Harris recurrent, then (1.4) holds for all $\boldsymbol{x}^{(0)}$, as does (1.5) if it is also aperiodic.

1.3.3 Rates of convergence

Knowing that the chain converges is not the same as knowing that it converges quickly enough to be useful. It is therefore important to try to study rates of convergence. This is a challenge for practically useful chains in general state spaces.

Only in very rare cases can numerical bounds be found for rates of convergence, and when they can, they are often very discouraging. However, there have been several successful approaches to the qualitative study of convergence.

The chain is *geometrically ergodic* if

$$||P^t(\boldsymbol{x}^{(0)}, \cdot) - \pi(\cdot)|| \leq M(\boldsymbol{x}^{(0)})\rho^t$$

for finite $M(\boldsymbol{x})$, $\rho < 1$.

It is *uniformly ergodic* if for all $\boldsymbol{x}^{(0)}$,

$$||P^t(\boldsymbol{x}^{(0)}, \cdot) - \pi(\cdot)|| \leq M\rho^t.$$

Various conditions are known to imply uniform ergodicity, for example Doeblin's condition: there exists a probability measure ϕ and constants $\varepsilon < 1$, $\delta > 0$, t such that

$$\phi(A) > \varepsilon \Rightarrow P^t(\boldsymbol{x}, A) \geq \delta \quad \text{for all} \quad \boldsymbol{x}.$$

There are both positive and negative results about uniform or geometric ergodicity of popular MCMC recipes. For example, see Mengersen and Tweedie (1996), Roberts and Tweedie (1996), Roberts and Rosenthal (1999), and Mira, Møller and Roberts (1999).

A rather different approach to assessing speed of convergence is via computational complexity; for example, there are recent interesting results by Frigessi, Martinelli and Stander (1997).

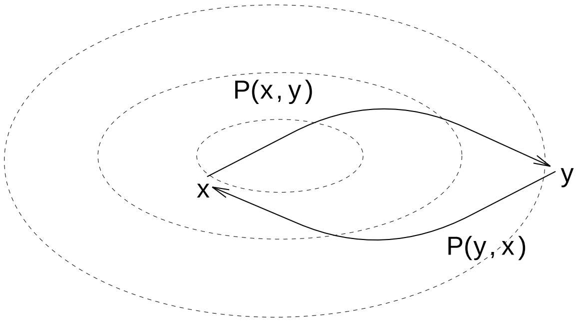

Figure 1.9 *Illustrating the idea of detailed balance. The transitions described by P are neutral with respect to the contours of probability of π.*

1.4 Recipes for constructing MCMC methods

One might think initially that to construct a Markov chain with a specified target as its limiting distribution would be a complicated matter. Fortunately, several standard 'recipes' are available to automate this task.

In this section, introducing the main recipes for MCMC methods, we assume the state space of our chain is countable, and work with a notation in which the target distribution π and the transition kernel P are expressed as densities with respect to counting measure, that is, as probability mass functions. Modifications to deal with other dominating measures, such as Lebesgue measure, are straightforward.

The key idea in most practical approaches to constructing MCMC methods is reversibility or detailed balance. The target π is invariant for P if we have detailed balance (time-reversibility):

$$\pi(\boldsymbol{x})P(\boldsymbol{x},\boldsymbol{y}) = \pi(\boldsymbol{y})P(\boldsymbol{y},\boldsymbol{x})$$

for all $\boldsymbol{x},\boldsymbol{y} \in \mathcal{X}$. Detailed balance is sufficient but not necessary for invariance; however it is far easier to work with. You can think of reversibility as requiring a balance in the flow of probability; see Figure 1.9.

We will ignore the issues of irreducibility and aperiodicity for the moment.

1.4.1 The Gibbs sampler

In the Gibbs sampler, the basic step is simple: discard the current value of a single component $\boldsymbol{x}_i$, and replace it by a value $\boldsymbol{y}_i$ drawn from the *full conditional* distribution induced by π:

$$\pi(\boldsymbol{x}_i|\boldsymbol{x}_{-i}),$$

keeping the current values of other variables: $\boldsymbol{y}_{-i} = \boldsymbol{x}_{-i}$ (where "$-i$" stands for $\{j : j \neq i\}$). Then we are using the kernel

$$P(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}_i | \boldsymbol{x}_{-i}) I[\boldsymbol{x}_{-i} = \boldsymbol{y}_{-i}],$$

and detailed balance holds because given $\boldsymbol{x}_{-i}$, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are independent, and identically distributed as $\pi(\boldsymbol{x}_i | \boldsymbol{x}_{-i})$.

This recipe was named the Gibbs sampler by Geman and Geman (1984), whose work brought the idea to the attention of spatial statisticians. However, it is earlier than that: it was well known as the 'heat bath' by statistical physicists, see for example, Creutz (1979), but the earliest appearance I know of is in statistics, in a Finnish Ph.D. thesis by Suomela (1976).

1.4.2 The Metropolis method

In the Metropolis method, we find a candidate new value (or "proposal") $\boldsymbol{y}$ by drawing $\boldsymbol{y}_i$ from an arbitrary density $q_i(\boldsymbol{y}_i; \boldsymbol{x})$ parameterised by $\boldsymbol{x}$, and setting $\boldsymbol{y}_{-i} = \boldsymbol{x}_{-i}$. We write $q_i(\boldsymbol{y}_i; \boldsymbol{x}) = q_i(\boldsymbol{x}, \boldsymbol{y})$, and impose the symmetry requirement $q_i(\boldsymbol{x}, \boldsymbol{y}) = q_i(\boldsymbol{y}, \boldsymbol{x})$. (Note the deliberate reversal of the order of arguments: $q_i(\boldsymbol{y}_i; \boldsymbol{x})$ is a density in $\boldsymbol{y}_i$ parameterised by $\boldsymbol{x}$, while $q_i(\boldsymbol{x}, \boldsymbol{y})$ is a transition kernel, and so the arguments are used in the conventional time-oriented order.)

This proposal is accepted as the next state of the chain with probability

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\} = \min\left\{1, \frac{\pi(\boldsymbol{y}_i | \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_i | \boldsymbol{x}_{-i})}\right\}, \tag{1.6}$$

and otherwise $\boldsymbol{x}$ is left unchanged.

This recipe is due to Metropolis, *et al.* (1953). Note that the target density π is only needed up to proportionality, and then only at two values, the current and proposed next states.

1.4.3 The Metropolis-Hastings sampler

In a paper astonishingly overlooked by statisticians for nearly 20 years, Hastings (1970) introduced an important generalisation of Metropolis, in which symmetry of q is not needed; the acceptance probability becomes:

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y}) q_i(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) q_i(\boldsymbol{x}, \boldsymbol{y})}\right\} = \min\left\{1, \frac{\pi(\boldsymbol{y}_i | \boldsymbol{y}_{-i}) q_i(\boldsymbol{x}_i; \boldsymbol{y})}{\pi(\boldsymbol{x}_i | \boldsymbol{x}_{-i}) q_i(\boldsymbol{y}_i; \boldsymbol{x})}\right\}. \tag{1.7}$$

The optimality in some senses of this particular choice of $\alpha(\boldsymbol{x}, \boldsymbol{y})$ over any other choice preserving detailed balance is demonstrated by Peskun (1973).

Note that Metropolis is the special case where q is symmetric, and Gibbs the special case where the proposal density $q_i(\boldsymbol{y}_i; \boldsymbol{x})$ is just the full conditional $\pi(\boldsymbol{y}_i | \boldsymbol{x}_{-i}) = \pi(\boldsymbol{y}_i | \boldsymbol{y}_{-i})$, so that the acceptance probability is 1.

1.4.4 Proof of detailed balance

The proof of correctness of each is the same: the choice of acceptance probability simply ensures that detailed balance is satisfied.

For $\boldsymbol{x} \neq \boldsymbol{y}$,

$$\begin{aligned}\pi(\boldsymbol{x})P(\boldsymbol{x},\boldsymbol{y}) &= \pi(\boldsymbol{x}_{-i})\pi(\boldsymbol{x}_i|\boldsymbol{x}_{-i})q_i(\boldsymbol{y}_i;\boldsymbol{x})\alpha(\boldsymbol{x},\boldsymbol{y}) \\ &= \pi(\boldsymbol{x}_{-i})\min\{R(\boldsymbol{x},\boldsymbol{y}),R(\boldsymbol{y},\boldsymbol{x})\},\end{aligned}$$

from (1.7), where $R(\boldsymbol{x},\boldsymbol{y}) = \pi(\boldsymbol{x}_i|\boldsymbol{x}_{-i})q_i(\boldsymbol{y}_i;\boldsymbol{x})$. The term R and hence the whole expression above is symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$ (recall that $\boldsymbol{x}_{-i} = \boldsymbol{y}_{-i}$). So detailed balance holds. (Note that we have only used the fact that

$$\frac{\alpha(\boldsymbol{x},\boldsymbol{y})}{\alpha(\boldsymbol{y},\boldsymbol{x})} = \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\frac{q_i(\boldsymbol{y},\boldsymbol{x})}{q_i(\boldsymbol{x},\boldsymbol{y})}.$$

The argument for the particular choice of $\alpha(\boldsymbol{x},\boldsymbol{y})$ in (1.7) will be made in Section 1.6.2.)

This argument for the Hastings method obviously covers Gibbs and Metropolis *a fortiori*.

1.4.5 Updating several variables at once

Each of the Gibbs, Metropolis and Hastings methods is equally valid if a group of variables $\boldsymbol{x}_A = \{\boldsymbol{x}_j : j \in A\}$ is updated simultaneously; each uses the full conditional $\pi(\boldsymbol{x}_A|\boldsymbol{x}_{-A})$. You could update *all* variables at once in Metropolis or Hastings. (It is a subtle question whether it is a good idea to update many variables.)

An important special case arises where the variables in $\boldsymbol{x}_A$ are *conditionally independent* (under the full conditional). They can then be updated in parallel.

1.4.6 The role of the full conditionals

All of the basic methods use the full conditionals $\pi(\boldsymbol{x}_A|\boldsymbol{x}_{-A})$, where A indexes the variables being updated. In Gibbs, you have to *draw* from this distribution; in Metropolis and Hastings, you only have to *evaluate* it (up to a multiplicative constant) at the old and new values.

1.4.7 Combining kernels to make an ergodic sampler

All of the methods above satisfy detailed balance, and hence preserve the equilibrium distribution: if

$$\boldsymbol{x} \sim \pi$$

before the transition, then so it will afterwards.

To ensure that this is also the limiting distribution of the chain (ergodicity), we must combine such kernels to make a Markov chain transition mechanism that is irreducible (and aperiodic).

To do that, scan over the available kernels (indexed by i or A) either systematically or randomly, or in various other ways that are valid, provided you visit each variable often enough. You can use different recipes (Gibbs, Metropolis,...) for different A. The most common strategies for combining kernels $P_1, P_2, \ldots, P_m$ are the systematic *cyclic* combination giving an overall kernel

$$P = P_1 P_2 \cdots P_m$$

or the equally-weighted random or *mixture* kernel

$$P = \frac{1}{m} \sum_{i=1}^{m} P_i.$$

Time in a MCMC simulation is usually measured in *sweeps*, the smallest period such that the chain is time-homogeneous, for example, after m individual transitions if the cyclic kernel is being used.

Note that the mixture kernel preserves detailed balance, while the cyclic one does not, so that reversibility at the sweep time scale is lost; of course π remains invariant for both combinations.

1.4.8 Common choices for proposal distribution

The user has a completely free choice of proposal distribution; there is no need even to worry about dividing by zero in (1.7), since $\boldsymbol{y}_i$ with $q_i(\boldsymbol{y}_i; \boldsymbol{x}) = 0$ will (almost surely) not get proposed! Nevertheless, typically, one of a small number of standard specifications is very often used.

Independence Metropolis-Hastings. If the proposed new state $\boldsymbol{y}$ is independent of the current $\boldsymbol{x}$ (so in particular we are proposing to update all components of the state simultaneously), then $q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y})$, say, and the acceptance probability simplifies to

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \min \left\{ 1, \frac{w(\boldsymbol{y})}{w(\boldsymbol{x})} \right\},$$

where $w(\boldsymbol{x}) = \pi(\boldsymbol{x})/q(\boldsymbol{x})$.

This choice is of little use in practical terms (except perhaps in splitting, see Section 1.6.2) but often yields kernels amenable to theoretical investigation.

Random walk Metropolis. If $q_i(\boldsymbol{x}, \boldsymbol{y}) = q_i(\boldsymbol{y}_i - \boldsymbol{x}_i)$ where $q_i(\cdot)$ is a density function symmetric about 0, then

$$\frac{q_i(\boldsymbol{y}, \boldsymbol{x})}{q_i(\boldsymbol{x}, \boldsymbol{y})} = 1$$

so the acceptance probability simplifies; the proposal amounts to adding a random walk increment $\sim q_i$ to the current $\boldsymbol{x}_i$.

Random walk Metropolis on the log scale. When a component $\boldsymbol{x}_i$ of the state vector is necessarily positive, it may be convenient to only propose changes to its value that leave it positive, in which case a multiplicative rather than additive update is suggested. If the proposed increment to $\log \boldsymbol{x}_i$ has any distribution symmetric about 0, then we find

$$\frac{q_i(\boldsymbol{y},\boldsymbol{x})}{q_i(\boldsymbol{x},\boldsymbol{y})} = \frac{\boldsymbol{y}_i}{\boldsymbol{x}_i}.$$

1.4.9 Comparing Metropolis-Hastings to rejection sampling

There is a superficial resemblance of Metropolis-Hastings to ordinary rejection sampling, which may cause confusion. Recall that in rejection sampling, to sample from π, we first draw $\boldsymbol{y}$ from a density q, and then accept this value with probability $\pi(\boldsymbol{y})/(Mq(\boldsymbol{y}))$, where M is any constant such that $M \geq \sup_{\boldsymbol{y}} \pi(\boldsymbol{y})/q(\boldsymbol{y})$. If the generated $\boldsymbol{y}$ is not accepted, this procedure is repeated until it is. As with Metropolis-Hastings, π and q are needed only up to proportionality. The crucial differences are that in Metropolis-Hastings: (a) π/q need not be bounded, (b) you do *not* repeat if the proposal is rejected, and (c) you end up with a Markov chain, not an independent sequence.

1.4.10 Example: Weibull/Gamma experiment

Let us consider a different but still very simple example, where Gibbs sampling would not be straightforward. Our data will be a random sample, possibly censored, from the Weibull(ρ, κ) distribution:

$$p(Y|\rho,\kappa) = \kappa^m \rho^{m\kappa} \textstyle\prod_U Y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum Y_i^\kappa\right)$$

where m and $\prod_U$ are the number of and product over uncensored observations. We place independent Gamma priors on ρ and κ:

$$p(\rho,\kappa) \propto \rho^{\alpha-1} e^{-\beta\rho} \kappa^{\gamma-1} e^{-\delta\kappa}$$

The resulting posterior is

$$\begin{aligned} p(\rho,\kappa|Y) \quad \propto \quad & \kappa^m \rho^{m\kappa} \textstyle\prod_U Y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum Y_i^\kappa\right) \\ & \rho^{\alpha-1} e^{-\beta\rho} \kappa^{\gamma-1} e^{-\delta\kappa} \end{aligned}$$

which is not a standard distribution.

Let us define a Markov chain with states $\boldsymbol{x} = (\rho, \kappa)$ and limiting distribution $\pi(\boldsymbol{x}) = p(\rho, \kappa|Y)$. The full conditionals for the two parameters are

$$p(\rho|\kappa) \quad \propto \quad \rho^{m\kappa} \exp\left(-\rho^\kappa \textstyle\sum Y_i^\kappa\right) \rho^{\alpha-1} e^{-\beta\rho}$$

$$p(\kappa|\rho) \quad \propto \quad \kappa^m \rho^{m\kappa} \textstyle\prod_U Y_i^{\kappa-1} \exp\left(-\rho^\kappa \sum Y_i^\kappa\right) \kappa^{\gamma-1} e^{-\delta\kappa}.$$

This is hardly of standard form, so Gibbs is problematical, but the full conditionals are easily evaluated for a Metropolis or Hastings algorithm.

An easily implemented Metropolis method for this setting would consist of the following ingredients:

1. alternate between updating ρ and κ,
2. propose a new value for the parameter from a distribution symmetric about its present value,
3. reject the update if the result is negative,
4. otherwise, accept it with probability (e.g.) $\min\{1, p(\rho'|\kappa)/p(\rho|\kappa)\}$.

1.4.11 Cyclones example, continued

Model 6: another hyperparameter

Let's now suppose α is also unknown, with, *a priori*,

$$\alpha \sim \Gamma(c, d)$$

for fixed constants c and d. (In our analysis of the cyclones data, we took $c = d = 2$.) This last change means that Gibbs sampling is not enough. In a Markov chain with states $\boldsymbol{x} = (x_0, x_1, \ldots, x_k, \alpha, \beta)$, we can update α using a random walk Metropolis move, on the $\log(\alpha)$ scale: the acceptance ratio is

$$\min\left\{1, \frac{p(\log\alpha'|\cdots)}{p(\log\alpha|\cdots)}\right\}$$

which simplifies to

$$\min\left\{1, \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha')}\right)^{k+1} \left(\frac{\alpha'}{\alpha}\right)^{c} \left(e^{-d}\beta^{k+1} \prod x_j\right)^{\alpha'-\alpha}\right\}$$

Model 7: unknown change points

If $x_0, x_1, \ldots, x_k$ are unknown, so probably are the times of the change points $T_1 < T_2 < \cdots < T_k$. The state vector is now $\boldsymbol{x} = (x_0, x_1, \ldots, x_k, T_1, T_2, \ldots, T_k, \alpha, \beta)$.

Let us assume *a priori*

$$p(T_1, T_2, \ldots, T_k) \propto T_1(T_2 - T_1)\ldots(T_k - T_{k-1})(L - T_k),$$

a joint density providing a gentle preference against two changes occurring too closely in succession (this is actually the joint distribution of the even-numbered order statistics for a sample of size $2k+1$ from $U(0, L)$).

The posterior marginal or joint conditional distributions are quite complex, for this or any prior, so Metropolis-Hastings is needed. The details

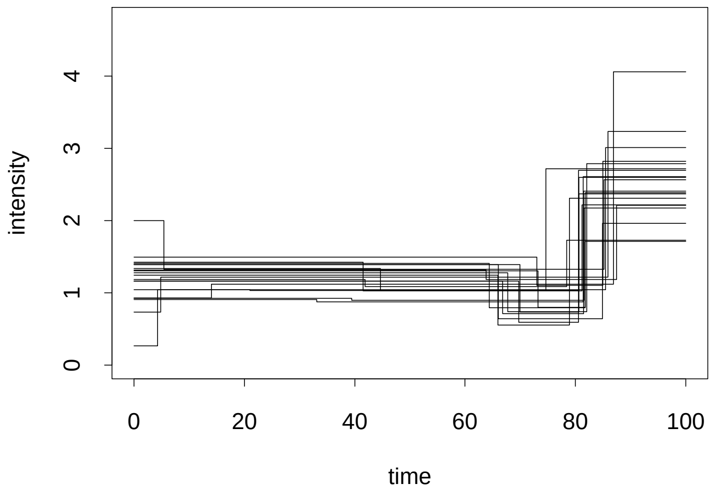

Figure 1.10 *Posterior sample of step functions $x(t)$ for model 7 with $k = 2$, applied to cyclones data.*

are a little messy but straightforward. For a proposal drawing T'_j uniformly from $[T_{j-1}, T_{j+1}]$, the acceptance probability is

$$\min\left\{1, (\text{likelihood ratio})\frac{(T'_j - T_{j-1})(T_{j+1} - T'_j)}{(T_j - T_{j-1})(T_{j+1} - T_j)}\right\}.$$

A sample of step functions drawn from the resulting MCMC sample is shown in Figure 1.10.

1.5 The role of graphical models

Graphical modelling provides a powerful language for specifying and understanding statistical models.

Graphs consist of vertices representing variables, and edges (directed or otherwise) that express conditional dependence properties. For a full treatment of the theory, see Lauritzen (1996).

1.5.1 Directed acyclic graphs

The DAG (directed acyclic graph) — a graph in which all edges are directed, and there are no directed loops — expresses the natural factorisa-

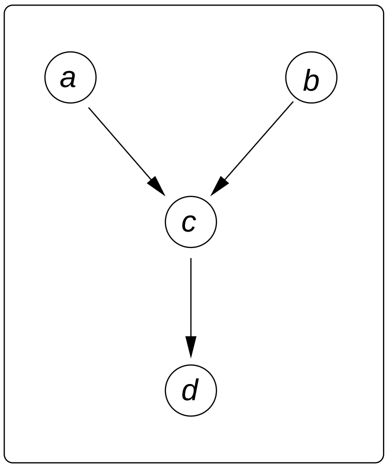

Figure 1.11 *A simple directed acyclic graph on four variables.*

tion of a joint distribution into factors each giving the joint distribution of a variable $\boldsymbol{x}_v$ given the values of its *parents* $\boldsymbol{x}_{\mathrm{pa}(v)}$; for example, in Figure 1.11,

$$\pi(a, b, c, d) = \pi(a)\pi(b)\pi(c|a, b)\pi(d|c)$$

In general, we can write

$$\pi(\boldsymbol{x}) = \prod_{v \in V} \pi(\boldsymbol{x}_v | \boldsymbol{x}_{\mathrm{pa}(v)}) \tag{1.8}$$

(see Figure 1.12), which in turn implies a *Markov property*, that variables are conditionally independent of their non-descendants, given their parents.

From the perspective of setting up MCMC methods, the graphical structure assists in identifying which terms need be included in a full conditional. Equation (1.8) implies

$$\pi(\boldsymbol{x}_v | \boldsymbol{x}_{-v}) \propto \pi(\boldsymbol{x}_v | \boldsymbol{x}_{\mathrm{pa}(v)}) \prod_{w : v \in \mathrm{pa}(w)} \pi(\boldsymbol{x}_w | \boldsymbol{x}_{\mathrm{pa}(w)})$$

where the right hand side has one term for the variable of interest itself, and one for each of its children.

Graphical modelling, the construction of MCMC methods through full conditional distributions, and good practice in statistical model building all exploit the same modular structure.

A concrete example of this modularity has already been seen implicitly; in Section 1.4.7, we discussed how an ergodic kernel might be assembled (by cycling or mixing) from a collection of kernels $P_1, P_2, \ldots$ that were individually in detailed balance but not irreducible.

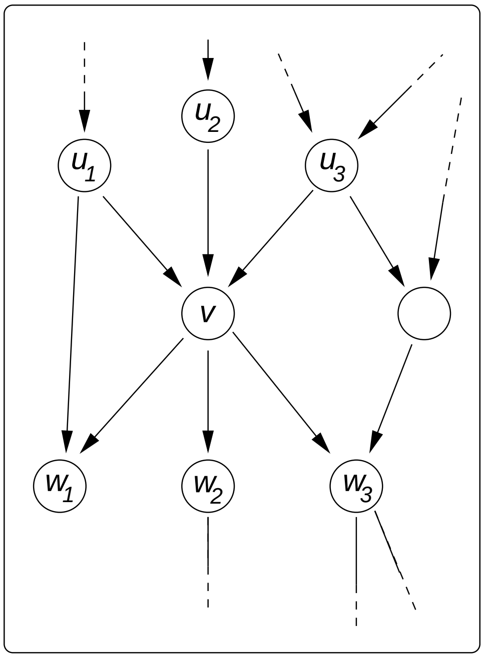

Figure 1.12 *A larger directed acyclic graph: the vertices labelled* $\{u_i\}$ *are the parents of* v*, and* $\{w_i\}$ *are its children.*

1.5.2 Undirected graphs, and spatial modelling

Directed acyclic graphs are a natural representation of the way we usually *specify* a statistical model (directionally, disease $\rightarrow$ symptom, past $\rightarrow$ future, parameters $\rightarrow$ data), but

- sometimes (e.g. spatial models) there is *no natural direction*;
- in *understanding associations* between variables implied by a model, however specified, directions can confuse; and
- these associations represent the full conditionals needed in *setting up MCMC* methods.

To form the conditional independence graph for a multivariate distribution, draw an (undirected) edge between variables α and β if they are **not** conditionally independent given all other variables.

Markov properties

The Markov property is familiar from temporal stochastic processes, where we learn that it may be expressed in several equivalent ways. For variables located on an arbitrary graph, the situation is more subtle: we can distinguish four related properties, each capturing an aspect of Markovness.

P: Pairwise Non-adjacent pairs of variables are conditionally independent given the rest (see definition of graph).

L: Local Conditional only on adjacent variables (neighbours), each variable is independent of all others (so that full conditionals are simplified).

G: Global Any two subsets of variables separated by a third are conditionally independent given the values of the third subset.

F: Factorisation The joint distribution factorises as a product of functions on cliques (that is, maximal complete subgraphs).

The four properties are illustrated in Figures 1.13 and 1.14.

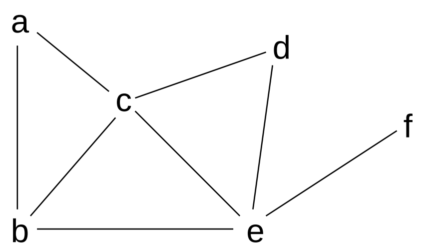

Figure 1.13 *Illustrating the pairwise, local and factorisation Markov properties:* $P : c \perp f|(a,b,d,e)$, $L : d \perp (a,b,f)|(c,e)$ *and* $F : \pi(a,b,c,d,e,f) = \psi_1(a,b,c)\psi_2(b,c,e)\psi_3(c,d,e)\psi_4(e,f)$.

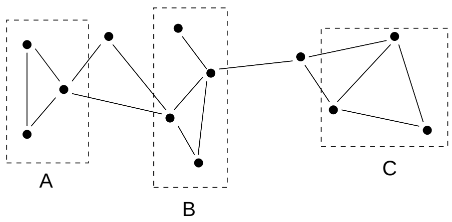

Figure 1.14 *Illustrating the global Markov property:* $G : A \perp C|B$.

It is always true that $F \Rightarrow G \Rightarrow L \Rightarrow P$, but these four Markov properties are in general different (there are easy counter-examples for each of the reverse implications). However, in many statistical contexts, the four properties are the same; a sufficient but not necessary condition is that the joint distribution has the positivity property ("any values realisable individually are realisable jointly"). This result includes the Clifford-Hammersley theorem (Markov random field = Gibbs distribution, $L = F$). See, for example, Besag (1974), Clifford (1990). A typical context in which the Markov properties may not coincide is where there are logical implications between some subsets of variables.

For directed acyclic graphs, the situation is simpler: the directed local Markov property is *always* equivalent to the directed graph factorisation criterion: $DL = DF$ (subject to existence of a dominating product measure).

Modelling directly with an undirected graph

With a DAG, because of the acyclicity, *any* set of conditional distributions $\pi(\boldsymbol{x}_v|\boldsymbol{x}_{\mathrm{pa}(v)})$ combine to form a consistent joint distribution.

In an undirected graph, however, we need consistency conditions on the full conditionals $\pi(\boldsymbol{x}_v|\boldsymbol{x}_{-v})$ (using L, this is equal to $\pi(\boldsymbol{x}_v|\boldsymbol{x}_{\partial v})$, where ∂v denotes the neighbours of v). The only safe strategy is to use property F, to model the joint distribution as a product of functions on cliques

$$\pi(\boldsymbol{x}) = \prod_C \psi_C(\boldsymbol{x}_C)$$

We can then use property L, the local Markov property, to read off the full conditionals needed to set up MCMC:

$$\pi(\boldsymbol{x}_v|\boldsymbol{x}_{-v}) = \prod_{C:v\in C} \psi_C(\boldsymbol{x}_C) = \pi(\boldsymbol{x}_v|\boldsymbol{x}_{\partial v}).$$

Most of the applications in Besag, *et al.* (1995) have a spatial flavour, and provide illustrations of this style of modelling.

1.5.3 Chain graphs

In hierarchical spatial models, we need a hybrid modelling strategy: there will be some directed and some undirected edges. If there are no one-way cycles, the graph can be arranged to form a DAG with composite nodes called *chain components* Δ_t, that are the connected subgraphs remaining when all directed edges are removed: we call this a *chain graph*.

Model specification uses an appropriate combination of the two approaches; this builds a joint distribution

$$\begin{aligned}\pi(\boldsymbol{x}) &= \prod_t \pi(\boldsymbol{x}_{\Delta_t}|\boldsymbol{x}_{\mathrm{pa}(\Delta_t)}) \\ &= \prod_t \prod_{C\in\mathcal{C}_t} \psi_C(\boldsymbol{x}_C)\end{aligned}$$

where $\mathcal{C}_t$ are the cliques in an undirected graph with nodes $(\Delta_t, \mathrm{pa}(\Delta_t))$ and undirected edges consisting of (a) those already in Δ_t, (b) the links between Δ_t and parents, with directions dropped, and (c) links between all members of $\mathrm{pa}(\Delta_t)$.

1.6 Performance of MCMC methods

There are two main issues to consider when evaluating the performance of a Markov chain used for Monte Carlo calculations, for example in choosing between alternative chains for a particular target, or in assessing if a particular run of a particular chain is adequate for its purpose:

- Convergence (how quickly does the distribution of $\boldsymbol{x}^{(t)}$ approach $\pi(\boldsymbol{x})$?);
- Efficiency (how well are functionals of $\pi(\boldsymbol{x})$ estimated from $\{\boldsymbol{x}^{(t)}\}$?) .

In both cases, performance will be measured in relation to the computing effort expended, and of course this effort should be measured in seconds, not sweeps, although this does beg questions about whether for example, two rival methods have been coded comparably efficiently.

In this section, we will review some of the issues involved in these assessments, and some of the methods proposed. However, we should not lose sight of a third factor:

- Simplicity (how convenient is the method to code reliably and to use?)

We return to some issues of implementation in Section 1.11.

Contrary to a popular misconception, it should not be supposed that Gibbs is necessarily superior to other methods on *any* of these three criteria, so it does not provide a gold standard for comparison.

1.6.1 Monitoring convergence

An active and important subfield of MCMC research has aimed at investigating and developing methods for analysis of a Markov chain realisation, to determine empirically whether the chain can safely be said to 'have converged', and to provide a reliable basis for estimation of aspects of the target distribution.

It is undoubtedly important in practice to obtain some reassurance on these issues, and grossly irresponsible, for example, to accept at face value a statistical analysis of an important real-world problem, where this analysis is computed by a MCMC sampler whose performance on the model in question is unknown. However, there is a limit to the degree of reassurance that can be obtained from an empirical analysis, and this should always be supplemented by a sound understanding of the qualitative form of the target distribution, with an eye to the possible presence of features that the chosen MCMC sampler may have difficulty with.

Attempts to place the activity of convergence monitoring on a firm logical footing seem unconvincing. Apart from some contrived exceptional cases, no finite segments of Markov chain path are truly in equilibrium, so the question is not a deterministic decision problem. But it is also wrong to regard the issue as one of hypothesis testing. We *know* the sample is not 'in equilibrium', so the logic of testing that is aimed at detection of departures from the null hypothesis is not relevant. Whether the sample is large enough to enable such detection is inevitably bound up with the quantity – the closeness of the approximation to equilibrium – that is being measured.

Finally, of course, there can never be any protection based on an empirical analysis alone against the possibility that immediately after monitoring ceases, the chain jumps into a part of the parameter space that it has not previously visited!

Notwithstanding all these caveats, diagnostic techniques, intelligently used, are valuable, and the reader is referred to Brooks and Gelman (1998) for a thorough guide to the topic.

Some researchers have expressed optimism in the last year or two that perfect (or exact) simulation – the organisation of a MCMC simulation so that it delivers a sample guaranteed to be an exact draw from the target – will make reliance on diagnostics redundant. This may or may not happen, but it is still in the future! For an introduction to the role of *coupling from the past* in perfect simulation, see Section 1.9.

1.6.2 Monte Carlo standard errors

Since any Monte Carlo method is used to provide numerical estimates of *deterministic* quantities, even if these quantities arise in a stochastic model, it is important to be aware of, and in general to evaluate, the *Monte Carlo standard error*, of estimated quantities, which should not of course be confused with the standard deviation of the posterior!

Because of Markov dependence, this is not quite straightforward, even though we (mostly) just use empirical averages as estimates.

Consider a Markov chain in equilibrium. Estimating $E_\pi(g) = \int g(\boldsymbol{x})\pi(d\boldsymbol{x})$ by $N^{-1}\sum_{t=1}^{N} g(\boldsymbol{x}^{(t)}) = \bar{g}_N$, we find

$$\begin{aligned}\mathrm{var}(\bar{g}_N) &= N^{-2}\sum_{t=-N+1}^{N-1}(N-|t|)\gamma_t \\ &\sim N^{-1}\sum_{t=-\infty}^{\infty}\gamma_t\end{aligned}$$

where $\gamma_t = \mathrm{cov}_{\pi,P}\{g(\boldsymbol{x}^{(s)}), g(\boldsymbol{x}^{(s+t)})\}$ (note that unlike the equilibrium mean and variance, which depend only on π, the autocovariances depend also on the kernel P). This quantity is equivalently written

$$N^{-1}\mathrm{var}_\pi(g)\sum_{t=-\infty}^{\infty}\rho_t = N^{-1}\mathrm{var}_\pi(g)\tau(g) = N^{-1}v(g,\pi,P)$$

where ρ_t is the corresponding autocorrelation at lag t. The factor $\tau(g)$ by which the variance of the sample mean exceeds the value obtained in independent sampling is sometimes called the integrated autocorrelation time; it depends on π, g and the transition kernel P.

Several possibilities for estimating $\mathrm{var}(\bar{g}_N)$, $\tau(g)$ or $v(g, \pi, P)$ have been proposed, most of which are in common use:

- Blocking (also known as batching) (Hastings, 1970)
- Time-series methods (e.g. Sokal, 1989)
- Initial series estimates (Geyer, 1992)
- Regeneration (Mykland, Tierney and Yu, 1995)

There are also Central Limit theorems for Markov chain averages, of the form

$$\sqrt{N}\,(\bar{g}_N - E_\pi(g)) \xrightarrow{D} N(0, v(g, \pi, P)).$$

The theorems take various forms, but broadly speaking, we need ergodicity of the Markov chain, a finite variance of g and sufficiently good mixing that $v(g, \pi, P)$ is finite. Kipnis and Varadhan (1986) give such a result assuming reversibility, while Gordon and Lifšic (1978) do not need this condition, but make stronger assumptions elsewhere.

There are results comparing $v(g, \pi, P)$ for different kernels P. The best known is due to Peskun (1973), proved for a general state space setting by Tierney (1998); this states that if P and Q are two kernels with the same invariant distribution π, with P dominating Q off the diagonal — that is, $P(x, B) \geq Q(x, B)$ for all B not containing x — then $v(g, \pi, P) \leq v(g, \pi, Q)$ for all $g \in L_2(\pi)$, so that P is preferable. In particular, among all Metropolis-Hastings methods for a given π and proposal mechanism, that maximising the acceptance probability $\alpha(\boldsymbol{x}, \boldsymbol{y})$ is best: this explains the almost universal use of the acceptance probability formula (1.7). There are other recent interesting results on ordering Markov chains in Mira and Geyer (1999).

Blocking (or batching)

After satisfying ourselves that our Markov chain is in equilibrium, we divide a run of length N into b blocks of k consecutive samples. Then if k is large, so that block means are approximately independent, and b is also large, so that between-block variability can be estimated adequately, we have

$$\mathrm{var}(\bar{g}_N) \approx \{b(b-1)\}^{-1} \sum_{i=1}^{b} \{\bar{g}_{k,i} - \bar{g}_{N,1}\}^2$$

where

$$\bar{g}_{k,i} = k^{-1} \sum_{j=(i-1)k+1}^{ik} g(\boldsymbol{x}^{(j)})$$

is the mean of the i^{th} block of length k.

This extends to nonlinear functionals of expectations; see Aykroyd and Green (1991).

Using empirical covariances

As is well-known from the time-series literature, we cannot estimate $\sum_{-\infty}^{\infty} \gamma_t$ consistently by $\sum_{-\infty}^{\infty} \widehat{\gamma}_t$, where $\widehat{\gamma}_t$ is the lag-t product-moment autocovariance of $g(\boldsymbol{x}^{(t)})$: we should, for example, use some kind of windowed estimate $\sum_{-\infty}^{\infty} w(t)\widehat{\gamma}_t$ instead. Since $\sum_{-\infty}^{\infty} \gamma_t$ is proportional to the spectral density function evaluated at 0, this is a well-studied problem. See, for example, Priestley (1981, p. 225). A convenient estimator of $\tau(g)$ in practice is the truncated periodogram estimator of Sokal (1989): $\widehat{\tau} = \sum_{t=-M}^{M} \widehat{\gamma}_t / \widehat{\gamma}_0$, where M is the smallest integer $\geq 3\widehat{\tau}$.

Initial series estimators

Geyer (1992) observes that, for a reversible ergodic chain, $\gamma_{2t} + \gamma_{2t+1}$ is non-negative, decreasing and convex in t. This suggests a class of estimators obtained by truncating $\sum_{t:|t|<M} \widehat{\gamma}_t$ when one or other of these properties is first violated.

Regeneration

Regeneration points in the Markov chain path are times $\{\tau_i, i = 1, 2, \ldots\}$ such that the *tours* $(\boldsymbol{x}^{(\tau_{i-1}+1)}, \boldsymbol{x}^{(\tau_{i-1}+2)}, \ldots, \boldsymbol{x}^{(\tau_i)})$ are independent and identically distributed for $i = 1, 2, \ldots$. If we can find such times, then renewal theory and ratio estimation give estimates of posterior expectations, and simulation standard errors that are valid without quantifying Markov dependence.

More specifically, let

$$L_i = (\tau_i - \tau_{i-1}), \qquad G_i = \sum_{t=\tau_{i-1}+1}^{\tau_i} g(\boldsymbol{x}^{(t)})$$

be the length of the i^{th} tour, and the total of a function g evaluated at the states visited in the tour, then (L_i, G_i) are i.i.d. pairs, and

$$\frac{\sum_{i=1}^{n} G_i}{\sum_{i=1}^{n} L_i} \stackrel{a.s.}{\to} E_\pi(g) \quad \text{as } n \to \infty$$

by the renewal theorem.

Finding such regeneration times is easy in a discrete state space chain, since the chain regenerates at visits to any specified state. For general state space chains, the process of finding regeneration points is facilitated by use of Nummelin's splitting technique.

Regeneration using Nummelin's splitting

Suppose the transition kernel $P(\boldsymbol{x}, A)$ satisfies

$$P(\boldsymbol{x}, A) \geq s(\boldsymbol{x})\nu(A)$$

where ν is a probability measure, and s is a non-negative function such that $\int s(\boldsymbol{x})\pi(d\boldsymbol{x}) > 0$.

Let $r(\boldsymbol{x}, \boldsymbol{y})$ denote the Radon-Nikodym derivative

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{s(\boldsymbol{x})\nu(d\boldsymbol{y})}{P(\boldsymbol{x}, d\boldsymbol{y})} \leq 1.$$

Now, given a realisation $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots$ from P, construct conditionally independent 0/1 random variables $S^{(0)}, S^{(1)}, \ldots$ with

$$P(S^{(t)} = 1 | \ldots) = r(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)})$$

Then by simple probability calculus we find

$$P(S^{(t)} = 1 | \boldsymbol{x}^{(\leq t)}, S^{(<t)}) = s(\boldsymbol{x}^{(t)})$$

and

$$P(\boldsymbol{x}^{(t+1)} \in A | \boldsymbol{x}^{(\leq t)}, S^{(<t)}, S^{(t)} = 1) = \nu(A)$$

that is, we can post-process the chain stochastically to generate binary 'splitting variables'. Whenever $S^{(t)} = 1$, the next state $\boldsymbol{x}^{(t+1)}$ is drawn from ν, independently of the past! The chain regenerates.

The problem with using the technique in practice is that in the Markov chains we tend to create for Bayesian computation, $P(\boldsymbol{x}, A)$ is difficult to handle algebraically, and/or impossible to bound below by $s(\boldsymbol{x})\nu(A)$ as required. Mykland, Tierney and Yu (1995) examine the possibilities of exploiting splitting in practical MCMC. Their perspective introduces another role for naive MCMC methods such as Independence Metropolis-Hastings (see Section 1.4.8), which although of limited efficiency may be amenable to algebraic manipulation to discover the required bounds.

1.7 Reversible jump methods: Metropolis-Hastings in a more general setting

The formulation of Metropolis-Hastings given in Subsection 1.4.3 is the standard one, and close to the original specification of Hastings (1970). It is already fairly general in that the densities $\pi(\boldsymbol{x})$ and $q(\boldsymbol{x}, \boldsymbol{y})$ appearing there may be with respect to an arbitrary measure on $\mathcal{X}$, so that both discrete and continuous distributions in any finite number of dimensions are covered. However, the formulation is a little restrictive when we come to consider MCMC samplers for certain new tasks, most notably problems where the dimension of the parameter varies, so that there is no elementary dominating measure for the target distribution.

The more general Metropolis-Hastings method we define here addresses this wider range of problems, but also offers a new perspective on the standard formulation, one that has certain pedagogical merits, and also may sometimes be more straightforward to implement. This reversible jump approach is based on Green (1995); see also Tierney (1998).

The detailed balance condition for a general transition kernel P and its invariant distribution π is written in integral form as

$$\int_{(\boldsymbol{x},\boldsymbol{y})\in A\times B} \pi(d\boldsymbol{x})P(\boldsymbol{x},d\boldsymbol{y}) = \int_{(\boldsymbol{x},\boldsymbol{y})\in A\times B} \pi(d\boldsymbol{y})P(\boldsymbol{y},d\boldsymbol{x}) \tag{1.9}$$

for all Borel sets $A, B \subset \mathcal{X}$. If P is constructed in two steps, according to the Metropolis-Hastings paradigm, we make a transition by first drawing a proposed new state $\boldsymbol{y}$ from the proposal measure $q(\boldsymbol{x}, d\boldsymbol{y})$ and then accepting it with probability $\alpha(\boldsymbol{x}, \boldsymbol{y})$. If we reject, we stay in the current state, so that $P(\boldsymbol{x}, d\boldsymbol{y})$ has an atom at $\boldsymbol{x}$. This makes an equal contribution to each side of equation(1.9), so can be neglected, and we are left with the requirement

$$\int_{(\boldsymbol{x},\boldsymbol{y})\in A\times B} \pi(d\boldsymbol{x})\alpha(\boldsymbol{x},\boldsymbol{y})q(\boldsymbol{x},d\boldsymbol{y}) = \int_{(\boldsymbol{x},\boldsymbol{y})\in A\times B} \pi(d\boldsymbol{y})\alpha(\boldsymbol{y},\boldsymbol{x})q(\boldsymbol{y},d\boldsymbol{x}). \tag{1.10}$$

When can we 'solve' this collection of equations of measures to give an explicit equation for the function $\alpha(\boldsymbol{x}, \boldsymbol{y})$? Suppose that $\pi(d\boldsymbol{x})q(\boldsymbol{x}, d\boldsymbol{y})$ is dominated by a *symmetric* measure μ on $\mathcal{X} \times \mathcal{X}$, and has density (Radon-Nikodym density) f with respect to this μ. Then (1.10) becomes

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

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

$$\alpha(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{x},\boldsymbol{y}) = \alpha(\boldsymbol{y},\boldsymbol{x})f(\boldsymbol{y},\boldsymbol{x}).$$

As with the standard Metropolis-Hastings method, we usually take the acceptance probabilities as large as possible subject to detailed balance, so

$$\alpha(\boldsymbol{x},\boldsymbol{y}) = \min\left\{1, \frac{f(\boldsymbol{y},\boldsymbol{x})}{f(\boldsymbol{x},\boldsymbol{y})}\right\}. \tag{1.11}$$

If we wrote this rather more informally as

$$\alpha(\boldsymbol{x},\boldsymbol{y}) = \min\left\{1, \frac{\pi(d\boldsymbol{y})q(\boldsymbol{y},d\boldsymbol{x})}{\pi(d\boldsymbol{x})q(\boldsymbol{x},d\boldsymbol{y})}\right\} \tag{1.12}$$

then the similarity with the usual expression using densities (1.7) is apparent, but we must not forget that the meaning of the ratio of measures derives from equation (1.11), and assumes the existence and symmetry of μ.

The formulation in this section applies to a completely general state space Markov chain. For the particular context of spatial point processes, a very similar development was given by Geyer and Møller (1994), providing an alternative to the usual spatial birth-and-death process approach. In this setting, although the dominating measure for the target distribution is not as familiar as Lebesgue, it is perfectly explicit: models are expressed via their densities with respect to a unit rate Poisson process. Detailed balance can therefore be established directly. In other situations, the dominating measure is much less explicit, and the constructions of the following two subsections very often prove useful.

1.7.1 Explicit representation using random numbers

The general Metropolis-Hastings method of the preceding subsection hardly lives up to the claim that it offers advantages in implementation, as it seems rather abstract. Fortunately, in many cases the dominating measure and Radon-Nikodym derivatives can be generated almost automatically.

To see this, imagine 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 ν_d. At the current state $\boldsymbol{x}$, the program-writer will generate, say, r random numbers $\boldsymbol{u}$ 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: $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{u})$. The left-hand side of (1.10) becomes:

$$\int_{(\boldsymbol{x},\boldsymbol{y})\in A\times B} \pi(\boldsymbol{x})g(\boldsymbol{u})\alpha(\boldsymbol{x},\boldsymbol{y})\nu_d(d\boldsymbol{x})\nu_r(d\boldsymbol{u})$$

Now consider how the reverse transition from $\boldsymbol{y}$ to $\boldsymbol{x}$ would be made, with the aid of random numbers $\boldsymbol{u}' \sim g$ giving $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{u}')$. If the transformation from $(\boldsymbol{x}, \boldsymbol{u})$ to $(\boldsymbol{y}, \boldsymbol{u}')$ is a bijection, and if both it and its inverse are differentiable, then by the standard change-of-variable formula, the $(d+r)$-dimensional integral equality (1.10) holds if

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

whence a valid choice for α is

$$\alpha(\boldsymbol{x},\boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})g(\boldsymbol{u}')}{\pi(\boldsymbol{x})g(\boldsymbol{u})} \left| \frac{\partial(\boldsymbol{y},\boldsymbol{u}')}{\partial(\boldsymbol{x},\boldsymbol{u})} \right| \right\}. \tag{1.13}$$

It is often easier to work with this expression than the usual one, equation (1.7).

1.7.2 MCMC for variable dimension problems

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

There is a huge variety of statistical problems of this kind, where the parameter dimension is not fixed, and itself subject to inference. Examples range from classical statistical tasks such as variable selection, mixture estimation, change-point analysis, and model determination in general, through to the kinds of problem raised in modern applications of stochastic modelling to gene-mapping, analysis of ion channel data, image segmentation and object recognition.

For a fully Bayesian analysis based on a single simulation run, we need a MCMC sampler that jumps between parameter subspaces of differing dimensions: given the reversible jump framework of the previous subsection, this is now a fairly modest generalisation. Our state variable $\boldsymbol{x}$ now lives in a union of spaces of differing dimension: $\mathcal{X} = \cup_k \mathcal{X}_k$.

We will use a range of *move types* m, each providing a transition kernel P_m, and insist on detailed balance for each:

$$\int_{\boldsymbol{x} \in A} \pi(d\boldsymbol{x}) P_m(\boldsymbol{x}, B) = \int_{\boldsymbol{y} \in B} \pi(d\boldsymbol{y}) P_m(\boldsymbol{y}, A)$$

for all Borel sets $A, B \subset \mathcal{X}$. The idea of a family of move types is implicit even in the simplest formulation of Metropolis-Hastings, where we have a different proposal density q_i for each component i, but compute the acceptance probability using the joint target distribution (equation (1.7)). In the present more elaborate context, there may be a richer variety of move types, recognising that different approaches may be needed to enable transitions between different pairs of spaces $\mathcal{X}_k, \mathcal{X}_{k'}$.

The Metropolis-Hastings idea still works, but you need to work a bit harder to make the acceptance ratio make sense. The proposal measure q is now the *joint* distribution of move type m and proposed destination $\boldsymbol{y}$, so for each $\boldsymbol{x} \in \mathcal{X}$, $\sum_m \int_{\boldsymbol{y} \in \mathcal{X}} q_m(\boldsymbol{x}, d\boldsymbol{y}) \leq 1$ (allowing a positive probability of not attempting a move, if required). The detailed balance condition (see (1.10)) becomes

$$\int_{(\boldsymbol{x},\boldsymbol{y}) \in A \times B} \pi(d\boldsymbol{x}) \alpha_m(\boldsymbol{x}, \boldsymbol{y}) q_m(\boldsymbol{x}, d\boldsymbol{y}) = \int_{(\boldsymbol{x},\boldsymbol{y}) \in A \times B} \pi(d\boldsymbol{y}) \alpha_m(\boldsymbol{y}, \boldsymbol{x}) q_m(\boldsymbol{y}, d\boldsymbol{x}).$$

for all m, A, B. This leads to the formal solution

$$\alpha_m(\boldsymbol{x}, \boldsymbol{y}) = \min \left\{ 1, \frac{\pi(d\boldsymbol{y}) q_m(\boldsymbol{y}, d\boldsymbol{x})}{\pi(d\boldsymbol{x}) q_m(\boldsymbol{x}, d\boldsymbol{y})} \right\}$$

as in (1.12).

Apart from the addition of the subscript m, this is just a special case of the earlier general Metropolis-Hastings method, and the ratio of measures makes sense subject to the existence of a symmetric dominating measure μ_m for $\pi(d\boldsymbol{y})q_m(\boldsymbol{y}, d\boldsymbol{x})$.

Again, this is most easily understood in the concrete terms of the preceding subsection: we need a differentiable bijection between $(\boldsymbol{x}, \boldsymbol{u})$ and $(\boldsymbol{y}, \boldsymbol{u}')$, where $\boldsymbol{u}, \boldsymbol{u}'$ are the vectors of random numbers used to go between $\boldsymbol{x}$ and $\boldsymbol{y}$ in each direction. Suppose these have densities $g_m(\boldsymbol{u}; \boldsymbol{x})$ and $g_m(\boldsymbol{u}'; \boldsymbol{y})$. In the variable dimension context, move type m might use transitions between $\mathcal{X}_{k_1}$ and $\mathcal{X}_{k_2}$; if these spaces have dimensions d_1 and d_2, and π is absolutely continuous with respect to ν_{d_1} and ν_{d_2} in the respective spaces, then the dimensions of $\boldsymbol{u}$ and $\boldsymbol{u}'$, r_1 and r_2 say, must satisfy the dimension-balancing condition

$$d_1 + r_1 = d_2 + r_2.$$

We can then write

$$\alpha_m(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})g_m(\boldsymbol{u}'; \boldsymbol{y})}{\pi(\boldsymbol{x})g_m(\boldsymbol{u}; \boldsymbol{x})}\left|\frac{\partial(\boldsymbol{y}, \boldsymbol{u}')}{\partial(\boldsymbol{x}, \boldsymbol{u})}\right|, \right\} \tag{1.14}$$

The ratio is of joint densities with the same degrees of freedom, together with the Jacobian needed to account for the change of variable.

Apart from the illustrative applications in Green (1995), this methodology has been widely implemented for problems with a variable-dimension parameter, for example Richardson and Green (1997), Uimari and Hoeschele (1997), Denison, Mallick and Smith (1998), Heikkinen and Arjas (1998), Holmes and Mallick (1998), Pievatolo and Green (1998), Green and Richardson (1999), Hodgson (1999), Hodgson and Green (1999), and Rue and Hurn (1999).

1.7.3 Example: step functions

Let us illustrate the methods of the preceding subsection in what is almost the simplest setting possible, by studying the situation where the state variable $\boldsymbol{x}$ represents a step function, as might be part of the parameterisation of a model for change-point analysis in regression or point processes. We can readily evaluate each of the factors in (1.14), and end up with a useful sampler that — with a little modification — will find application in the next subsection.

A simple prior model for a step function on $[0, L)$ would parameterise the function in terms of its number of steps k, the positions $\{s_1 < s_2 < \cdots < s_k\}$ of those steps, and the heights $\{h_0, h_1, \ldots, h_k\}$, h_j being the value of the function on the interval $[s_{j-1}, s_j)$. For illustration here, we assume that the number of steps is drawn from an arbitrary $p(k)$, and that given k, the step heights are i.i.d. from some density $f_H(\cdot)$, and that the step positions are drawn as the order statistics from a uniform distribution on

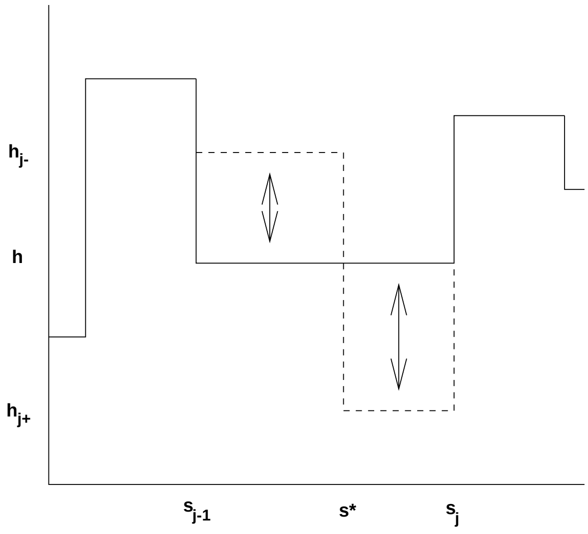

Figure 1.15 *Split and merge of a step.*

the observation interval:

$$p(s_1, s_2, \ldots, s_k) = \frac{k!\, I\,[0 < s_1 < s_2 < \cdots < s_k < L]}{L^k}.$$

As with ordinary Metropolis-Hastings, you have freedom to use intuition in designing proposals; validity is ensured by using the correct acceptance probability (1.14).

Consider a move which allows the number of steps to change, by 'birth' and 'death'. When $\boldsymbol{x}$ has k steps, we propose birth with probability b_k, draw two random numbers u_1 and u_2 from $g(u_1, u_2)$, and use them to split an existing step interval into two. Let the new step position be $s^* = u_1$, located between s_{j^*-1} and s_{j^*} say, and use u_2 to divide the current step height h_{j^*} into two values with weighted average h_{j^*}: $h_{j-} = h_{j^*} + u_2/w_-$ and $h_{j+} = h_{j^*} - u_2/w_+$, where $w_- = s^* - s_{j^*-1}$ and $w_+ = s_{j^*} - s^*$. Turning now to death of a step: this is proposed with probability d_k, and we choose a step at random to delete; if step $j^\dagger$ is deleted, then the new step height for the interval $[s_{j^\dagger-1}, s_{j^\dagger+1})$ is the weighted average $\{(s_{j^\dagger} - s_{j^\dagger-1})h_{j^\dagger} + (s_{j^\dagger+1} - s_{j^\dagger})h_{j^\dagger+1}\}/(s_{j^\dagger+1} - s_{j^\dagger-1})$. This precisely reverses the effect of the birth just described.

Note that this formulation has the dimension balance we require: when there are k steps, there are k positions and $k+1$ heights, making $d_1 = 2k+1$ variables in all. With a birth, we are proposing a move to $d_2 = 2k+3$. The dimensions of the random numbers $\boldsymbol{u}, \boldsymbol{u}'$ are $r_1 = 2$ and $r_2 = 0$ respectively, and indeed $d_1 + r_1 = d_2 + r_2$. (An alternative choice, equally valid, would be

$r_1 = 3, r_2 = 1$, which would be obtained if we had dropped the requirement of preserving the weighted average on birth and death, and generating new random heights as needed, for example independently of the current state.)

Suppose we let $\boldsymbol{x}, \boldsymbol{y}$ denote the states of the chain before the split is proposed, and the state as modified by the birth proposal. Then

$$\begin{aligned}
\pi(\boldsymbol{x}) &\propto p(Y|\boldsymbol{x})p(k)\prod_{j=0}^{k} f_H(h_j)\frac{k!\, I\,[0<s_1<s_2<\cdots<s_k<L]}{L^k} \\
\pi(\boldsymbol{y}) &\propto p(Y|\boldsymbol{y})p(k+1)\prod_{j\neq j^*} f_H(h_j)f_H(h_{j-})f_H(h_{j+}) \\
&\quad \frac{(k+1)!\, I\,[0<s_1<s_2<\cdots<s_{j^*-1}<s^*<s_{j^*}<\cdots<s_k<L]}{L^{k+1}},
\end{aligned}$$

the constant of proportionality being the same in each case. The proposal terms are

$$\begin{aligned}
g_m(\boldsymbol{u};\boldsymbol{x}) &= b_k g(u_1,u_2) \\
g_m(\boldsymbol{u}';\boldsymbol{y}) &= \frac{d_{k+1}}{k+1},
\end{aligned}$$

reflecting the described mechanism for choosing to propose birth or death, the drawing of (u_1, u_2), and the random choice of a step to delete. Finally, the Jacobian we need is an order $2k+3$ determinant, but with many of the components of the state vector unaltered by the transformation, it reduces to

$$\left|\frac{\partial(h_{j-}, h_{j+}, s^*)}{\partial(h_{j^*}, u_1, u_2)}\right| = \frac{w_- + w_+}{w_- w_+}.$$

We can now compute the acceptance probability for a birth from (1.14):

$$\begin{aligned}
\alpha &= \min\left\{1, \Lambda\frac{p(k+1)}{p(k)}\frac{(k+1)}{L}\frac{f_H(h_{j-})f_H(h_{j+})}{f_H(h_{j^*})}\right. \\
&\qquad \left.\frac{d_{k+1}/(k+1)}{b_k g(u_1,u_2)}\frac{w_- + w_+}{w_- w_+}\right\}
\end{aligned}$$

where Λ is the likelihood ratio $p(Y|\boldsymbol{y})/p(Y|\boldsymbol{x})$.

1.7.4 Cyclones example, continued

Model 8: unknown number of change points

What if the *number* of change points, k, is also unknown? We might place a prior on k, say Poisson(λ):

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

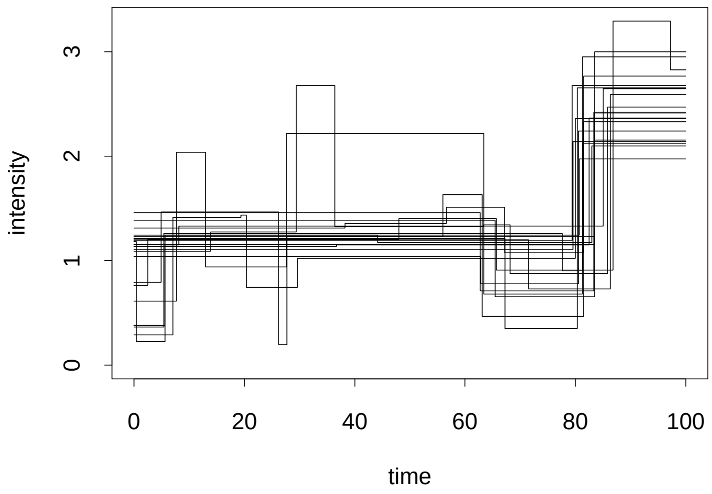

Figure 1.16 *Posterior sample of step functions $x(t)$ for model 8, applied to cyclones data.*

and then make Bayesian inference about all unknowns: $\boldsymbol{x} = (k, \alpha, \beta, T_1, \ldots, T_k, x_0, \ldots, x_k)$. There are $2k+4$ parameters: the number of things you don't know is one of the things you don't know!

For a MCMC solution, the only additional ingredient we need over model 7 is a birth/death move to allow a variable number of steps. This follows closely the setup of the preceding subsection, except that since the step function represents an intensity and is necessarily non-negative, we arranged to preserve the weighted *geometric* mean:

$$h_{j-}^{w_-} h_{j+}^{w_+} = h_{j^*}^{w_- + w_+}.$$

Also the joint density of the step positions is now $\propto \prod_j (s_j - s_{j-1})$ with a corresponding change to the acceptance ratio. The moves for this sampler are described in detail in Green (1995), except that here we have slightly extended the model to include variable hyperparameters α and β, as seen in Sections 1.2.4 and 1.4.11.

We applied this model to the cyclones data, using $\lambda = 3$; a small sample from the posterior distribution of the step function $x(\cdot)$ is shown in Figure 1.16. Various aspects of the posterior distribution can be summarised by appropriate analysis and display of much larger MCMC samples; see, for example, Figures 1.17, 1.18 and 1.19.

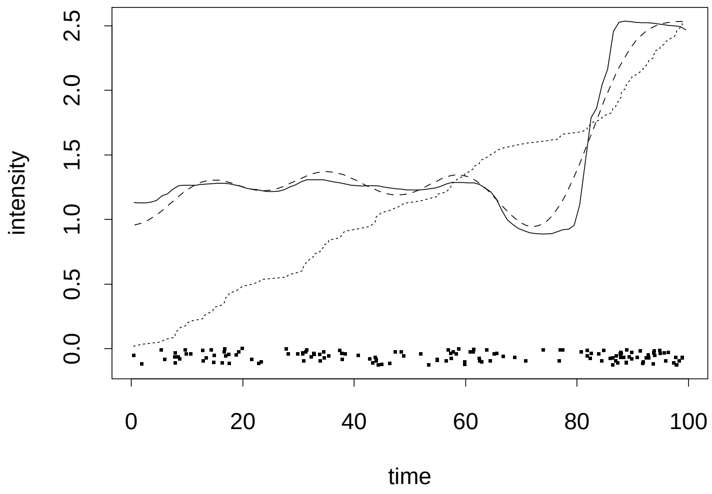

Figure 1.17 *Posterior mean of step function $x(t)$ for model 8 (solid line) and kernel estimate of $x(t)$ (broken line), for cyclones data.*

Model 9: with a cyclic component

Finally, here, as further illustration of the flexibility in modelling allowed by the approach, we include another ingredient, that will be justified in many real time-series point process problems: periodicity.

This could be handled in various ways — parametric, nonparametric, with known and unknown period(s) — but the simplest is to take a simple sinusoid, and assume that the data are generated from a Poisson process with instantaneous rate

$$x(t)\left\{1+\gamma\cos(2\pi ft)+\delta\sin(2\pi ft)\right\},$$

where $x(t)$ is the step function defined above, and f denotes the assumed (known) frequency).

If *a priori* (γ,δ) are taken uniform on the unit disc, then a simultaneous Metropolis update is easily implemented.

A small sample from the posterior for the cyclic component is shown in Figure 1.20.

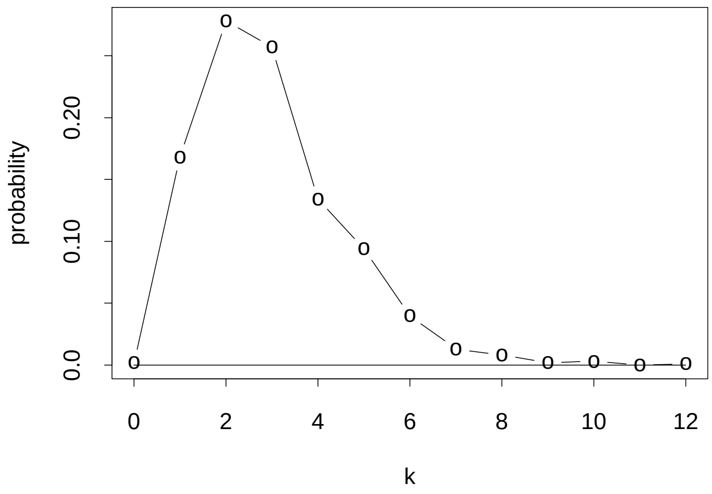

Figure 1.18 *Posterior distribution of number of change points k for model 8, applied to cyclones data.*

1.7.5 Bayesian model determination

It is wrong to behave as if the statistical model for our data was not subject to question.

Suppose we have a (countable) collection of models that we wish to entertain: $M_1, M_2, \ldots, M_k, \ldots$. *A priori*, we assign probabilities to these: $p(k)$.

For each model, there is a parameter vector $\theta = \theta_k \in \mathcal{R}^{n_k}$ say, with a prior: $p(\theta_k|k)$, and a likelihood for the observed data Y: $p(Y|k, \theta_k)$. The joint distribution of all variables is

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

(Note that in this section, the subscript k on θ indicates the model to which θ_k belongs, not the k^{th} element of a vector θ.)

Observing Y provides information about both the model indicator k and the corresponding parameter vector θ_k, through their posterior distributions:

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

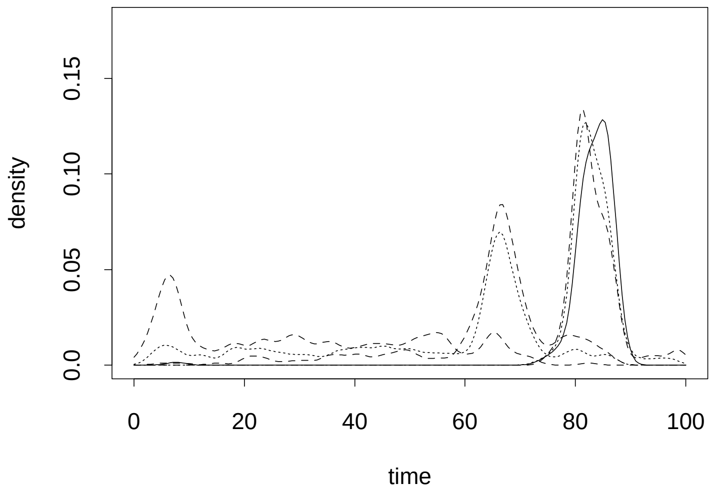

Figure 1.19 *Estimates of posterior density for change point positions for model 8, applied to cyclones data: $k = 1$ (solid line), $k = 2$ (dotted lines) and $k = 3$ (broken lines).*

and

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

involving integrals that as usual seem to need MCMC! There are two main approaches: within-model and across-model simulation.

Within-model simulation

Here we treat each model M_k separately.

The posterior for the parameters θ_k is in any case a within-model notion:

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

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 *Bayes factor* for model M_{k_1} vs. M_{k_0}), it is sufficient to estimate the

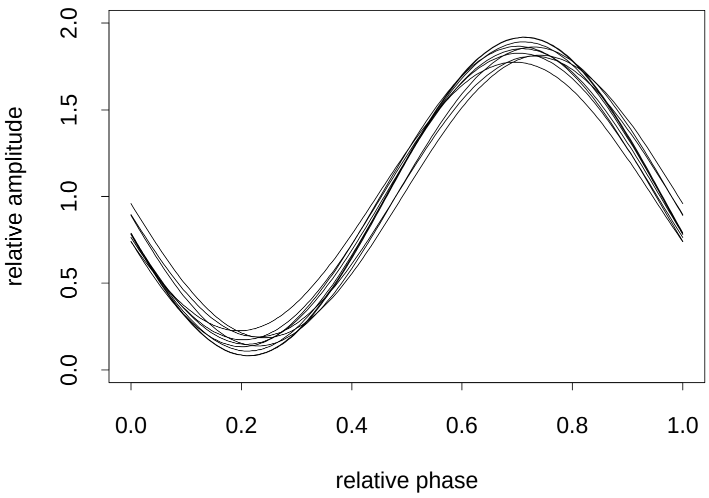

Figure 1.20 *Posterior samples from the harmonic component, model 9, applied to cyclones data.*

marginal likelihoods

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

separately for each k, using individual MCMC runs.

Estimating the marginal likelihood

There are many 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).

Across-model simulation

Here we conduct a *single* simulation that traverses the entire (k, θ_k) space. Since the dimension n_k of θ_k typically varies with k, this requires a MCMC sampler that works in more general spaces than $\mathcal{R}^d$. The reversible jump sampler of Section 1.7.2 is an obvious candidate. Applications of this approach include Giudici and Green (1999), and Nobile and Green (2000).

See also Madigan and Raftery (1994), Carlin and Chib (1995), Phillips and Smith (1996) and George and McCulloch (1997) for other recent approaches to Bayesian computation for model determination.

1.8 Some tools for improving performance

1.8.1 Tuning a MCMC simulation

Having implemented an MCMC sampler, there are various quite simple techniques available to amend the algorithm to try to improve performance.

Most Metropolis-Hastings methods involve proposal distributions with freely chosen parameters – the spread of the perturbation distribution in a random-walk Metropolis method, for example. As the parameter is varied, different acceptance rates will be obtained. Of course, 100% acceptance is not necessarily desirable; in random-walk Metropolis it is only achieved in the limit as the spread goes to 0. As this is approached, the sample path will exhibit increasingly high autocorrelation. In the 'bold' opposite to this 'timid' strategy, the steps taken will be large, but they will be taken rarely. The right balance, where convergence may be faster and autocorrelation less, will be in the middle.

There is an interesting theoretical study of optimal acceptance rates for random walk Metropolis in Gelman, Roberts and Gilks (1996), based on the artificial case of multivariate normal target and proposal distributions; this has been quite influential in establishing a 'rule of thumb' advising aiming for 20-40% acceptance generally, but this study is perhaps a rather narrow basis for such a sweeping conclusion.

Metropolis-Hastings methods can be designed for updating single variables, or groups of any size. Larger groups offer the possibility of allowing the sampler to beat the restrictions on performance imposed by high correlations between variables in the target distribution, but may carry a burden in cumbersome tuning of a multivariate proposal distribution. A careful study of the merits of grouping variables is one of the themes of Roberts and Sahu (1997); see also Besag et al (1995, p. 10).

Another opportunity for reducing correlation between variables is to consider re-parameterisation; a hierarchical centring formalism in the linear mixed models context is introduced by Gelfand, Sahu and Carlin (1995), and there is a broader discussion in Gilks and Roberts (1996).

1.8.2 *Antithetic variables and over-relaxation*

The essential idea of antithetic variables in ordinary static Monte Carlo is one of the classic ideas for variance reduction: to aim to introduce negative correlation among some of the summands in an empirical average $N^{-1}\sum_1^N g(\boldsymbol{x}^{(t)})$ by using coupled pairs $(u, 1-u)$ of uniform random numbers in generating pairs of state vectors $\boldsymbol{x}$. Of course, the success of the method requires some monotonicity in both the mapping from u to $\boldsymbol{x}$, and in the function g.

As applied to MCMC, the aim would be to choose an update of $\boldsymbol{x}^{(t)}$ that has detailed balance, as usual, but also introduces negative *serial* autocorrelation in the process $g(\boldsymbol{x}^{(t)})$, or at least reduces the value of a positive autocorrelation.

Barone and Frigessi (1989) studied the effect of antithetic variables on the convergence of samplers for Gaussian processes. The full conditional for a single variable x_i in a multivariate Gaussian distribution is of course a normal distribution, $N(\mu_i, \sigma_i^2)$, say. It is easy to check that drawing the updated variable x_i' from $N((1+\theta)\mu_i - \theta x_i, (1-\theta^2)\sigma_i^2)$ is in detailed balance for any $\theta \in (-1, 1)$; $\theta = 0$ gives the Gibbs sampler, and if $\theta > 0$ then x_i and x_i' are conditionally negatively correlated. They show that in the case of entirely positive association between the variables, the spectral radius of the corresponding Markov chain is a decreasing function of θ at $\theta = 0$; thus convergence is improved by using the dynamic version of antithetic variables, $\theta > 0$.

Green and Han (1992) (see also Besag and Green, 1993) examine the effect of this antithetic modification on the autocorrelation time, and show that it is reduced by a factor $(1-\theta)/(1+\theta)$. They also propose using antithetically-modified Gaussian approximations to full conditionals as proposal distributions for Metropolis-Hastings for non-Gaussian targets, although the empirical evidence assessing this idea suggests that convergence is not always improved. Barone, Sebastiani and Stander (1998) have developed the idea further. Neal (1998) has proposed a related method, based on order statistics, that seems much more widely applicable.

This whole topic has close parallels with the theory of over-relaxation in the iterative solution of simultaneous equations in numerical analysis.

1.8.3 *Augmenting the state space*

Perhaps counter-intuitively, it is sometimes possible to improve MCMC performance by augmenting the state vector to include additional components. Two particularly successful recipes are those in which the original model appears as a *conditional* distribution in an augmented model (simulated tempering) and in which it appears as a *marginal* (auxiliary variables); these approaches are described in the next two subsections.

Two other devices might also be bracketed under the heading of augmentation. In multigrid methods, spatial problems are treated on a variety of spatial scales, sometimes by coupling together several different models, sometimes merely by using a family of MCMC samplers that update groups of variables together, the sizes of the groups varying with sweep. In hybrid MCMC, additional variables are introduced, bearing a relationship to the original ones analogous to that between momentum and position variables in dynamics. The MCMC updates maintain this physical analogy.

1.8.4 Simulated tempering

The approach here is to combat slow mixing by embedding the desired model in a family of models, indexed say by α, and treat α now as an additional dynamic variable. Thus the target is changed from $\pi(\boldsymbol{x})$ to $\pi^\star(\boldsymbol{x}, \alpha_0)$. The family $\{\pi^\star(\boldsymbol{x}, \alpha)\}$ is designed so that for some α, a much better-mixing chain can be found than for the original target. We run MCMC on $\pi^\star(\boldsymbol{x}, \alpha)$, and condition on $\alpha = \alpha_0$ by selecting from the output.

This 'serial' approach can be compared with the 'parallel' one of Metropolis-coupled MCMC (Geyer, 1991; see also Gilks and Roberts, 1996).

Simulated tempering, by changing the temperature

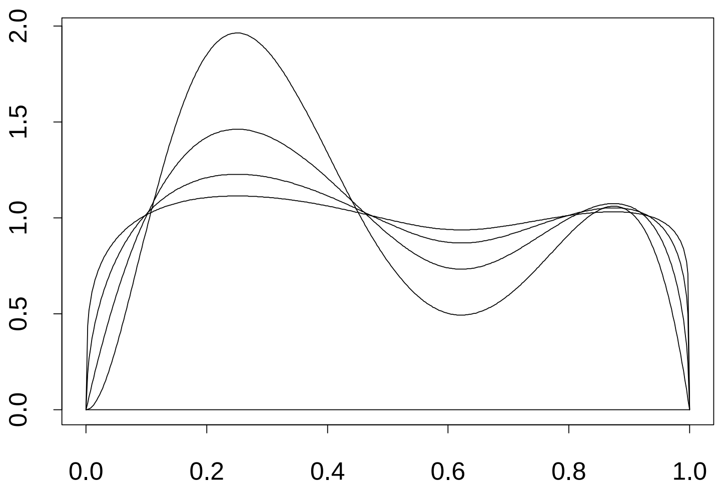

Figure 1.21 *The effect of tempering on a univariate full conditional: the beta mixture density* $0.7Be(3,7) + 0.3Be(8,2)$*, and the results after raising to the powers* $\alpha = 0.5, 0.25, 0.125$ *and renormalising.*

This was the original idea of Marinari and Parisi (1992), independently derived by Geyer and Thompson (1995); we set

$$\pi^\star(\theta, \alpha) \propto \{\pi(\theta)\}^\alpha$$

where $\alpha = \alpha_0 = 1$ corresponds to the original model, and $\alpha \to 0$ makes the probability surface 'flatter', or in physical terms, 'warmer'. A graphical illustration of the effect of the α power on a univariate density can be seen in Figure 1.21.

The full conditionals change in the same way as the joint distribution:

$$\pi^\star(\theta_i | \theta_{-i}, \alpha) \propto \{\pi(\theta_i | \theta_{-i})\}^\alpha$$

so implementation is very easy.

We normally place a (discrete) artificial prior on α so that the *marginal* for α is approximately uniform.

Simulated tempering, by inventing models

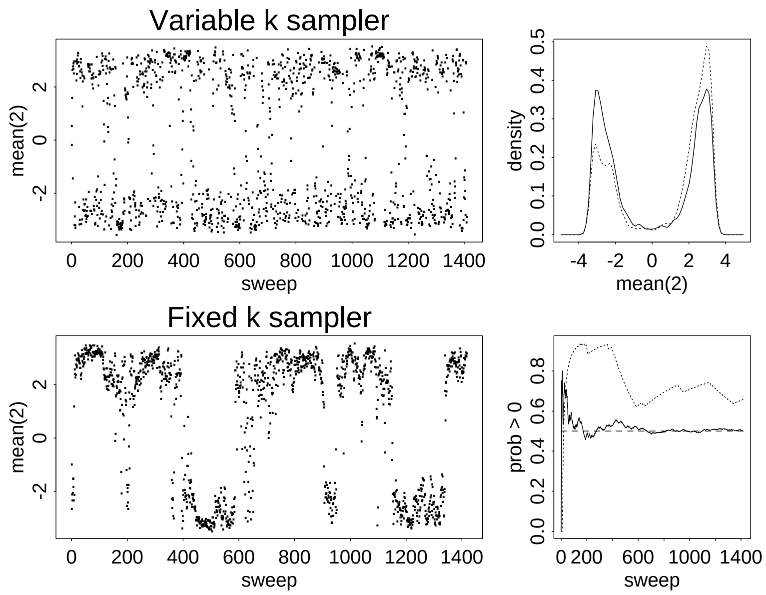

Figure 1.22 *Better mixing with variable dimensions, illustrated by a mixture analysis application (from Richardson and Green (1997)).*

A more general perspective on what tempering achieves and how it works can be obtained by envisaging it as embedding the target into a bigger model space, and there may be many ways to do that. For example, a model indicator k may be allowed to vary, although in truth its value is known, or at least fixed. An example from mixture analysis is shown in Figure 1.22. The left hand panels show the sample paths for one component of the parameter vector, which has a strongly bimodal distribution under the target; two samplers are compared, one (bottom) in which the model indicator k (in this case the number of mixture components) is held fixed, the other (top) in which it varies but we condition on its value by selecting from the output. In the right hand side panels we see (top) the resulting estimates of the marginal density of this parameter and (bottom) the evolution of the ergodic average estimating the probability that the parameter is positive; from the symmetry of the setup of the experiment, this is known to be 0.5. Results for the variable-k sampler are shown in solid lines, those for fixed k are dotted. Allowing the number of components to vary can give much better mixing. See Richardson and Green (1997) for details.

1.8.5 Auxiliary variables

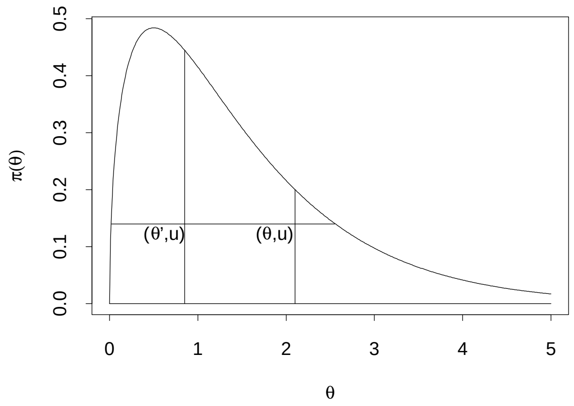

Figure 1.23 *The slice sampler.*

Edwards and Sokal (1988) proposed a way to improve mixing by augmenting the state space so that the original target appears as the *marginal* equilibrium distribution. The following interpretation of their approach in statistical language can be found in Besag and Green (1993).

Starting from $\pi(\boldsymbol{x})$, introduce some additional variables u, with $\pi(u|\boldsymbol{x})$ arbitrarily chosen. Then the joint is $\pi(\boldsymbol{x}, u) = \pi(\boldsymbol{x})\pi(u|\boldsymbol{x})$, for which $\pi(\boldsymbol{x})$ is certainly the marginal for $\boldsymbol{x}$.

We could now run a MCMC method for the *joint* target $\pi(\boldsymbol{x}, u)$ (usually a method that updates $\boldsymbol{x}$ and u alternately), and simply ignore the u variables in extracting information from the simulation.

When might this idea be useful? Suppose $\pi(\boldsymbol{x})$ factorises as:

$$\pi(\boldsymbol{x}) = \pi_0(\boldsymbol{x})b(\boldsymbol{x})$$

where $\pi_0(\boldsymbol{x})$ is a (possibly unnormalised) distribution that is easy to simulate from, and $b(\boldsymbol{x})$ is the awkward part, often representing the 'interactions' between variables that are slowing down the chain.

Then take a one-dimensional u with $u|\boldsymbol{x} \sim U[0, b(\boldsymbol{x})]$: we find

$$\pi(\boldsymbol{x}, u) = \pi(\boldsymbol{x})\pi(u|\boldsymbol{x}) = \pi_0(\boldsymbol{x})b(\boldsymbol{x})\frac{I[0 \leq u \leq b(\boldsymbol{x})]}{b(\boldsymbol{x})}$$

so that

$$\pi(\boldsymbol{x}|u) \propto \pi_0(\boldsymbol{x})$$

restricted to (conditional on) the event $\{\boldsymbol{x} : b(\boldsymbol{x}) \geq u\}$. At least when this $\pi(\boldsymbol{x}|u)$ can be sampled without rejection, we can easily implement a Gibbs sampler, drawing u and $\boldsymbol{x}$ in turn.

This method has recently been popularised under the name of the 'slice sampler', a picturesque but otherwise unnecessary name, reflecting the fact that if $\pi_0(\boldsymbol{x}) =$ constant, $\pi(\boldsymbol{x}|u)$ is a uniform distribution, corresponding to a horizontal slice through the graph of $\pi(\boldsymbol{x})$. For statistical applications of the idea, see Neal (1997) and Damien, Wakefield and Walker (1999), and for a detailed analysis of the method, see Roberts and Rosenthal (1999).

The original applications of auxiliary variable methods were to statistical physics problems, where in particular the Swendsen-Wang method (Swendsen and Wang, 1987) has had a profound influence; see also Edwards and Sokal (1988) and Sokal (1989).

The Swendsen-Wang method is a MCMC method for the Potts model on an arbitrary graph (V, E), the target distribution

$$\pi(\boldsymbol{x}) \propto \exp\left\{-\beta \sum_{(v,w)\in E} I[\boldsymbol{x}_v \neq \boldsymbol{x}_w]\right\} = \prod_{e\in E} b_e(\boldsymbol{x}),$$

say. We define one auxiliary variable u_e for each edge e, conditionally independent given $\boldsymbol{x}$, with $u_e|\boldsymbol{x} \sim U(0, b_e(\boldsymbol{x}))$. If $u_e > e^{-\beta}$ we say the edge e is 'on', otherwise 'off'. It is easy to see that in drawing $\boldsymbol{u}$ given $\boldsymbol{x}$, edges are on with probability $1 - e^{-\beta}$ if $\boldsymbol{x}_v = \boldsymbol{x}_w$, always off if $\boldsymbol{x}_v \neq \boldsymbol{x}_w$. Simple manipulation shows that $\pi(\boldsymbol{x}|\boldsymbol{u})$ is a random uniform colouring on the clusters determined by the on bonds.

Figure 1.24 illustrates one sweep of the Swendsen-Wang algorithm, applied to the Potts model on a small graph.

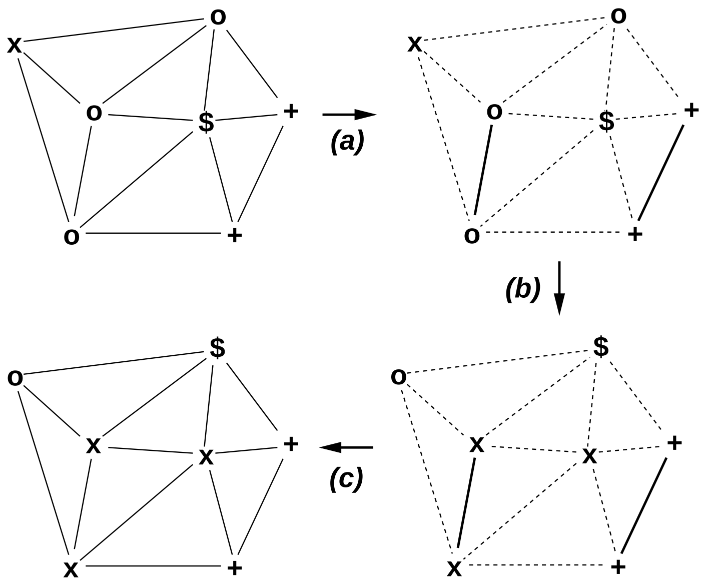

Figure 1.24 *Illustrating the Swendsen-Wang algorithm: (a) bond variables between like-coloured nodes are 'on' with probability* $1 - e^{-\beta}$*, always 'off' between unlike-coloured ones; (b) clusters formed by 'on' bonds are re-coloured uniformly at random; (c) the new colouring.*

1.9 Coupling from the Past (CFTP)

Coupling from the Past (CFTP) is a beautiful idea due to Propp and Wilson (1996): it provides a way of organising a Markov chain simulation so that after a finite but random amount of work, it *exactly* delivers a sample from the target distribution! (Another such protocol, based on an elaborate form of rejection sampling was given by Fill (1998).)

Since the CFTP idea first appeared in preprint form, it has generated much excitement among MCMC researchers, keen both to understand and generalise the basic formulation, and to discover the practical potential for computation in stochastic processes and statistical applications.

For an example of Propp and Wilson's construction, consider the partial simulation of a symmetric random walk with reflecting barriers shown in Figure 1.25. To appreciate the message of this figure, it is not necessary to know anything about the order in which the displayed steps were generated, nor anything at all about any steps not displayed. All that we need is that the successive steps along each partially-drawn path are independent, and have the correct law: equally probably ± 1, except where steps attempting to go outside the interval $[1, 5]$ are suppressed.

One can see that *for the random numbers used to make this simulation*,

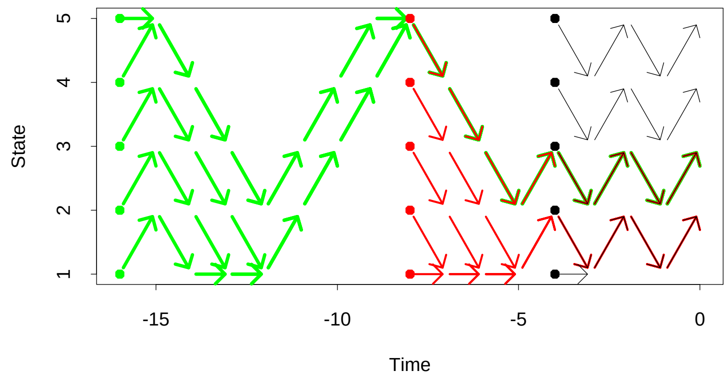

Figure 1.25 *Monotone CFTP for a simple random walk.*

and regarding the figure as part of a conceptual simulation of paths from *all* initial states at *all* initial times < 0, all paths of the chain starting at time $-\infty$ have the same state (*viz.*, 3) at time 0. This state, $\boldsymbol{x}^{(0)}$, must be drawn from π!

Generally, imagine multiple coupled paths of a Markov chain run from all initial states in the indefinite past, and look at the state at time 0, $\boldsymbol{x}^{(0)}$. If this is unique, then $\boldsymbol{x}^{(0)} \sim \pi$. For this to be of any practical consequence in computing, we must be able to conduct this conceptually infinite amount of simulation in a finite time. But, we can that Figure 1.25 was indeed constructed with a finite amount of work — fewer than 60 steps are shown.

Generalising from this example, if there exists a (random) initial time $-T$ such that for all initial states $\boldsymbol{x}_{-T}$, $\boldsymbol{x}^{(0)}$ is the same, then $\boldsymbol{x}^{(0)} \sim \pi$. We do not even need to find T exactly, since coalescence occurs from all initial times $< -T$. So we can just try a decreasing sequence of initial times $-1, -2, -4, -8, \ldots$ until we discover coalescence.

1.9.1 Is CFTP of any use in statistics?

There have been some spectacular successes in finding CFTP implementations for certain models in statistical physics and spatial processes possessing a lot of symmetry, even with huge numbers of variables (4 million in one case).

But it seems much harder to make it work for even quite low-dimensional continuous distributions without symmetry.

1.9.2 The Rejection Coupler

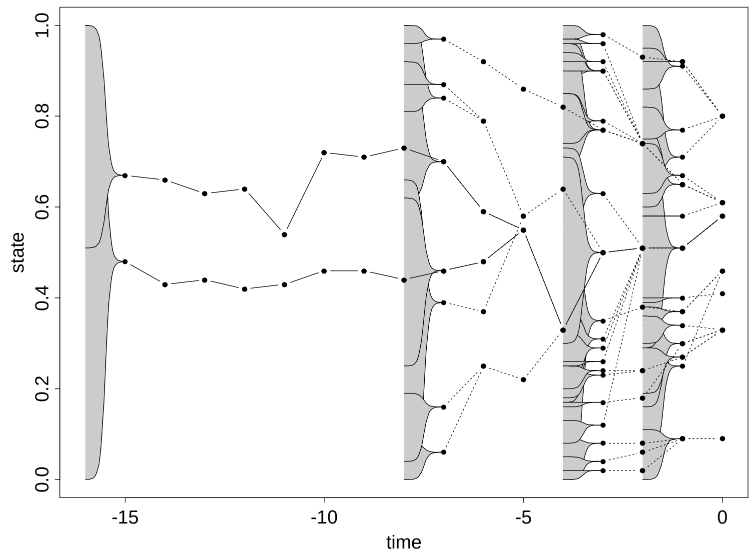

Figure 1.26 *Example realisation of rejection coupler for* $f(y|x) = (\delta + 1)\min(\{y/x\}^{\delta}, \{(1-y)/(1-x)\}^{\delta})$ *with* $\delta = 6$.

Here is a simple approach to CFTP for a continuous state space, namely the unit interval, from Murdoch and Green (1998). It is more of a 'proof of existence' (of a CFTP method in a continuous state space) than a practical method, for we have to suppose we know $f(y|x)$, where

$$P\{X_{t+1} \leq y | X_t = x\} \propto \int_{-\infty}^{y} f(u|x)du,$$

and that the (not necessarily normalised) densities $f(y|x)$ are bounded above by an integrable $h(y)$. We cannot expect the transition density to be available for a practically-useful MCMC method.

Recall the familiar rejection sampler, expressed in geometrical terms. To sample from the (not necessarily normalised) $f(\cdot|x)$, we repeatedly draw (Y, Z) uniformly under the graph of h until $Z < f(Y|x)$. The rejection coupler generalises this scheme. To sample from $f(\cdot|x)$ for all x, again we repeatedly draw (Y_i, Z_i) uniformly under the graph of h. Let $A_i = \{x : Z_i < f(Y_i|x)\}$; then Y_i is a valid update for all $x \in A_i$. We continue until $\cup_i A_i = \chi$, obtaining a random-length list $\{Y_i\}$.

When incorporated into the CFTP protocol, this procedure gives partial realisations of a continuum of coupled paths exemplified by the simulation shown in Figure 1.26. This shows a single realisation of CFTP using rejection coupling, for the kernel density $f(y|x) = (1+\delta)\min(\{y/x\}^\delta, \{(1-y)/(1-x)\}^\delta)$ with $\delta = 6$, which is bounded above by the envelope function $h(y) = 1+\delta$ for $0 < y < 1$. The solid lines indicate the paths ultimately followed by all realisations starting from the indefinite past.

1.9.3 Towards generic methods for Bayesian statistics

In contrast to the rejection coupler, a practical technique for Bayesian CFTP should be based only on the target distribution, and created by some generic recipe, just as is the case for standard MCMC.

Evidence that this will become possible is still quite unconvincing, although this is an extremely active research area, and success may be obtained soon. Some experiments in this direction are the random walk Metropolis coupler of Green and Murdoch (1998), the methods using Gibbs sampling and (anti-) monotonicity of Møller (1999), perfect slice sampling (Mira, Møller and Roberts, 1999) and the perfect simulated tempering approach of Møller and Nicholls (1999).

One general reason for pessimism about the future of CFTP in Bayesian statistics is found by noting that much of the success of ordinary MCMC in this field is based on its modularity: as a model is elaborated, the parameter vector is augmented, and the current sampler is supplemented by new moves for the new components. Existing methods for perfect simulation are not modular.

1.10 Miscellaneous topics

1.10.1 Diffusion methods

A number of MCMC methods have been developed recently, inspired by the Langevin stochastic differential equation

$$d\boldsymbol{x}_t = d\boldsymbol{B}_t + \frac{1}{2}\nabla \log \pi(\boldsymbol{x}_t) dt$$

where $\boldsymbol{B}_t$ denotes Brownian motion on $\mathcal{X}$. Here we describe only the case where $\mathcal{X} = \mathcal{R}$. This diffusion has invariant distribution π, and suggests use of the discrete-time chain

$$\boldsymbol{x}^{(t+\varepsilon)}|\boldsymbol{x}^{(\leq t)} \sim N\left(\boldsymbol{x}^{(t)} + \frac{1}{2}\varepsilon\nabla \log \pi(\boldsymbol{x}^{(t)}), \varepsilon\right), \tag{1.15}$$

where the time increment is ε, not 1.

Unfortunately, this simple discretisation is too crude: not only does it only, at best, deliver an approximation to π as its invariant distribution, it can actually create a transient chain!

However, that can be fixed by using the *Metropolis-adjusted* Langevin algorithm (Besag, 1994), in which (1.15) is used simply as a proposal distribution, with acceptance determined as usual by (1.7).

Among examples of practical methodology using diffusion-based MCMC are the jump-diffusion methods of Grenander and Miller (1994), combining (unadjusted) Langevin diffusion with dimension-jumping moves to address variable-dimension problems, and the work of Phillips and Smith (1996) applying this approach in various statistical settings.

The Metropolis-adjusted Langevin method is known to fail to be geometrically ergodic for heavy-tailed targets, a problem addressed by the richer class of *'self-targetting'* Metropolis-adjusted Langevin algorithms due to Stramer and Tweedie (1998), in which the proposal distribution is

$$\boldsymbol{x}^{(t+\varepsilon)}|\boldsymbol{x}^{(\leq t)} \sim N\left(\boldsymbol{x}^{(t)} + \varepsilon\mu(\boldsymbol{x}^{(t)}), \varepsilon\sigma^2(\boldsymbol{x}^{(t)})\right)$$

where

$$\mu(\boldsymbol{x}) = \frac{1}{2}\sigma^2(\boldsymbol{x})\nabla \log \pi(\boldsymbol{x}) + \sigma(\boldsymbol{x})\nabla\sigma(\boldsymbol{x}).$$

This is derived from the diffusion generated by

$$d\boldsymbol{x}_t = \sigma(\boldsymbol{x}_t)d\boldsymbol{B}_t + \mu(\boldsymbol{x}_t)dt.$$

Stramer and Tweedie (1998) discuss the extent to which their theory for these methods can be extended to the practically important cases where $X = \mathcal{R}^d, d > 1$.

1.10.2 Sensitivity analysis via MCMC

In responsible Bayesian inference, it is important to assess the effect on the posterior of changes to the model, especially variations in the prior. Suppose that, having completed a MCMC-based analysis using a prior $\pi_0(\boldsymbol{\theta})$ and likelihood $f(Y|\boldsymbol{\theta})$, we wish to entertain an alternative model built from $\pi_0^*(\boldsymbol{\theta})$ and $f^*(Y|\boldsymbol{\theta})$.

We could just repeat MCMC computation on the new model: note that even where the base model is rather tractable (for example, $\pi_0(\boldsymbol{\theta})$ conjugate to $f(Y|\boldsymbol{\theta})$), we should consider alternatives that are not. Thus MCMC may be needed in sensitivity analysis even where exact analytic calculation handles the standard model, or we may need Metropolis where Gibbs was sufficient in the standard case.

As an alternative to treating the revised model as a completely fresh problem, we may be able to make use of importance sampling to assess sensitivity using only the original simulation. This uses the importance sampling identity

$$E_{\pi^*}(g) = E_\pi\left(\frac{\pi^*}{\pi}g\right)$$

showing that expectations under π^* can be estimated from an MCMC run aimed at π, by

$$\frac{\sum_{t=1}^{N} w(\boldsymbol{x}^{(t)})g(\boldsymbol{x}^{(t)})}{\sum_{t=1}^{N} w(\boldsymbol{x}^{(t)})}$$

where

$$w(\boldsymbol{x}) \propto \frac{\pi^*(\boldsymbol{x})}{\pi(\boldsymbol{x})}.$$

There are several practical examples of MCMC-based sensitivity analysis in Besag, et al. (1995).

One problem of the importance sampling approach is that, except in very low dimensional problems or where π and π^* are very similar, $w(\boldsymbol{x}^{(t)})/\sum_t w(\cdot)$ is effectively concentrated on very few samples, implying very poor efficiency. This can sometimes be mitigated by considering infinitesimal perturbations instead: $\pi_\varepsilon(\boldsymbol{x}) \propto (\pi(\boldsymbol{x}))^{(1-\varepsilon)}(\pi^*(\boldsymbol{x}))^\varepsilon$, or, of course, by running another chain.

1.10.3 Bayes with a loss function

We have seen the tremendous advantages that MCMC offers to the practising Bayesian through the opportunities it gives for computing posterior distributions. However, the complete Bayesian agenda for statistical analysis does not stop at computing posteriors — in the full decision theoretic framework, a loss function is introduced, and optimal Bayes estimates and decisions determined by minimising the expectation of the loss under the posterior distribution.

Writing the posterior $p(\boldsymbol{x} \in \cdot|Y)$ in the generic $\pi(\cdot)$ notation, we wish to choose an action $\boldsymbol{z}$ to minimise

$$E(L(\boldsymbol{x}, \boldsymbol{z})|Y) = \int L(\boldsymbol{x}, \boldsymbol{z})\pi(d\boldsymbol{x}),$$

where $L(\boldsymbol{x}, \boldsymbol{z})$ is the loss incurred through taking action $\boldsymbol{z}$ when the true state of nature is $\boldsymbol{x}$.

When the posterior is computed using MCMC, the expectation is replaced by an empirical average over the realisation $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$:

$$E(L(\boldsymbol{x}, \boldsymbol{z})|Y) \approx \frac{1}{N}\sum_{t=1}^{N} L(\boldsymbol{x}^{(t)}, \boldsymbol{z}).$$

The difficulty with this approach lies in the interplay between the averaging and the optimisation with respect to $\boldsymbol{z}$.

One class of loss functions that can be easily handled is that of finite sums of separable loss functions, where

$$L(\boldsymbol{x}, \boldsymbol{z}) = \sum_r a_r(\boldsymbol{x})b_r(\boldsymbol{z}). \qquad (1.16)$$

Then the MCMC computation and the optimisation separate, since

$$E(L(\boldsymbol{x}, \boldsymbol{z})|Y) \approx \sum_r A_r b_r(\boldsymbol{z}) \quad \text{where} \quad A_r = \frac{1}{N} \sum_{t=1}^{N} a_r(\boldsymbol{x}^{(t)}).$$

The optimisation can even sometimes be done analytically, as for example in the elementary case of squared-error loss for some functional $g(\boldsymbol{x})$: $L(\boldsymbol{x}, z) = (g(\boldsymbol{x}) - z)^2$. Then $A_r, r = 0, 1, 2$ are the 0^{th}, 1^{st} and 2^{nd} empirical moments of g, and $b_r(z) = z^2, -2z, 1$ for $r = 0, 1, 2$, leading of course to the optimal $z = A_1/A_0$, the MCMC estimate of the posterior mean.

More commonly, numerical optimisation is necessary. Several research papers recently (for example, Rue and Hurn (1999), Rue and Syversveen (1998)) have used simulated annealing, and have exploited the representation (1.16). The A_r are first computed by MCMC, and then a second, annealing, simulation, set up for the artificial probability distributions

$$p_T(\boldsymbol{z}) \propto \exp\left(\frac{-1}{T} \sum_r A_r b_r(\boldsymbol{z})\right),$$

where the temperature T is sent to 0 on some suitable schedule.

1.11 Some notes on programming MCMC

1.11.1 The Bugs *software*

The only software I am aware of that provides MCMC computation for a wide range of statistical models, without requiring the user to code the sampling algorithms is Bugs (Gilks, Thomas and Spiegelhalter, 1994). The model is specified in a high-level specification language – or, if using the WinBugs version, a graphical interface – a few options controlling the simulation are entered, and the system does the rest. Particularly for rather well-understood standard models, hierarchical versions of generalised linear models, for example, the facilities are easy to use, and the system extremely effective. The suite of implemented and documented examples distributed as part of the software release demonstrates the remarkable flexibility of the system, and some very complex models can be handled. However, facilities for using anything other than single-variable Gibbs sampling are very limited, so for some models, Bugs may be inefficient or even completely incapable. The authors are careful to stress that setting up a MCMC sampler and interpreting the output, even with Bugs, requires knowledge of the user beyond appreciation of the statistical model itself.

Bugs is very useful for many practitioners, but may be too limited for most MCMC researchers.

1.11.2 Your own code

Programming your own MCMC method from scratch is much less daunting than it might first appear, and provides flexibility and (run-time) efficiency that cannot really be matched by package software. The basic recipes are simple in structure, and may be coded following the algebraic notation almost exactly. I do not usually bother with Gibbs sampling unless the full conditionals are entirely standard distributions for which I have a random number generator. (The adaptive rejection sampling method of Gilks and Wild (1992) provides a means of extending Gibbs sampling to a wider range of full conditionals.) In Metropolis-Hastings algorithms, it is necessary to take some care with floating point arithmetic, as in complex models, there may be many multiplicative factors entering the acceptance ratio, with a wide numerical range. I find it convenient to accumulate the sum of the logarithms of the factors, and then truncate onto a safe range before exponentiating.

High and low level languages

The poor performance of looping code in most high-level statistical languages such as **S** precludes their use in coding all but the smallest problems. I always use Fortran or C. On the other hand, the flexibility of control and the availability of a wide range of statistical and graphical procedures in **S**, and similar systems, is absolutely invaluable in analysis of MCMC in a research environment. My usual strategy is to dump large quantities of raw MCMC output into a collection of files, with structured filenames, and then employ a suite of **S** functions to read, display and analyse these.

Validating your code

It is absolutely essential to check and double-check MCMC code. The very nature of the output of the computation — simulation results in a context where other numerical methods are not available for cross-checking — makes this problematical, especially in Bayesian statistical contexts, where problems are one-off, and data subject to sampling variation. Testing on simulated data-sets with known parameter values does not tell you very much.

I find two particular tricks extremely useful. First, I always use restartable random number streams, so that I can conveniently and reliably duplicate a run, with additional diagnostic output, if I suspect a bug. This is also often useful to compare results before and after a minor edit. Second, in programs implementing posterior simulation for a Bayesian model in which the variables are organised as a directed acyclic graph, and in which the observed data have no ‘children’, I always include a ‘prior’ option, which ignores the data and the likelihood terms. The posterior simulation program, largely unaltered, is then actually simulating from the prior distribution, and typically many marginal and conditional aspects thereof may be directly checked as the true distributions are known.

Many useful hints on the practical details of algorithm design, including matters such as thinning and burn-in, will be found in Geyer (1992) and Gilks and Roberts (1996).

1.12 Conclusions

1.12.1 Some strengths of MCMC

MCMC is evidently a very powerful and flexible tool for computation with complex multivariate distributions. Its availability has transformed practical Bayesian statistics, and it is making an important if less dramatic impact on other areas of computational statistics.

In the Bayesian context, its power derives from the two kinds of flexibility it offers. First there is flexibility in modelling, permitting the analyst to get much closer to his or her understanding of the reality of the process generating the data, and liberating the modelling process from the constraints only imposed for the sake of tractability. A desirable by-product is the encouragement to model builders to think in graphical terms, as MCMC is particularly well-adapted for models defined on sparse graphs.

Secondly, there is freedom in inference; in principle, there are no limits to what features of the target distribution may be estimated by MCMC, although one needs always to be aware of the Monte Carlo errors unavoidable in such estimates: some features of the target can be computed much more reliably than others. MCMC addresses questions only posed after simulation completed (e.g. ranking and selection) and offers opportunities for simultaneous inference. It allows and even facilitates sensitivity analysis, and addresses questions of model comparison, criticism and choice.

1.12.2 Some weaknesses and dangers

MCMC is not a panacea. In the end it is only a numerical method, and does not displace the need for careful thought about modelling, and about the probable reliability of numerical results obtained in the given context. When other methods are available, MCMC can be relatively extremely expensive; hence the common preference in fields with large data-sets such as signal and image processing for approximations to the full Bayesian paradigm that are amenable to fast numerical calculations for particular outputs of interest.

In qualitative terms, a problem that is insurmountable (at least in estimating expectations and probabilities) is the order $\sqrt{N}$ precision of any simulation method, and for MCMC, the possibility of slow convergence, especially when it is not diagnosable.

Use of MCMC imposes serious responsibilities on the careful researcher, for there is the risk that fitting technology runs ahead of statistical science, so that models are fitted that are not understood, and the risk of overusing the flexibility allowed in inference, leading to undisciplined, selective presentation of posterior information.

1.12.3 Some important lines of continuing research

MCMC remains an important, exciting and challenging field for further research. It is impossible to predict how the field will develop over the next few years, but I believe that the most interesting questions for exploration at present include:

a. Adaptive methods, and other possibilities for automation;

b. Perfect simulation: will these become useful in statistical practice?

c. Getting quantitative results from theoretical analysis;

d. Learning even more from physics.

Acknowledgements

I have presented tutorial lectures on MCMC on several occasions over the last 5 years, to different kinds of audience. Most of these have been in partnership with David Spiegelhalter, who in addition to his wonderful expository powers has enlivened the lecture series with biomedical applications and with material on the BUGS software for Gibbs sampling. I am indebted to him.

I am also grateful 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.

Some of this work was supported by the EPSRC.

1.13 References and further reading

Aykroyd, R. G. and Green, P. J. (1991). Global and local priors, and the location of lesions using gamma-camera imagery. *Phil. Trans. R. Soc.*, A, **337**, 323–342.

Barone, P. and Frigessi, A. (1989). Improving stochastic relaxation for Gaussian random fields. *Probability in the Engineering and Informational Sciences*, **4**, 369–389.

Barone, P., Sebastiani, G. and Stander, J. (1998) *Over-relaxation methods and Metropolis-Hastings coupled Markov chains for Monte Carlo simulation.* Technical report, University of Plymouth.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley, Chichester.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, B, **36**, 192–236.

Besag, J. (1994). Contribution to the discussion of paper by Grenander and Miller (1994). *Journal of the Royal Statistical Society*, B, **56**, 591–592.

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.

Besag, J. and York, J. C. (1989). Bayesian restoration of images. In *Analysis of Statistical Information*, (ed. T. Matsunawa), pp. 491–507. Inst. Statist. Math., Tokyo.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

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

Clifford, P. (1990). Markov random fields in statistics. In *Disorder in physical systems* (eds. G. R. Grimmett and D. J. A. Welsh), pp. 19–32. Clarendon Press, Oxford.

Creutz, M. (1979). Confinement and the critical dimensionality of space-time. *Physical Review Letters*, **43**, 553–556.

Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society*, B, **61**, 331–344.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society*, B, **60**, 333-350.

Edwards, R. G. and Sokal, A. D. (1988). Generalization of the Fortuin-Kastelyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review*, D, **38**, 2009–2012.

Fill, J. A. (1998). An interruptible algorithm for exact sampling via Markov chains. *Annals of Applied Probability*, **8**, 131–162.

Frigessi, A., Martinelli, F. and Stander, J. (1997). Computational complexity of Markov chain Monte Carlo methods for finite Markov random fields. *Biometrika*, **84**, 1–18.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika* **82**, 479–488.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Roberts, G. O. and Gilks, W. R. (1996) Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 599–607. Clarendon Press, Oxford.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (ed. E. M. Keramidas), pp. 156–163. Interface Foundation, Fairfax Station.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **8**, 473–483.

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

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90**, 909–920.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with discussion). *Journal of the Royal Statistical Society*, B, **55**, 39–52.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds.) (1996). *Markov chain Monte Carlo in practice*, Chapman and Hall, London.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC. Chapter 6 (pp. 89–114) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, **43**, 169–178.

Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, **86**, 785–801.

Gordon, M. I. and Lifšic, B. A. (1978) The central limit theorem for stationary Markov processes. *Soviet Math. Dokl.*, **19**, 392–394.

Green, P. J. (1994). Contribution to the discussion of paper by Grenander and Miller (1994). *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 Han, X.-L. (1992). Metropolis methods, Gaussian proposals and antithetic variables. In *Stochastic models, Statistical Methods and Algorithms in Image Analysis, Lect. Notes Statist.*, **74**, 142–164, Springer-Verlag, Berlin.

Green, P. J. and Murdoch, D. J. (1998). Exact sampling for Bayesian inference: towards general purpose algorithms. In *Bayesian Statistics 6*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 301–321. Clarendon Press, Oxford.

Green, P. J. and Richardson, S. (1998). *Modelling heterogeneity with and without the Dirichlet process.* Submitted.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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.

Hodgson, M. E. A. and Green, P. J. (1999). Bayesian choice among Markov models of ion channels using Markov chain Monte Carlo. *Proceedings of the Royal Society of London* A, **455**, 3425–3448.

Holmes, C. C. and Mallick, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation*, **10**, 1217–1233.

Kass, R. E., Tierney, L. and Kadane, J. B. (1988). Asymptotics in Bayesian computation (with discussion). In *Bayesian Statistics 3* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 261-278. Clarendon Press, Oxford.

Kipnis, C. and Varadhan, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, **104**, 1–19.

Lauritzen, S. L. (1996). *Graphical models*, Clarendon Press, Oxford.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society*, B, **50**, 157–224.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, **19**, 451–458.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, **24**, 101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953).Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability.* Springer-Verlag, New York.

Mira, A. and Geyer, C. J. (1999). *Ordering Monte Carlo Markov chains.* Technical Report No. 632, School of Statistics, University of Minnesota.

Mira, A., Møller, J. and Roberts, G. O. (1999). Perfect slice samplers. Technical report R-99-2020, Department of Mathematical Sciences, Aalborg University, Denmark.

Møller, J. (1999). Perfect simulation of some conditionally-specified models. *Journal of the Royal Statistical Society*, B **61**, 251–264.

Møller, J. and Nicholls, G. K. (1999). Perfect simulation for sample-based inference. Technical report R-99-2011, Department of Mathematical Sciences, Aalborg University, Denmark.

Mooley, D.A. (1981). Applicability of the Poisson probability model to the severe cyclonic storms striking the coast around the Bay of Bengal. *Sankhya*, **43** B, 187–197.

Murdoch, D. J. and Green, P. J. (1998). Exact sampling from a continuous state space. *Scandinavian Journal of Statistics*, **25**, 483–502.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, **90**, 233–241.

Neal, R. M. (1997). *Markov chain Monte Carlo methods based on 'slicing' the density function.* Technical report #9722, University of Toronto.

Neal, R. M. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation, in M. I. Jordan (editor) *Learning in Graphical Models* pp. 205–228, Kluwer Academic Publishers, Dordrecht.

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.

Nobile, A. and Green, P. J. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, **87**, (in press).

O'Hagan, A. (1994). *Bayesian Inference (Kendall's Advanced theory of Statistics,* **2 B***)*, Edward Arnold, London.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, **60**, 607–612.

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.

Pievatolo, A. and Green, P. J. (1998). Boundary detection through dynamic polygons. *Journal of the Royal Statistical Society*, B **60**, 609–626.

Priestley, M. (1981). *Spectral Analysis and Time Series.* Academic Press, London.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, **9**, 223–252.

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

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains. *Journal of the Royal Statistical Society*, B **61**, 643–660.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society*, B **59**, 291–317.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–100.

Rue, H. and Hurn, M. A. (1999). Bayesian object identification. *Biometrika*, **86**, 649–660.

Rue, H. and Syversveen, A. R. (1998). Bayesian object recognition using Baddeley's delta loss. *Advances in Applied Probability*, **30**, 64–84.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society*, B, **55**, 3–23.

Sokal, A. D. (1989). Monte Carlo methods in statistical mechanics: foundations and new algorithms. *Cours de Troisiéme Cycle de la Physique en Suisse Romande*. Lausanne.

Stramer, O. and Tweedie, R. L. (1998). Langevin-type models II: self-targetting candidates for MCMC algorithms. Methodology and Computing in Applied Probability, **1**, 307–328.

Suomela, P. (1976). Unpublished Ph.D. thesis. University of Jyväskylä, Finland.

Swendsen, R. H. and Wang, J.-S. (1987). Non-universal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, **58**, 86–88.

Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics*, **22**, 1701–1762.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. Chapter 4 (pp. 59–74) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

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

Uimari, P. and Hoeschele, I. (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo. *Genetics*, **146**, 734–743.

Contents