Markov chain Monte Carlo in action: a tutorial

Peter J. Green
University of Bristol, Department of Mathematics, Bristol, BS8 1TW, UK.
P.J.Green@bristol.ac.uk

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.

2. Cyclones example: point processes and change points

We will 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 interval $[0, L)$. We suppose the events occur as a Poisson process — but at a possibly non-uniform rate: say $x(t)$ per unit time, at time t; we wish to make inference about $x(t)$. We consider 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 100 years.

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 $\widehat{x} = \frac{N}{L}$, the *maximum likelihood estimator* of x under the assumption that N has a Poisson distribution, with mean xL.

Model 2: constant rate, the Bayesian way. For a Bayesian approach to this example, suppose that we have prior information about x (from previous studies, for example). Suppose we can model this by $x \sim \Gamma(\alpha, \beta)$.
Then we find that *a posteriori* 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.
There is no need for MCMC in this model: you can calculate the posterior exactly, and recognise it as a standard distribution; it only worked like this because we used a *conjugate* prior.

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 (a formulation that makes more sense in our next formulation, model 4).
Now $p(x|N, \alpha, e, f)$ no longer has an explicit form, but the *full conditionals* $p(x|N, \alpha, \beta, e, f)$ 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

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

What happens if we generate a sample of (x, β) pairs by alternately drawing x and β from these distributions?
Figure 1 shows the first few moves of this process applied to model 3 on the cyclones data; we took $e = 1$ and $f = N/L = 1.41$. The marginal distribution for x, as accumulated from the first 1000 sweeps of this process is also displayed in Figure 1.

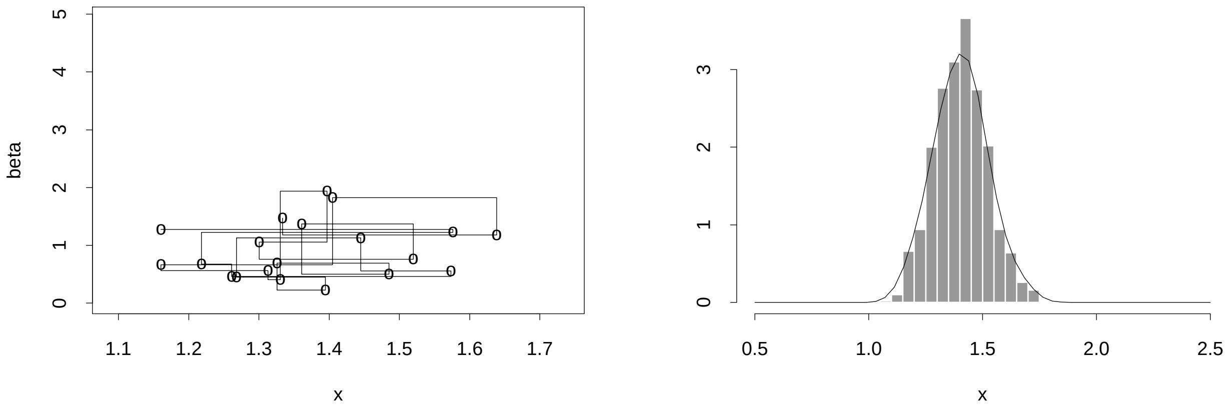

Figure 1: First few moves of the Gibbs sampler, and marginal distribution for β from 1000 sweeps, for the cyclones data, model 3.

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 is the (unique) invariant distribution of the chain. Standard theorems imply that the chain converges to this invariant distribution in several useful senses, so that we can treat the realised values as a sample from the posterior.

Model 4: constant rate, with change point. Now let us suppose $x(t)$ is a step function (a suitable model if we postulate one or more *change points*; the process is completely random, but the rate switches between levels). Let us take k change points at *known* times $T_1, T_2, \ldots, T_k$ with

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

Suppose that $x_0, x_, \ldots, x_k$ are *a priori* independently drawn from Gamma distributions, as before: $x_j \sim \Gamma(\alpha, \beta)$. Then if $N_0, N_1, \ldots, N_k$ are the numbers of events between adjacent $\{T_j\}$, the above method extends to sampling in turn from

$$x_j | \cdots \sim \Gamma(\alpha + N_j, \beta + T_{j+1} - T_j), j = 0, 1 \ldots, k \text{ and}$$

$$\beta | \cdots \sim \Gamma(e + (k+1)\alpha, f + \sum_{i=o}^{k} x_i),$$

forming a Markov chain with a $(k+2)$-dimensional state space $\{(x_0, x_1, \ldots, x_k, \beta)\}$. Note that we write '$|\cdots$' to mean 'given all other variables' — including the data.

The hierarchical model using random β allows 'borrowing strength' in estimation from all the data together: the x_j are conditionally independent given β, but are *un*conditionally *dependent*. In inference their values will be shrunk together.

3. Beyond the Gibbs sampler – MCMC in general

Having motivated the idea of MCMC by use of the Gibbs sampler in a very basic problem, we are now in a position to discuss the subject from a rather more general perspective.
The objective is to construct a discrete time Markov chain whose state space is $\mathcal{X}$ (the parameter space in Bayesian statistics), and whose limiting distribution is a specified target (e.g. a Bayesian posterior). That is, we want a transition kernel P such that

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

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.
A number of standard recipes have been developed (see, for example, Besag, *et al.*, 1995), including the Gibbs sampler, and the Metropolis and Hastings methods; the latter is very general, and can even be extended to cases where the parameter space is not of fixed dimension. We illustrate some of the extensions to our point process model that can also be handled.

Model 5: another hyperparameter. Let us now suppose α is also unknown, with, *a priori*, $\alpha \sim \Gamma(c, d)$ for fixed constants c and d. (For 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 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 6: 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* $p(T_1, T_2, \ldots, T_k) \propto T_1(T_2 - T_1)\ldots(T_k - T_{k-1})(L - T_k)$.
The posterior marginal or joint conditional distributions are quite complex, for this or any prior, so Metropolis-Hastings is needed. The details are a little messy but straightforward. The acceptance probability for a proposal that T_j' be drawn uniformly from $[T_{j-1}, T_{j+1}]$ is

$$\min\left\{1, (\text{likelihood ratio})\frac{(T_j' - T_{j-1})(T_{j+1} - T_j')}{(T_j - T_{j-1})(T_{j+1} - T_j)}\right\}.$$

Model 7: 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):

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

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!

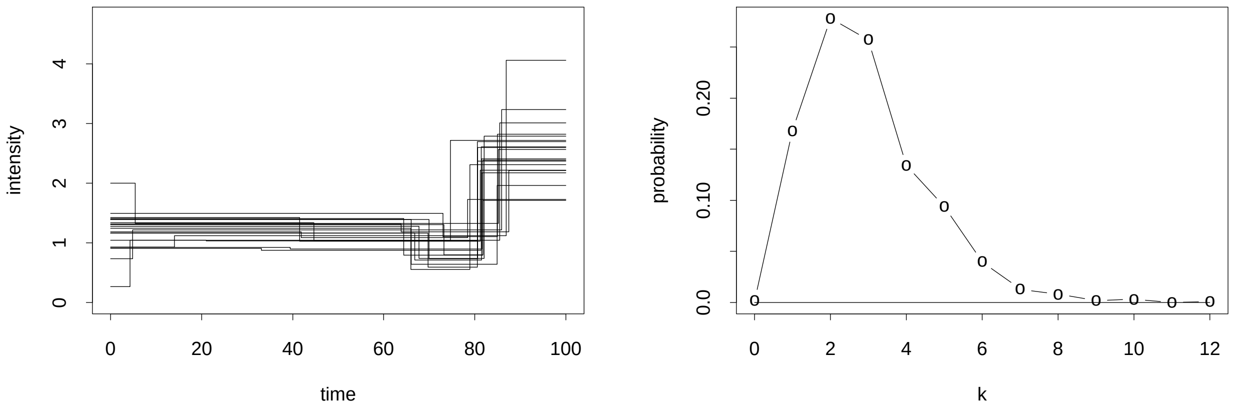

Figure 2: Posterior sample of step functions $x(t)$ for model 6 with $k = 2$, and posterior for k in model 7, applied to cyclones data.

A variable-dimension Metropolis-Hastings algorithm was applied to this problem, setting the hyperparameter $\lambda = 3$, and one aspect of the resulting analysis is displayed in Figure 2.

References and Further Reading

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

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

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

Green, P. J. (1999) *A primer on Markov chain Monte Carlo.* To appear (a preprint version can be found at `http://www.stats.bris.ac.uk/~peter/papers/semstat.ps`).

Résumé

Les méthode de Monte Carlo par Châines de Markov datent d'environ cinquante ans et ont été dévelopées et utilisées en Physique pendant les quarante dernières années. Cependant leur impact le plus spectaculaire et leur influence en Statistique et Probabilité ne remontent qu'à la fin des années 80.
A l'heure actuelle, ces méthodes sont omniprésentes dans les calculs statistiques, en particulier dans l'inférence bayésienne et tout spécialement dans l'analyse des systèmes stochastiques complexes.