Markov chain Monte Carlo

by

Gareth O. Roberts[1] and Jeffrey S. Rosenthal[2]

(April 2003.)

1 Introduction

One of the simplest and most powerful practical uses of the ergodic theory of Markov chains is in Markov chain Monte Carlo (MCMC). Suppose we wish to simulate from a probability density π (which will be called the *target* density) but that direct simulation is either impossible or practically infeasible (possibly due to the high dimensionality of π). This generic problem occurs in diverse scientific applications, for instance Statistics, Computer Science, and Statistical Physics.

Markov chain Monte Carlo offers an indirect solution based on the observation that it is much easier to construct an ergodic Markov chain with π as a stationary probability measure, than to simulate directly from π. This is because of the ingenious *Metropolis-Hastings* algorithm which takes an arbitrary Markov chain and adjusts it using a simple accept-reject mechanism to ensure the stationarity of π for the resulting process.

The algorithms was introduced by Metropolis *et al.* (1953) in a Statistical Physics context, and was generalised by Hastings (1970). It was considered in the context of image analysis (Geman and Geman, 1984) data augmentation (Tanner and Wong, 1987). However, its routine use in Statistics (especially for Bayesian inference) did not take place until its popularisation by Gelfand and Smith (1990). For modern discussions of MCMC, see e.g. Tierney (1994), Smith and Roberts (1993), Gilks *et al.* (1996), and Roberts and Rosenthal (1998b).

The number of financial applications of MCMC is rapidly growing (see for example the reviews of Kim *et al.*, 1996 and Johannes and Polson, 2003). In this area, important problems revolve around the need to impute latent (or imperfectly observed) time-series such as stochastic volatility processes. Modern developments have often combined the use of MCMC methods with filtering or particle filtering methodology. In Actuarial Sciences, MCMC appears to have huge potential in hitherto intractabile inference problems, much of this untapped as yet (though see Scollnik, 2001, Ntzoufras and Dellaportas, 2002, and Bladt *et al.*, 2003).

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K. LA1 4YF. E-mail: g.o.roberts@lancaster.ac.uk. Web: http://www.maths.lancs.ac.uk/dept/people/robertsg.html

[2]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Supported in part by NSERC of Canada. E-mail: jeff@math.toronto.edu. Web: http://probability.ca/jeff/

2 The Basic Algorithms

Suppose that π is a (possibly unnormalised) density function, with respect to some reference measure (e.g. Lebesgue measure, or counting measure) on some state space $\mathcal{X}$. Assume that π is so complicated, and $\mathcal{X}$ is so large, that direct numerical integration is infeasible. We now describe several MCMC algorithms, which allow us to approximately sample from π. In each case, the idea is to construct a Markov chain update to generate X_{t+1} given X_t, such that π is a *stationary distribution* for the chain, i.e. if X_t has density π, then so will X_{t+1}.

2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm proceeds in the following way. An initial value X_0 is chosen for the algorithm. Given X_t, a *candidate transition* Y_{t+1} is generated according to some fixed density $q(X_t, \cdot)$, and is then accepted with probability $\alpha(X_t, Y_{t+1})$, given by

$$\alpha(x,y) = \begin{cases} \min\{\frac{\pi(y)}{\pi(x)}\frac{q(y,x)}{q(x,y)}, 1\} & \pi(x)q(x,y) > 0 \\ 1 & \pi(x)q(x,y) = 0 \ , \end{cases}$$

otherwise it is rejected. If Y_{t+1} is accepted, then we set $X_{t+1} = Y_{t+1}$. If Y_{t+1} is rejected, we set $X_{t+1} = X_t$. By iterating this procedure, we obtain a Markov chain realisation $X = \{X_0, X_1, X_2, \ldots\}$.

The formula (2.1) was chosen precisely to ensure that, if X_t has density π, then so does X_{t+1}. Thus, π is stationary for this Markov chain. It then follows from the ergodic theory of Markov chains (see Section 3) that, under mild conditions, for large t, the distribution of X_t will be approximately that having density π. Thus, for large t, we may regard X_t as a *sample observation* from π.

Note that this algorithm requires only that we can simulate from the density $q(x, \cdot)$ (which can be chosen essentially arbitrarily), and that we can compute the probabilities $\alpha(x, y)$. Further note that this algorithm only ever requires the use of *ratios* of π values, which is convenient for application areas where densities are usually known only up to a normalisation constant, including Bayesian Statistics and Statistical Physics.

2.2 Specific versions of the Metropolis-Hastings algorithm

The simplest and most widely applied version of the Metropolis-Hastings algorithm is the so-called *symmetric random walk Metropolis* algorithm (RWM). To describe this method, assume that $\mathcal{X} = \mathbf{R}^d$, and let q denote the transition density of a random walk with spherically symmetric transition density: $q(x, y) = g(\|y - x\|)$ for some g. In this case, $q(x, y) = q(y, x)$, so (2.1) reduces to

$$\alpha(x,y) = \begin{cases} \min\{\frac{\pi(y)}{\pi(x)}, 1\} & \pi(x)q(x,y) > 0 \\ 1 & \pi(x)q(x,y) = 0 \ . \end{cases}$$

Thus all moves to regions of larger π values are accepted, whereas all moves to lower values of π are potentially rejected. Thus the accept/reject mechanism 'biases' the random walk in favour of areas of larger π values.

Even for RWM, it is difficult to know how to choose the spherically symmetric function g. However, it is proved by Roberts *et al.* (1997) that, if the dimension d is large, then under appropriate conditions, g should be scaled so that the asymptotic acceptance rate of the algorithm is about 0.234, and the required running time is $O(d)$; see also Roberts and Rosenthal (2001).

Another simplification of the general Metropolis-Hastings algorithm is the *independence sampler*, which sets the proposal density $q(x, y) = q(y)$ to be independent of the current state. Thus the proposal choices just form an i.i.d. sequence from the density q, though the derived Markov chain gives a dependent sequence as the accept/reject probability still depends on the current state.

Both the Metropolis and independence samplers are *generic* in the sense that the proposed moves are chosen with no apparent reference to the target density π. One Metropolis algorithm which *does* depend on the target density is the *Langevin algorithm*, based on discrete approximations to diffusions, first developed in the Physics literature (Rossky *et al.*, 1978). Here $q(x, \cdot)$ is the density of a normal distribution with variance δ and mean $x + (\delta/2)\nabla\pi(x)$ (for small fixed $\delta > 0$), thus pushing the proposal in the direction of increasing values of π, hopefully speeding up convergence. Indeed, it is proved by Roberts and Rosenthal (1998a) that, in large dimensions, under appropriate conditions, δ should be chosen so that the asymptotic acceptance rate of the algorithm is about 0.574, and the required running time is only $O(d^{1/3})$; see see also Roberts and Rosenthal (2001).

2.3 Combining different algorithms: hybrid chains

Suppose $P_1, \ldots, P_k$ are k different Markov chain updating schemes, each of which leaves π stationary. Then we may *combine* the chains in various ways, to produce a new chain which still leaves π stationary. For example, we can run them in sequence to produce the *systematic-scan* Markov chain given by $P = P_1 P_2 \ldots P_k$. Or, we can select one of them uniformly at random at each iteration, to produce the *random-scan* Markov chain given by $P = (P_1 + P_2 + \ldots + P_k)/k$.

Such combining strategies can be used to build more complicated Markov chains (sometimes called *hybrid chains*), out of simpler ones. Under some circumstances, the hybrid chain may have good convergence properties (see e.g. Roberts and Rosenthal, 1997, 1998c). In addition, such combining are the essential idea behind the Gibbs sampler, discussed next.

2.4 The Gibbs sampler

Assume that $\mathcal{X} = \mathbf{R}^d$, and that π is a density function with respect to $d-$dimensional Lebesgue measure. We shall write $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$ for an element of $\mathbf{R}^d$ where $\mathbf{x}^{(i)} \in \mathbf{R}$, for $1 \leq i \leq d$. We shall also write $\mathbf{x}^{(-i)}$ for any vector produced by omitting the ith component, $\mathbf{x}^{(-i)} = (x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots x^{(d)})$, from the vector x.

The idea behind the Gibbs sampler is that even though direct simulation from π may not be possible, the 1-dimensional conditional densities $\pi_i(\cdot \mid \mathbf{x}^{(-i)})$, for $1 \leq i \leq d$, may be much more amenable to simulation. This is a very common situation in many simulation examples, such as those arising from the Bayesian analysis of *hierarchical models* (see e.g. Gelfand and Smith, 1990).

The Gibbs sampler proceeds as follows. Let P_i be the Markov chain update that, given $X_t = \mathbf{x}$, samples $X_{t+1} \sim \pi_i(\cdot \mid \mathbf{x}^{(-i)})$, as above. Then the *systematic-scan Gibbs sampler* is given by $P = P_1 P_2 \ldots P_d$, while the *random-scan Gibbs sampler* is given by $P = (P_1 + P_2 + \ldots + P_d) / d$.

The following example from Bayesian Statistics illustrates the ease with which a fairly complex model can be fitted using the Gibbs sampler. It is a simple example of a *hierarchical model*.

Example. Suppose that for $1 \leq i \leq n$. we observe data $\mathbf{Y}$ which we assume is independent from the model $Y_i \sim \text{Poisson}(\lambda_i)$. The λ_is are termed individual level *random effects*. As a hierarchical prior structure, we assume that conditional on a parameter θ, the λ_i are independent with distribution $\lambda_i|\theta \sim$Exponential(θ), and impose an exponential prior on θ, say $\theta \sim$Exponential(1).

The multivariate distribution of $(\theta, \boldsymbol{\lambda}|\mathbf{Y})$ is complex, possibly high-dimensional, and lacking in any useful symmetry to help simulation or calculation (essentially because data will almost certainly vary). However the Gibbs sampler for this problem is easily constructed by noticing that $(\theta|\boldsymbol{\lambda}, \mathbf{Y}) \sim \text{Gamma}(n, \sum_{i=1}^n \lambda_i)$ and that $(\lambda_i|\theta, \text{other } \lambda_j\text{s}, \mathbf{Y}) \sim \text{Gamma}(Y_i + 1, \theta + 1)$.

Thus the algorithm iterates the following procedure. Given $(\boldsymbol{\lambda}_t, \theta_t)$,

- for each $1 \leq i \leq n$, we replace $\lambda_{i,t}$ by $\lambda_{i,t+1} \sim \text{Gamma}(Y_i + 1,\ \theta_t + 1)$;
- then we replace θ_t by $\theta_{t+1} \sim \text{Gamma}(n,\ \sum_{i=1}^n \lambda_{i,t+1})$;

thus generating the vector $(\boldsymbol{\lambda}_{t+1}, \theta_{t+1})$.

The Gibbs sampler construction is highly dependent on the choice of coordinate system, and indeed its efficiency as a simulation method can vary wildly for different parameterisations; this is explored in e.g. Roberts and Sahu (1997).

While many more complex algorithms have been proposed and have uses in hard simulation problems, it is remarkable how much flexibility and power is provided by just the Gibbs sampler, RWM, and various combinations of these algorithms.

3 Convergence

3.1 Ergodicity

MCMC algorithms are all constructed to have π as a stationary distribution. However, we require extra conditions to ensure that they converge in distribution to π. Consider for instance the following examples.

1. Consider the Gibbs sampler for the density π on $\mathbf{R}^2$ corresponding to the uniform density on the subset

$$S = ([-1, 0] \times [-1, 0]) \ \cup \ ([0, 1] \times [0, 1]) \, .$$

   For positive X values, the conditional distribution of $Y|X$ is supported on $[0, 1]$. Similarly, for positive Y values, the conditional distribution of $X|Y$ is supported on $[0, 1]$. Therefore, started in the positive quadrant, the algorithm will never reach $[-1, 0]\times[-1, 0]$ and therefore must be *reducible*. In fact for this problem, although π is a stationary distribution, there are infinitely many different stationary distributions corresponding to arbitrary convex mixtures of the uniform distributions on $[-1, 0]\times[-1, 0]$ and on $[0, 1] \times [0, 1]$.

2. Let $\mathcal{X} = \{0, 1\}^d$ and suppose that π is the uniform distribution on $\mathcal{X}$. Consider the Metropolis algorithm which takes at random one of the d dimensions and proposes to switch its value, i.e.

$$\mathbf{P}(x_1, \ldots x_d; \quad x_1, \ldots x_{i-1}, 1 - x_i, x_{i+1}, \ldots x_d) = \frac{1}{d}$$

   for each $1 \leq i \leq d$. Now it is easy to check that in this example, all proposed moves are accepted, and the algorithm is certainly irreducible on $\mathcal{X}$. However

$$\mathbf{P}\left(\sum_{i=1}^{d} X_{i,t} \text{ is even } \mid \mathbf{X}_0 = (0, \ldots 0)\right) = \begin{cases} 1, & d \text{ even} \\ 0, & d \text{ odd} \, . \end{cases}$$

   Therefore the Metropolis algorithm is *periodic* in this case.

On the other hand, call a Markov chain *aperiodic* if there do not exist disjoint non-empty subsets $\mathcal{X}_1, \ldots, \mathcal{X}_r \subseteq \mathcal{X}$ for some $r \geq 2$, with $\mathbf{P}(X_{t+1} \in \mathcal{X}_{i+1} \,|\, X_t) = 1$ whenever $X_t \in \mathcal{X}_i$ for $1 \leq i \leq r - 1$, and $\mathbf{P}(X_{t+1} \in \mathcal{X}_1 \,|\, X_t) = 1$ whenever $X_t \in \mathcal{X}_r$. Furthermore, call a Markov chain *ϕ-irreducible* if there exists a non-zero measure ϕ on $\mathcal{X}$, such that for all $x \in \mathcal{X}$ and all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, there is positive probability that the chain will eventually hit A if started at x. Call a chain *ergodic* if it is both ϕ-irreducible and aperiodic, with stationary density function π. Then it is well known (see e.g. Nummelin, 1984; Meyn and Tweedie, 1993; Tierney, 1994; Smith and Roberts, 1993; Rosenthal, 2001) that, for an ergodic Markov

chain on the state space $\mathcal{X}$ having stationary density function π, the following convergence theorem holds. For any $B \subseteq \mathcal{X}$ and π-a.e. $x \in \mathcal{X}$,

$$\lim_{t\to\infty} \mathbf{P}(\mathbf{X}_t \in B \,|\, X_0 = x) \;=\; \int_B \pi(y)\, dy \;; \tag{1}$$

and for any function f with $\int_{\mathcal{X}} |f(x)|\pi(x)\, dx < \infty$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \int_{\mathcal{X}} f(x)\pi(x)\, dx\,, \quad a.s.\;. \tag{2}$$

In particular, π is the unique stationary probability density function for the chain.

3.2 Geometric ergodicity and CLT's

Under slightly stronger conditions (e.g. *geometric ergodicity*, meaning the convergence in (1) is exponentially fast, together with $\int |g|^{2+\epsilon}\,\pi < \infty$), a *Central Limit Theorem* (CLT) will hold, wherein $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (f(X_t) - \int_{\mathcal{X}} f(x)\pi(x)dx)$ will converge in distribution to a normal distribution with mean 0, and variance $\sigma_f^2 \equiv Var_\pi(f(X_0)) + 2\sum_{i=1}^{\infty} Cov_\pi(f(X_0),\, f(X_i))$, where the subscript π means we start with X_0 having density π (see e.g. Geyer, 1992; Tierney, 1994; Chan and Geyer, 1994):

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} f(X_t) \;\Rightarrow\; N(m_f,\, \sigma_f^2)\,.$$

In the i.i.d. case, of course $\sigma_f^2 = Var_\pi(f(X_0))$. For general Markov chains, σ_f^2 usually cannot be computed directly, however its estimation can lead to useful error bounds on the MCMC simulation results (see e.g. Roberts and Gilks, 1996). Furthermore, it is known (see e.g. Geyer, 1992; Meyn and Tweedie, 1993; Roberts and Rosenthal, 1997; Jarner and Roberts, 2002) that under suitable regularity conditions

$$\lim_{T\to\infty} Var(\frac{1}{\sqrt{T}} \sum_{i=1}^{T} f(X_i)) = \sigma_f^2$$

regardless of the distribution of X_0.

The quantity $\tau_f \equiv \sigma_f^2 / Var_\pi(f(X_0))$ is known as the *integrated auto-correlation time* for estimating $\mathbf{E}_\pi(f(X))$ using this particular Markov chain. It has the interpretation that a Markov chain sample of length $T\tau_f$ gives asymptotically the same Monte Carlo variance as an i.i.d. sample of size T; that is, we require τ_f dependent samples for every one independent sample. Here $\tau_f = 1$ in the i.i.d. case, while for slowly-mixing Markov chains, τ_f may be quite large.

There are numerous general results giving conditions for a Markov chain to be ergodic or geometrically ergodic; see e.g. Tierney (1994), Smith and Roberts (1993), Schervish and

Carlin (1992), and Baxter and Rosenthal (1995). For example, it is proved by Roberts and Smith (1994) that if π is a continuous, bounded, and positive probability density with respect to Lebesgue measure on $\mathbf{R}^d$, then the Gibbs sampler on π using the standard coordinate directions is ergodic. Furthermore, RWM with continuous and bounded proposal density q is also ergodic provided that q satisfies the property $q(\mathbf{x}) > 0$ for $|\mathbf{x}| \leq \epsilon$, for some $\epsilon > 0$.

4 Burn-in Issues

Classical Monte Carlo simulation produces i.i.d. simulations from the target distribution π, $X_1, \ldots X_T$, and attempts to estimate, say $\mathbf{E}_\pi(f(X))$, using the Monte Carlo estimator $e(f,T) = \sum_{t=1}^{T} f(X_t)/T$. The elementary variance result, $\mathrm{Var}(e(f,T)) = \mathrm{Var}_\pi(f(X))/T$, allows the Monte Carlo experiment to be constructed (i.e., T chosen) in order to satisfy any prescribed accuracy requirements.

Despite the convergence results of Section 3, the situation is rather more complicated for dependent data. For instance, for small values of t, X_t is unlikely to be distributed as (or even similar to) π, so that it makes practical sense to omit the first few iterations of the algorithm when computing appropriate estimates. Therefore we often use the estimator $e_B(f,T) = \frac{1}{T-B}\sum_{t=B+1}^{T} f(X_t)$, where $B \geq 0$ is called the *burn-in* period. If B is too small, then the resulting estimator will be overly influenced by the starting value X_0. On the other hand, if B is too large, e_B will average over too few iterations leading to lost accuracy.

The choice of B is a complex problem. Often B is estimated using *convergence diagnostics*, where the Markov chain output (perhaps starting from multiple initial values X_0) is analysed to determine approximately at what point the resulting distributions become "stable"; see e.g. Gelman and Rubin (1992), Cowles and Carlin (1995), and Brooks and Roberts (1996).

Another approach is to attempt to prove analytically that for appropriate choice of B, the distribution of X_B will be within ϵ of π; see e.g. Meyn and Tweedie (1994), Rosenthal (1995), Roberts and Tweedie (1999), and Douc *et al.* (2001). This approach has had success in various specific examples, but it remains too difficult for widespread routine use.

In practice, the burn-in B is often selected in an *ad hoc* manner. However, as long as B is large enough for the application of interest, this is usually not a problem.

5 Perfect Simulation

Recently, algorithms have been developed which use Markov chains to produce an *exact* sample from π, thus avoiding the burn-in issue entirely. The two main such algorithms are the Coupling from the Past (CFTP) algorithm of Propp and Wilson (1996), and Fill's Markov chain rejection algorithm (Fill, 1998; see also Fill *et al.*, 2000).

To define CFTP, let us assume that we have an ergodic Markov chain $\{X_n\}_{n\in\mathbf{Z}}$ with

transition kernel $P(x, \cdot)$ on a state space $\mathcal{X}$, and a probability measure π on $\mathcal{X}$, such that π is stationary for P (i.e. $(\pi P)(dy) \equiv \int_{\mathcal{X}} \pi(dx)\, P(x, dy) = \pi(dy)$). Let us further assume that we have defined the Markov chain as a *stochastic recursive sequence*, so there is a function $\phi : \mathcal{X} \times \mathbf{R} \to \mathcal{X}$ and an i.i.d. sequence of random variables $\{U_n\}_{n \in \mathbf{Z}}$, such that we always have $X_{n+1} = \phi(X_n, U_n)$.

CFTP involves considering *negative* times n, rather than positive times. Specifically, let

$$\phi^{(n)}(x; u_{-n}, \ldots, u_{-1}) \;=\; \phi(\phi(\phi(\ldots \phi(x,\, u_{-n}),\, u_{-n+1}),\, u_{-n+2}),\, \ldots),\, u_{-1})\,.$$

Then CFTP proceeds by considering various increasing choices of $T > 0$, in the search for a value $T > 0$ such that

$$\phi^{(T)}(x; U_{-T}, \ldots, U_{-1}) \text{ does not depend on } x \in \mathcal{X}\,,$$

i.e. such that the chain has *coalesced* in the time interval from time $-T$ to time 0. (Note that the values $\{U_n\}$ should be thought of as being fixed in advance, even though of course they are only computed as needed. In particular, crucially, all previously-used values of $\{U_n\}$ must be used again, unchanged, as T is increased.)

Once such a T has been found, the resulting value

$$W \equiv \phi^{(T)}(x; U_{-T}, \ldots, U_{-1})$$

(which does not depend on x) is the output of the algorithm. Note in particular that, because of the backward composition implicit in (5), $W = \phi^{(n)}(y; U_{-n}, \ldots, U_{-1})$ for any $n \geq T$ and any $y \in \mathcal{X}$. In particular, letting $n \to \infty$, it follows by ergodicity that $W \sim \pi(\cdot)$. That is, this remarkable algorithm uses the Markov chain to produce a sample W which has density function *exactly* equal to π.

Despite the elegance of perfect simulation methods, and despite their success in certain problems in Spatial Statistics (see e.g. Møller, 1999), it remains difficult to implement perfect simulation in practice. Thus, most applications of MCMC continue to use conventional Markov chain simulation, together with appropriate burn-in periods.

6 Other Recent Developments

MCMC continues to be an active research area in terms of applications methodology and theory. It's impossible to even attempt to describe the diversity of current applications of the technique which extend throughout natural life social and mathematical sciences. In some areas, problem-specific MCMC methodology needs to be developed, though (as stated earlier) in many applications it is remarkable how effective generic techniques such as RWM or the Gibbs sampler can be.

More generally, in Statistical Model Choice problems, it is often necessary to try and construct samplers which can effectively jump between spaces of different dimensionalities

(model spaces), and for this purpose Green (1995) devised *trans-dimensional* algorithms (also called *reversible jump* algorithms). Though such algorithms can be thought of as specific examples of the general Metropolis-Hastings procedure described in Subsection 2.1, great care is required in the construction of suitable 'between-model' jumps. The construction of reliable methods for implementing reversible jump algorithms remains an active and important research area (see for example Brooks *et al.*, 2003).

REFERENCES

J.R. Baxter and J.S. Rosenthal (1995), Rates of convergence for everywhere-positive Markov chains. Stat. Prob. Lett. **22**, 333–338.

M. Bladt, A. Gonzales and S.L. Lauritzen (2003), The estimation of phase-type related functionals through Markov chain Monte Carlo methodology. To appear in Scand. J. Stat.

S.P. Brooks and G.O. Roberts (1996), Diagnosing Convergence of Markov Chain Monte Carlo Algorithms. Technical Report.

S.P. Brooks P. Giudici and G.O. Roberts (2003), Efficient construction of reversible jump MCMC proposal distributions (with discussion). J. Royal Stat. Soc., Series B **65**, 3–56.

M.K. Cowles and B.P. Carlin (1995), Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. J. Amer. Stat. Assoc. **91**, 883–904.

R. Douc, E. Moulines, and J.S. Rosenthal (2002), Quantitative bounds on convergence of time-inhomogeneous Markov Chains. Submitted for publication.

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

J.A. Fill (1998), An interruptible algorithm for perfect sampling via Markov chains. Ann. Appl. Prob. **8**, 131–162.

J.A. Fill, M. Machida, D.J. Murdoch, and J.S. Rosenthal (2000), Extension of Fill's perfect rejection sampling algorithm to general chains. Random Struct. Alg. **17**, 290–316.

A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.

A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci., Vol. **7**, No. **4**, 457–472.

S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.

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

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

S.F. Jarner and G.O. Roberts (2002), Polynomial convergence rates of Markov chains, Ann. Appl. Prob., **12**, 224–247.

M. Jerrum and A. Sinclair (1989), Approximating the permanent. SIAM J. Comput. **18**, 1149–1178.

M. Johannes and N. G. Polson (2002). MCMC methods for Financial Econometrics *Handbook of Financial Econometrics*.

S. Kim, N. Shephard and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models, Rev. Econ. Stud., **65**, 361–393.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

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

S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981–1011.

J. Møller (1999), Perfect simulation of conditionally specified models. J. Royal Stat. Soc., Series B **61**, 251–264.

I. Ntzoufras and P. Dellaportas (2002). Bayesian Prediction of Outstanding Claims. North American Actuarial Journal **6**, 113–136.

E. Nummelin (1984), General irreducible Markov chains and non-negative operators. Cambridge University Press.

G.O. Roberts, A. Gelman, and W.R. Gilks (1997), Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob. **7**, 110–120.

G.O. Roberts and W.R. Gilks (1996), chapter in Gilks *et al.*, above.

G.O. Roberts and N.G. Polson (1994), On the geometric convergence of the Gibbs sampler. J. Royal Stat. Soc. Ser. B, 377–384.

G.O. Roberts and J.S. Rosenthal (1997), Geometric ergodicity and hybrid Markov chains. Electronic Comm. Prob. **2**, Paper no. 2, 13–25.

G.O. Roberts and J.S. Rosenthal (1998a), Optimal scaling of discrete approximations to Langevin diffusions. J. Roy. Stat. Soc. Ser. B **60**, 255–268.

G.O. Roberts and J.S. Rosenthal (1998b), Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Canadian Journal of Statistics* **26**, 5–31.

G.O. Roberts and J.S. Rosenthal (1998c), Two convergence properties of hybrid samplers. Ann. Appl. Prob. **8**, 397–407.

G.O. Roberts and J.S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**, 351–367.

G.O. Roberts and S.K. Sahu (1997) Updating schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler. J. Roy. Statis. Soc., B, **59**, 291–317.

J.S. Rosenthal (1995), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

J.S. Rosenthal (2001), A review of asymptotic convergence for general state space Markov chains. *Far East Journal of Theoretical Statistics* **5** (2001), 37–50.

P.J. Rossky, J.D. Doll, and H.L. Friedman (1978), Brownian dynamics as smart Monte Carlo simulation. J. Chem. Phys. **69**, 4628–4633.

M.J. Schervish and B.P. Carlin (1992), On the convergence of successive substitution sampling, J. Comp. Graph. Stat. **1**, 111–127.

D.P.M. Scollnik (2001). Actuarial Modeling with MCMC and BUGS, North American Actuarial Journal, **5**, 96–124.

A.F.M. Smith and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3–24.

M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528–550.

L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.