Chapter 4
Markov Chain Monte Carlo Technology

Siddhartha Chib

4.1 Introduction

In the past fifteen years computational statistics has been enriched by a powerful, somewhat abstract method of generating variates from a target probability distribution that is based on Markov chains whose stationary distribution is the probability distribution of interest. This class of methods, popularly referred to as Markov chain Monte Carlo methods, or simply MCMC methods, have been influential in the modern practice of Bayesian statistics where these methods are used to summarize the posterior distributions that arise in the context of the Bayesian prior-posterior analysis (Besag et al. 1995; Chib and Greenberg 1995; Gelfand and Smith 1990; Smith and Roberts 1993; Tanner and Wong 1987; Tierney 1994, 1996; Carlin and Louis 2000; Chen et al. 2000; Chib 2001; Congdon 2001; Gammerman 1997; Gelman et al. 2003; Gilks et al. 1996; Liu 2001; Robert 2001; Robert and Casella 1999; Tanner 1996). MCMC methods have proved useful in practically all aspects of Bayesian inference, for example, in the context of prediction problems and in the computation of quantities, such as the marginal likelihood, that are used for comparing competing Bayesian models.

A central reason for the widespread interest in MCMC methods is that these methods are extremely general and versatile and can be used to sample univariate and multivariate distributions when other methods (for example classical methods that produce independent and identically distributed draws) either fail or are difficult to implement. The fact that MCMC methods produce dependent draws causes no substantive complications in summarizing the target distribution. For example, if $\{\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)}\}$ are draws from a (say continuous) target distribution $\pi(\boldsymbol{\psi})$, where $\boldsymbol{\psi} \in \Re^d$, then the expectation of $h(\boldsymbol{\psi})$ under π can be estimated by the average

S. Chib (✉)
Olin Business School, Washington University in St. Louis
St. Louis, MO 63130, USA
e-mail: chib@wustl.edu

J.E. Gentle et al. (eds.), *Handbook of Computational Statistics*, Springer Handbooks of Computational Statistics, DOI 10.1007/978-3-642-21551-3_4,

$$M^{-1}\sum_{j=1}^{M} h\left(\boldsymbol{\psi}^{(j)}\right) \ , \tag{4.1}$$

as in the case of random samples, because suitable laws of large numbers for Markov chains can be used to show that

$$M^{-1}\sum_{j=1}^{M} h\left(\boldsymbol{\psi}^{(j)}\right) \rightarrow \int_{\Re^d} h(\boldsymbol{\psi})\pi(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\psi} \ ,$$

as the simulation sample size M becomes large.

Another reason for the interest in MCMC methods is that, somewhat surprisingly, it is rather straightforward to construct one or more Markov chains whose limiting invariant distribution is the desired target distribution. One way to construct the appropriate Markov chain is by a method called the Metropolis method which was introduced by Metropolis et al. (1953) in connection with work related to the hydrogen bomb project. In this method, the Markov chain simulation is constructed by a recursive two step process. Given the current iterate $\boldsymbol{\psi}^{(j)}$, a proposal value $\boldsymbol{\psi}'$ is drawn from a distribution $q(\boldsymbol{\psi}^{(j)}, \cdot)$, such that $\boldsymbol{\psi}'$ is symmetrically distributed about the current value $\boldsymbol{\psi}^{(j)}$. In the second step, this proposal value is accepted as the next iterate $\boldsymbol{\psi}^{(j+1)}$ of the Markov chain with probability

$$\alpha\left(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}'\right) = \min\left\{1, \frac{\pi(\boldsymbol{\psi}')}{\pi\left(\boldsymbol{\psi}^{(j)}\right)}\right\} \ .$$

If the proposal value is rejected, then $\boldsymbol{\psi}^{(j+1)}$ is taken to be the current value. The method is simple to implement, even in multivariate settings, and was widely used by physicists in computational statistical mechanics and quantum field theory to sample the coordinates of a point in phase space. In those settings, and in subsequent statistical problems, it is helpful that the method can be implemented without knowledge of the normalizing constant of π since that constant cancels in the determination of the probability $\alpha(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}')$.

The requirement that the proposal distribution be symmetric about the current value was relaxed by Hastings (1970). The resulting method, commonly called the Metropolis–Hastings (M–H) method, relies on the same two steps of the Metropolis method except that the probability of move is given by

$$\alpha\left(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}'\right) = \min\left\{1, \frac{\pi\left(\boldsymbol{\psi}'\right)}{\pi\left(\boldsymbol{\psi}^{(j)}\right)}\frac{q\left(\boldsymbol{\psi}', \boldsymbol{\psi}^{(j)}\right)}{q\left(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}'\right)}\right\}$$

which clearly reduces to the Metropolis probability of move when the proposal distribution is symmetric in its arguments. Starting with an arbitrary value $\boldsymbol{\psi}^{(0)}$ in the support of the target distributions, iterations of this two step process produce the

(correlated) sequence of values

$$\{\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots\} \ .$$

Typically, a certain number of values (say n_0) at the start of this sequence are discarded and the subsequent (say) M values are used as variates from the target distribution.

In applications when the dimension of $\boldsymbol{\psi}$ is large it may be preferable to construct the Markov chain simulation by first grouping the variables $\boldsymbol{\psi}$ into smaller blocks. For simplicity suppose that two blocks are adequate and that $\boldsymbol{\psi}$ is written as $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$, with $\boldsymbol{\psi}_k \in \Omega_k \subseteq \Re^{d_k}$. In that case, the M–H chain can be constructed by:

- Updating $\boldsymbol{\psi}_1$ given $\left(\boldsymbol{\psi}_1^{(j)}, \boldsymbol{\psi}_2^{(j)}\right)$ to produce $\boldsymbol{\psi}_1^{(j)}$ and then
- Updating $\boldsymbol{\psi}_2$ given $\left(\boldsymbol{\psi}_1^{(j+1)}, \boldsymbol{\psi}_2^{(j)}\right)$ to produce $\boldsymbol{\psi}_2^{(j+1)}$,

which completes one cycle through two sub-moves. Chib and Greenberg (1995) who emphasized and highlighted such M–H chains have referred to them as multiple-block M–H algorithms.

Despite the long vintage of the M–H method, the contemporary interest in MCMC methods was sparked by work on a related MCMC method, the Gibbs sampling algorithm. The Gibbs sampling algorithm is one of the simplest Markov chain Monte Carlo algorithms and has its origins in the work of Besag (1974) on spatial lattice systems, Geman and Geman (1984) on the problem of image processing, and Tanner and Wong (1987) on missing data problems. The paper by Gelfand and Smith (1990) helped to demonstrate the value of the Gibbs algorithm for a range of problems in Bayesian analysis. In the Gibbs sampling method, the Markov chain is constructed by simulating the conditional distributions that are implied by $\pi(\boldsymbol{\psi})$. In particular, if $\boldsymbol{\psi}$ is split into two components $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, then the Gibbs method proceeds through the recursive sampling of the conditional distributions $\pi(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)$ and $\pi(\boldsymbol{\psi}_2|\boldsymbol{\psi}_1)$, where the most recent value of $\boldsymbol{\psi}_2$ is used in the first simulation and the most recent value of $\boldsymbol{\psi}_1$ in the second simulation. This method is most simple to implement when each conditional distribution is a known distribution that is easy to sample. As we show below, the Gibbs sampling method is a special case of the multiple block M–H algorithm.

4.1.1 Organization

The rest of the chapter is organized as follows. In Sect. 4.2 we summarize the relevant Markov chain theory that justifies simulation by MCMC methods. In particular, we provide the conditions under which discrete-time and continuous state space Markov chains satisfy a law of large numbers and a central limit theorem. The M–H algorithm is discussed in Sect. 4.3 followed by the Gibbs sampling algorithm

in Sect. 4.4. Section 4.5 deals with MCMC methods with latent variables and Sect. 4.6 with ways of estimating the marginal densities based on the MCMC output. Issues related to sampler performance are considered in Sect. 4.7 and strategies for improving the mixing of the Markov chains in Sect. 4.8. Section 4.9 concludes with brief comments about new and emerging directions in MCMC methods.

4.2 Markov Chains

Markov chain Monte Carlo is a method to sample a given multivariate distribution π^* by constructing a suitable Markov chain with the property that its limiting, invariant distribution, is the target distribution π^*. In most problems of interest, the distribution π^* is absolutely continuous and, as a result, the theory of MCMC methods is based on that of Markov chains on continuous state spaces outlined, for example, in Nummelin (1984) and Meyn and Tweedie (1993). Tierney (1994) is the fundamental reference for drawing the connections between this elaborate Markov chain theory and MCMC methods. Basically, the goal of the analysis is to specify conditions under which the constructed Markov chain converges to the invariant distribution, and conditions under which sample path averages based on the output of the Markov chain satisfy a law of large numbers and a central limit theorem.

4.2.1 Definitions and Results

A Markov chain is a collection of random variables (or vectors) $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_i : i \in T\}$ where $T = \{0, 1, 2, \ldots\}$. The evolution of the Markov chain on a space $\Omega \subseteq \Re^p$ is governed by the transition kernel

$$\begin{aligned} P(\boldsymbol{x}, A) &\equiv \Pr(\boldsymbol{\Phi}_{i+1} \in A | \boldsymbol{\Phi}_i = \boldsymbol{x}, \boldsymbol{\Phi}_j,\ j < i) \\ &= \Pr(\boldsymbol{\Phi}_{i+1} \in A | \boldsymbol{\Phi}_i = \boldsymbol{x}) \,, \quad \boldsymbol{x} \in \Omega \,, \quad A \subset \Omega \,, \end{aligned}$$

where the second line embodies the Markov property that the distribution of each succeeding state in the sequence, given the current and the past states, depends only on the current state.

Generally, the transition kernel in Markov chain simulations has both a continuous and discrete component. For some function $p(\boldsymbol{x}, \boldsymbol{y}) : \Omega \times \Omega \to \Re^+$, the kernel can be expressed as

$$P(\boldsymbol{x}, d\boldsymbol{y}) = p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} + r(\boldsymbol{x}) \delta_{\boldsymbol{x}}(d\boldsymbol{y}) \,, \tag{4.2}$$

where $p(\boldsymbol{x}, \boldsymbol{x}) = 0$, $\delta_{\boldsymbol{x}}(d\boldsymbol{y}) = 1$ if $\boldsymbol{x} \in d\boldsymbol{y}$ and 0 otherwise, $r(\boldsymbol{x}) = 1 - \int_\Omega p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y}$. This transition kernel specifies that transitions from $\boldsymbol{x}$ to $\boldsymbol{y}$ occur according to $p(\boldsymbol{x}, \boldsymbol{y})$ and transitions from $\boldsymbol{x}$ to $\boldsymbol{x}$ occur with probability $r(\boldsymbol{x})$.

The transition kernel is thus the distribution of $\boldsymbol{\Phi}_{i+1}$ given that $\boldsymbol{\Phi}_i = \boldsymbol{x}$. The nth step ahead transition kernel is given by

$$P^{(n)}(\boldsymbol{x}, A) = \int_{\Omega} P(\boldsymbol{x}, d\,\boldsymbol{y})\, P^{(n-1)}(\boldsymbol{y}, A) \ ,$$

where $P^{(1)}(\boldsymbol{x}, d\,\boldsymbol{y}) = P(\boldsymbol{x}, d\,\boldsymbol{y})$ and

$$P(\boldsymbol{x}, A) = \int_{A} P(\boldsymbol{x}, d\,\boldsymbol{y}) \ . \tag{4.3}$$

The goal is to find conditions under which the nth iterate of the transition kernel converges to the invariant distribution π^* as $n \to \infty$. The invariant distribution is one that satisfies

$$\pi^*(d\,\boldsymbol{y}) = \int_{\Omega} P(\boldsymbol{x}, d\,\boldsymbol{y})\pi(\boldsymbol{x})\, d\,\boldsymbol{x} \ , \tag{4.4}$$

where π is the density of π^* with respect to the Lebesgue measure. The invariance condition states that if $\boldsymbol{\Phi}_i$ is distributed according to π^*, then all subsequent elements of the chain are also distributed as π^*. Markov chain samplers are invariant by construction and therefore the existence of the invariant distribution does not have to be checked.

A Markov chain is reversible if the function $p(\boldsymbol{x}, \boldsymbol{y})$ in (4.2) satisfies

$$f(\boldsymbol{x})p(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{y})p(\boldsymbol{y}, \boldsymbol{x}) \ , \tag{4.5}$$

for a density $f(\cdot)$. If this condition holds, it can be shown that $f(\cdot) = \pi(\cdot)$ and has π^* as an invariant distribution (Tierney 1994). To verify this we evaluate the right hand side of (4.4):

$$\begin{aligned}
\int P(\boldsymbol{x}, A)\pi(\boldsymbol{x})\, d\,\boldsymbol{x} &= \int \left\{ \int_A p(\boldsymbol{x}, \boldsymbol{y})\, d\,\boldsymbol{y} \right\} \pi(\boldsymbol{x})\, d\,\boldsymbol{x} + \int r(\boldsymbol{x})\delta_{\boldsymbol{x}}(A)\pi(\boldsymbol{x})\, d\,\boldsymbol{x} \\
&= \int_A \left\{ \int p(\boldsymbol{x}, \boldsymbol{y})\pi(\boldsymbol{x})\, d\,\boldsymbol{x} \right\} d\,\boldsymbol{y} + \int_A r(\boldsymbol{x})\pi(\boldsymbol{x})\, d\,\boldsymbol{x} \\
&= \int_A \left\{ \int p(\boldsymbol{y}, \boldsymbol{x})\pi(\boldsymbol{y})\, d\,\boldsymbol{x} \right\} d\,\boldsymbol{y} + \int_A r(\boldsymbol{x})\pi(\boldsymbol{x})\, d\,\boldsymbol{x} \\
&= \int_A (1 - r(\boldsymbol{y}))\pi(\boldsymbol{y})\, d\,\boldsymbol{y} + \int_A r(\boldsymbol{x})\pi(\boldsymbol{x})\, d\,\boldsymbol{x} \\
&= \int_A \pi(\boldsymbol{y})\, d\,\boldsymbol{y} \ .
\end{aligned} \tag{4.6}$$

A minimal requirement on the Markov chain for it to satisfy a law of large numbers is the requirement of π^*-irreducibility. This means that the chain is able to visit all sets with positive probability under π^* from any starting point in Ω. Formally, a Markov chain is said to be π^*-irreducible if for every $\boldsymbol{x} \in \Omega$,

$$\pi^*(A) > 0 \Rightarrow P(\boldsymbol{\Phi}_i \in A | \boldsymbol{\Phi}_0 = \boldsymbol{x}) > 0$$

for some $i \geq 1$. If the space Ω is connected and the function $p(\boldsymbol{x}, \boldsymbol{y})$ is positive and continuous, then the Markov chain with transition kernel given by (4.3) and invariant distribution π^* is π^*-irreducible.

Another important property of a chain is aperiodicity, which ensures that the chain does not cycle through a finite number of sets. A Markov chain is aperiodic if there exists no partition of $\Omega = (D_0, D_1, \ldots, D_{p-1})$ for some $p \geq 2$ such that $P(\boldsymbol{\Phi}^i \in D_{i \bmod (p)} | \boldsymbol{\Phi}_0 \in D_0) = 1$ for all i.

These definitions allow us to state the following results from Tierney (1994) which form the basis for Markov chain Monte Carlo methods. The first of these results gives conditions under which a strong law of large numbers holds and the second gives conditions under which the probability density of the Mth iterate of the Markov chain converges to its unique, invariant density.

Theorem 1. *Suppose $\{\boldsymbol{\Phi}_i\}$ is a π^*-irreducible Markov chain with transition kernel $P(\cdot, \cdot)$ and invariant distribution π^*, then π^* is the unique invariant distribution of $P(\cdot, \cdot)$ and for all π^*-integrable real-valued functions h,*

$$\frac{1}{M} \sum_{i=1}^{M} h(\boldsymbol{\Phi}_i) \rightarrow \int h(\boldsymbol{x}) \pi(\boldsymbol{x}) d\boldsymbol{x} \quad as \quad M \rightarrow \infty \ , \ a.s.$$

Theorem 2. *Suppose $\{\boldsymbol{\Phi}_i\}$ is a π^*-irreducible, aperiodic Markov chain with transition kernel $P(\cdot, \cdot)$ and invariant distribution π^*. Then for π^*-almost every $\boldsymbol{x} \in \Omega$, and all sets A*

$$\parallel P^M(\boldsymbol{x}, A) - \pi^*(A) \parallel \rightarrow 0 \quad as \quad M \rightarrow \infty \ ,$$

where $\parallel \cdot \parallel$ denotes the total variation distance.

A further strengthening of the conditions is required to obtain a central limit theorem for sample-path averages. A key requirement is that of an ergodic chain, i.e., chains that are irreducible, aperiodic and positive Harris-recurrent (for a definition of the latter, see Tierney (1994)). In addition, one needs the notion of geometric ergodicity. An ergodic Markov chain with invariant distribution π^* is a geometrically ergodic if there exists a non-negative real-valued function (bounded in expectation under π^*) and a positive constant $r < 1$ such that

$$\parallel P^M(\boldsymbol{x}, A) - \pi^*(A) \parallel \leq C(\boldsymbol{x}) r^n$$

for all $\boldsymbol{x}$ and all n and sets A. Chan and Ledolter (1995) show that if the Markov chain is ergodic, has invariant distribution π^*, and is geometrically ergodic, then for all L^2 measurable functions h, taken to be scalar-valued for simplicity, and any initial distribution, the distribution of $\sqrt{M}(\hat{h}_M - \mathrm{E}h)$ converges weakly to a normal distribution with mean zero and variance $\sigma_h^2 \geq 0$, where

$$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i)$$

$$\mathrm{E}h = \int h(\boldsymbol{\Phi})\pi(\boldsymbol{\Phi})d\boldsymbol{\Phi}$$

and

$$\sigma_h^2 = \mathrm{Var}\, h(\boldsymbol{\Phi}_0) + 2\sum_{k=1}^{\infty} \mathrm{Cov}\ [\{h(\boldsymbol{\Phi}_0), h(\boldsymbol{\Phi}_k)\}]\ . \tag{4.7}$$

4.2.2 *Computation of Numerical Accuracy and Inefficiency Factor*

The square root of σ_h^2 is the numerical standard error of $\hat{h}_M$. To describe estimators of σ_h^2 that are consistent in M, let $Z_i = h(\boldsymbol{\Phi}_i)$ $(i \leq M)$. Then, due to the fact that $\{Z_i\}$ is a dependent sequence

$$\begin{aligned}
\mathrm{Var}\left(\hat{h}_M\right) &= M^{-2}\sum_{j,k} \mathrm{Cov}\,(Z_j, Z_k) \\
&= s^2 M^{-2}\sum_{j,k=1}^{M} \rho_{|j-k|} \\
&= s^2 M^{-1}\left\{1 + 2\sum_{s=1}^{M}\left(1 - \frac{s}{M}\right)\rho_s\right\}\ ,
\end{aligned}$$

where s^2 is the sample variance of $\{Z_i\}$ and ρ_s is the estimated autocorrelation at lag s (see Ripley 1987, Chap. 6). If $\rho_s > 0$ for each s, then this variance is larger than s^2/M which is the variance under independence. Another estimate of the variance can be found by consistently estimating the spectral density f of $\{Z_i\}$ at frequency zero and using the fact that $\mathrm{Var}\,(\hat{h}_M) = \tau^2/M$, where $\tau^2 = 2\pi f(0)$. Finally, a traditional approach to finding the variance is by the method of batch means. In this approach, the data $(Z_1, \ldots, Z_M)$ is divided into k batches of length m with means $B_i = m^{-1}[Z_{(i-1)m+1} + \ldots + Z_{im}]$ and the variance of $\hat{h}_M$ estimated as

$$\text{Var}\left(\hat{h}_M\right) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \left(B_i - \bar{B}\right)^2 , \tag{4.8}$$

where the batch size m is chosen to ensure that the first order serial correlation of the batch means is less than 0.05.

Given the numerical variance it is common to calculate the inefficiency factor, which is also called the autocorrelation time, defined as

$$\kappa_{\hat{h}} = \frac{\text{Var}\left(\hat{h}_M\right)}{s^2/M} . \tag{4.9}$$

This quantity is interpreted as the ratio of the numerical variance of $\hat{h}_M$ to the variance of $\hat{h}_M$ based on independent draws, and its inverse is the relative numerical efficiency defined in Geweke (1992). Because independence sampling produces an autocorrelation time that is theoretically equal to one and Markov chain sampling produces autocorrelation times that are bigger than one, the inefficiency factor serves to quantify the relative efficiency loss in the computation of $\hat{h}_M$ from correlated versus independent samples.

4.3 Metropolis–Hastings Algorithm

This powerful algorithm provides a general approach for producing a correlated sequence of draws from the target density that may be difficult to sample by a classical independence method. The goal is to simulate the d-dimensional distribution $\pi^*(\boldsymbol{\psi})$, $\boldsymbol{\psi} \in \Psi \subseteq \Re^d$ that has density $\pi(\boldsymbol{\psi})$ with respect to some dominating measure. To define the algorithm, let $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ denote a source density for a candidate draw $\boldsymbol{\psi}'$ given the current value ψ in the sampled sequence. The density $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is referred to as the proposal or candidate generating density. Then, the M–H algorithm is defined by two steps: a first step in which a proposal value is drawn from the candidate generating density and a second step in which the proposal value is accepted as the next iterate in the Markov chain according to the probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$, where

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \begin{cases} \min\left[\dfrac{\pi(\boldsymbol{\psi}')q(\boldsymbol{\psi}', \boldsymbol{\psi})}{\pi(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}')}, 1\right] & \text{if} \quad \pi(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}') > 0 ; \\ 1 & \text{otherwise} . \end{cases} \tag{4.10}$$

If the proposal value is rejected, then the next sampled value is taken to be the current value. In algorithmic form, the simulated values are obtained by the following recursive procedure.

Algorithm 1 Metropolis–Hastings

1. Specify an initial value $\boldsymbol{\psi}^{(0)}$:
2. Repeat for $j = 1, 2, \ldots, M$.

   (a) Propose
   $$\boldsymbol{\psi}' \sim q\left(\boldsymbol{\psi}^{(j)}, \cdot\right)$$

   (b) Let
   $$\boldsymbol{\psi}^{(j+1)} = \begin{cases} \boldsymbol{\psi}' & \text{if} \quad Unif(0,1) \leq \alpha\left(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}'\right) \ ; \\ \boldsymbol{\psi}^{(j)} & \text{otherwise} \ . \end{cases}$$

3. Return the values $\left\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\right\}$.

Typically, a certain number of values (say n_0) at the start of this sequence are discarded after which the chain is assumed to have converged to it invariant distribution and the subsequent draws are taken as approximate variates from π. Because theoretical calculation of the burn-in is not easy it is important that the proposal density is chosen to ensure that the chain makes large moves through the support of the invariant distribution without staying at one place for many iterations. Generally, the empirical behavior of the M–H output is monitored by the autocorrelation time of each component of $\boldsymbol{\psi}$ and by the acceptance rate, which is the proportion of times a move is made as the sampling proceeds.

One should observe that the target density appears as a ratio in the probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and therefore the algorithm can be implemented without knowledge of the normalizing constant of $\pi(\cdot)$. Furthermore, if the candidate-generating density is symmetric, i.e. $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}', \boldsymbol{\psi})$, the acceptance probability only contains the ratio $\pi(\boldsymbol{\psi}')/\pi(\boldsymbol{\psi})$; hence, if $\pi(\boldsymbol{\psi}') \geq \pi(\boldsymbol{\psi})$, the chain moves to $\boldsymbol{\psi}'$, otherwise it moves with probability given by $\pi(\boldsymbol{\psi}')/\pi(\boldsymbol{\psi})$. The latter is the algorithm originally proposed by Metropolis et al. (1953). This version of the algorithm is illustrated in Fig. 4.1.

Different proposal densities give rise to specific versions of the M–H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}' - \boldsymbol{\psi})$. The candidate $\boldsymbol{\psi}'$ is thus drawn according to the process $\boldsymbol{\psi}' = \boldsymbol{\psi} + z$, where z follows the distribution q. Since the candidate is equal to the current value plus noise, this case is called a random walk M–H chain. Possible choices for q include the multivariate normal density and the multivariate-t. The random walk M–H chain is perhaps the simplest version of the M–H algorithm (and was the one used by Metropolis et al. 1953) and popular in applications. One has to be careful, however, in setting the variance of z; if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution. In both cases the generated draws that will be highly serially correlated. Note that when q is symmetric, $q(z) = q(-z)$ and the probability of move only contains the

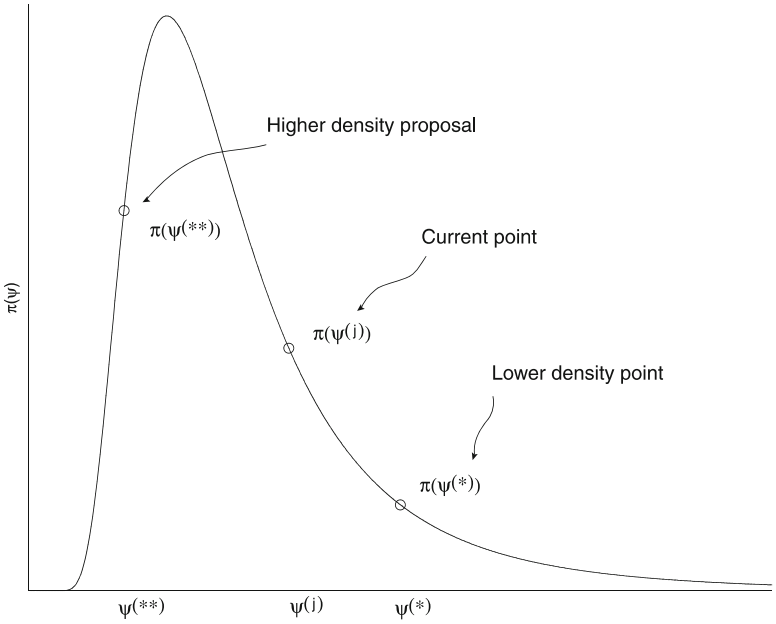

Fig. 4.1 Original Metropolis algorithm: higher density proposal is accepted with probability one and the lower density proposal with probability α

ratio $\pi(\boldsymbol{\psi}')/\pi(\boldsymbol{\psi})$. As mentioned earlier, the same reduction occurs if $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}', \boldsymbol{\psi})$.

Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{y})$. Tierney (1994) refers to this as an independence M–H chain because, in contrast to the random walk chain, the candidates are drawn independently of the current location $\boldsymbol{\psi}$. In this case, the probability of move becomes

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \min\left\{\frac{w(\boldsymbol{\psi}')}{w(\boldsymbol{\psi})}, 1\right\} \ ;$$

where $w(\boldsymbol{\psi}) = \pi(\boldsymbol{\psi})/q(\boldsymbol{\psi})$ is the ratio of the target and proposal densities. For this method to work and not get stuck in the tails of π, it is important that the proposal density have thicker tails than π. A similar requirement is placed on the importance sampling function in the method of importance sampling (Geweke 1989). In fact, Mengersen and Tweedie (1996) show that if $w(\boldsymbol{\psi})$ is uniformly bounded then the resulting Markov chain is ergodic.

Chib and Greenberg (1994, 1995) discuss a way of formulating proposal densities in the context of time series autoregressive-moving average models that has a bearing on the choice of proposal density for the independence M–H chain. They suggest matching the proposal density to the target at the mode by a multivariate normal or

multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode. Specifically, the parameters of the proposal density are taken to be

$$\begin{aligned} \boldsymbol{m} &= \arg\max \log \pi(\boldsymbol{\psi}) \quad \text{and} \\ \boldsymbol{V} &= \tau \left\{ -\frac{\partial^2 \log \pi(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\, \partial \boldsymbol{\psi}'} \right\}^{-1}_{\boldsymbol{\psi}=\hat{\psi}} , \end{aligned} \tag{4.11}$$

where τ is a tuning parameter that is adjusted to control the acceptance rate. The proposal density is then specified as $q(\boldsymbol{\psi}') = f(\boldsymbol{\psi}'|\boldsymbol{m}, \boldsymbol{V})$, where f is some multivariate density. This may be called a tailored M–H chain.

Another way to generate proposal values is through a Markov chain version of the accept-reject method. In this version, due to Tierney (1994), and considered in detail by Chib and Greenberg (1995), a pseudo accept-reject step is used to generate candidates for an M–H algorithm. Suppose $c > 0$ is a known constant and $h(\boldsymbol{\psi})$ a source density. Let $C = \{\boldsymbol{\psi} : \pi(\boldsymbol{\psi}) \leq ch(\boldsymbol{\psi})\}$ denote the set of value for which $ch(\boldsymbol{\psi})$ dominates the target density and assume that this set has high probability under π^*. Given $\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}$, the next value $\boldsymbol{\psi}^{(n+1)}$ is obtained as follows: First, a candidate value $\boldsymbol{\psi}'$ is obtained, independent of the current value $\boldsymbol{\psi}$, by applying the accept-reject algorithm with $ch(\cdot)$ as the "pseudo dominating" density. The candidates $\boldsymbol{\psi}'$ that are produced under this scheme have density $q(\boldsymbol{\psi}') \propto \min\{\pi(\boldsymbol{\psi}'), ch(\boldsymbol{\psi}')\}$. If we let $w(\boldsymbol{\psi}) = c^{-1}\pi(\boldsymbol{\psi})/h(\boldsymbol{\psi})$ then it can be shown that the M–H probability of move is given by

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \begin{cases} 1 & \text{if} \quad \boldsymbol{\psi} \in C \\ 1/w(\boldsymbol{\psi}) & \text{if} \quad \boldsymbol{\psi} \notin C\,, \quad \boldsymbol{\psi}' \in C \\ \min\{w(\boldsymbol{\psi}')/w(\boldsymbol{\psi}), 1\} & \text{if} \quad \boldsymbol{\psi} \notin C\,, \quad \boldsymbol{\psi}' \notin C \end{cases} . \tag{4.12}$$

4.3.1 Convergence Results

In the M–H algorithm the transition kernel of the chain is given by

$$P(\boldsymbol{\psi}, d\boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\, d\boldsymbol{\psi}' + r(\boldsymbol{\psi})\delta_{\boldsymbol{\psi}}(d\boldsymbol{\psi}') , \tag{4.13}$$

where $\delta_{\boldsymbol{\psi}}(d\boldsymbol{\psi}') = 1$ if $\boldsymbol{\psi} \in d\boldsymbol{\psi}'$ and 0 otherwise and

$$r(\boldsymbol{\psi}) = 1 - \int_{\Omega} q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\, d\boldsymbol{\psi}' .$$

Thus, transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ ($\boldsymbol{\psi}' \neq \boldsymbol{\psi}$) are made according to the density

$$p(\boldsymbol{\psi}, \boldsymbol{\psi}') \equiv q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}'), \quad \boldsymbol{\psi} \neq \boldsymbol{\psi}'$$

while transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}$ occur with probability $r(\boldsymbol{\psi})$. In other words, the density function implied by this transition kernel is of mixed type,

$$K(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') + r(\boldsymbol{\psi})\delta_{\boldsymbol{\psi}}(\boldsymbol{\psi}') \,, \tag{4.14}$$

having both a continuous and discrete component, where now, with change of notation, $\delta_{\boldsymbol{\psi}}(\boldsymbol{\psi}')$ is the Dirac delta function defined as $\delta_{\boldsymbol{\psi}}(\boldsymbol{\psi}') = 0$ for $\boldsymbol{\psi}' \neq \boldsymbol{\psi}$ and $\int_{\Omega} \delta_{\boldsymbol{\psi}}(\boldsymbol{\psi}')d\boldsymbol{\psi}' = 1$.

Chib and Greenberg (1995) provide a way to derive and interpret the probability of move $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$. Consider the proposal density $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$. This proposal density q is not likely to be reversible for π (if it were then we would be done and M–H sampling would not be necessary). Without loss of generality, suppose that $\pi(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}') > \pi(\boldsymbol{\psi}')q(\boldsymbol{\psi}', \boldsymbol{\psi})$ implying that the rate of transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ exceed those in the reverse direction. To reduce the transitions from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ one can introduce a function $0 \leq \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') \leq 1$ such that $\pi(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \pi(\boldsymbol{\psi}')q(\boldsymbol{\psi}', \boldsymbol{\psi})$. Solving for $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ yields the probability of move in the M–H algorithm. This calculation reveals the important point that the function $p(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is reversible by construction, i.e., it satisfies the condition

$$q(\boldsymbol{\psi}, \boldsymbol{\psi}')\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\pi(\boldsymbol{\psi}) = q(\boldsymbol{\psi}', \boldsymbol{\psi})\alpha(\boldsymbol{\psi}', \boldsymbol{\psi})\pi(\boldsymbol{\psi}') \,. \tag{4.15}$$

It immediately follows, therefore, from the argument in (4.6) that the M–H kernel has $\pi(\boldsymbol{\psi})$ as its invariant density.

It is not difficult to provide conditions under which the Markov chain generated by the M–H algorithm satisfies the conditions of Propositions 1–2. The conditions of Proposition 1 are satisfied by the M–H chain if $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is positive for $(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and continuous and the set $\boldsymbol{\psi}$ is connected. In addition, the conditions of Proposition 2 are satisfied if q is not reversible (which is the usual situation) which leads to a chain that is aperiodic. Conditions for ergodicity, required for use of the central limit theorem, are satisfied if in addition π is bounded. Other similar conditions are provided by Robert and Casella (1999).

4.3.2 Example

To illustrate the M–H algorithm, consider the binary response data in Table 4.1, taken from Fahrmeir and Tutz (1997), on the occurrence or non-occurrence of infection following birth by caesarean section. The response variable y is one if the caesarean birth resulted in an infection, and zero if not. There are three covariates: x_1, an indicator of whether the caesarean was non-planned; x_2, an indicator of whether risk factors were present at the time of birth and x_3, an indicator of whether

Table 4.1 Caesarean infection data

Y (1/0)	x_1	x_2	x_3
11/87	1	1	1
1/17	0	1	1
0/2	0	0	1
23/3	1	1	0
28/30	0	1	0
0/9	1	0	0
8/32	0	0	0

antibiotics were given as a prophylaxis. The data in the table contains information from 251 births. Under the column of the response, an entry such as 11/87 means that there were 98 deliveries with covariates $(1, 1, 1)$ of whom 11 developed an infection and 87 did not.

Suppose that the probability of infection for the ith birth ($i \leq 251$) is

$$\Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}) \,, \tag{4.16}$$

$$\boldsymbol{\beta} \sim N_4(0, 5\boldsymbol{I}_4) \,, \tag{4.17}$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^\top$ is the covariate vector, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ is the vector of unknown coefficients, Φ is the cdf of the standard normal random variable and $\boldsymbol{I}_4$ is the four-dimensional identity matrix. The target posterior density, under the assumption that the outcomes $\boldsymbol{y} = (y_1, y_2, \ldots, y_{251})$ are conditionally independent, is

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{251} \Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^{y_i} \left\{1 - \Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right\}^{(1-y_i)} \,,$$

where $\pi(\boldsymbol{\beta})$ is the density of the $N(0, 10\boldsymbol{I}_4)$ distribution.

Random Walk Proposal Density

To define the proposal density, let

$$\hat{\boldsymbol{\beta}} = (-1.093022 \;\; 0.607643 \;\; 1.197543 \;\; -1.904739)^\top$$

be the MLE found using the Newton–Raphson algorithm and let

$$\boldsymbol{V} = \begin{pmatrix} 0.040745 & -0.007038 & -0.039399 & 0.004829 \\ & 0.073101 & -0.006940 & -0.050162 \\ & & 0.062292 & -0.016803 \\ & & & 0.080788 \end{pmatrix}$$

Table 4.2 Caesarean data: Prior-posterior summary based on 5000 draws (beyond a burn-in of 100 cycles) from the random-walk M–H algorithm

	Prior		Posterior			
	Mean	Std dev	Mean	Std dev	Lower	Upper
β_0	0.000	3.162	−1.110	0.224	−1.553	−0.677
β_1	0.000	3.162	0.612	0.254	0.116	1.127
β_2	0.000	3.162	1.198	0.263	0.689	1.725
β_3	0.000	3.162	−1.901	0.275	−2.477	−1.354

be the symmetric matrix obtained by inverting the negative of the Hessian matrix (the matrix of second derivatives) of the log-likelihood function evaluated at $\hat{\boldsymbol{\beta}}$. Now generate the proposal values by the random walk:

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{\beta}^{(j-1)} + \boldsymbol{\varepsilon}^{(j)} \\ \boldsymbol{\varepsilon}^{(j)} &\sim \mathrm{N}_4(\mathbf{0}, \boldsymbol{V}) \ , \end{aligned} \tag{4.18}$$

which leads to the original Metropolis method. From a run of 5000 iterations of the algorithm beyond a burn-in of a 100 iterations we get the prior-posterior summary that is reported in Table 4.2, which contains the first two moments of the prior and posterior and the 2.5th (lower) and 97.5th (upper) percentiles of the marginal densities of $\boldsymbol{\beta}$.

As expected, both the first and second covariates increase the probability of infection while the third covariate (the antibiotics prophylaxis) reduces the probability of infection.

To get an idea of the form of the posterior density we plot in Fig. 4.1 the four marginal posterior densities. The density plots are obtained by smoothing the histogram of the simulated values with a Gaussian kernel. In the same plot we also report the autocorrelation functions (correlation against lag) for each of the sampled parameter values. The autocorrelation plots provide information of the extent of serial dependence in the sampled values. Here we see that the serial correlations start out high but decline to almost zero by lag twenty.

Tailored Proposal Density

To see the difference in results, the M–H algorithm is next implemented with a tailored proposal density. In this scheme one utilizes both $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{V}$ that were defined above. We let the proposal density be $f_T(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{V}, 15)$, a multivariate-$t$ density with fifteen degrees of freedom. This proposal density is similar to the random-walk proposal except that the distribution is centered at the fixed point $\hat{\boldsymbol{\beta}}$. The prior-posterior summary based on 5,000 draws of the M–H algorithm with this proposal density is given in Table 4.3. We see that the marginal posterior moments are similar to those in Table 4.1. The marginal posterior densities are

Table 4.3 Caesarean data: Prior-posterior summary based on 5,000 draws (beyond a burn-in of 100 cycles) from the tailored M–H algorithm

	Prior		Posterior			
	Mean	Std dev	Mean	Std dev	Lower	Upper
β_0	0.000	3.162	−1.080	0.220	−1.526	−0.670
β_1	0.000	3.162	0.593	0.249	0.116	1.095
β_2	0.000	3.162	1.181	0.254	0.680	1.694
β_3	0.000	3.162	−1.889	0.266	−2.421	−1.385

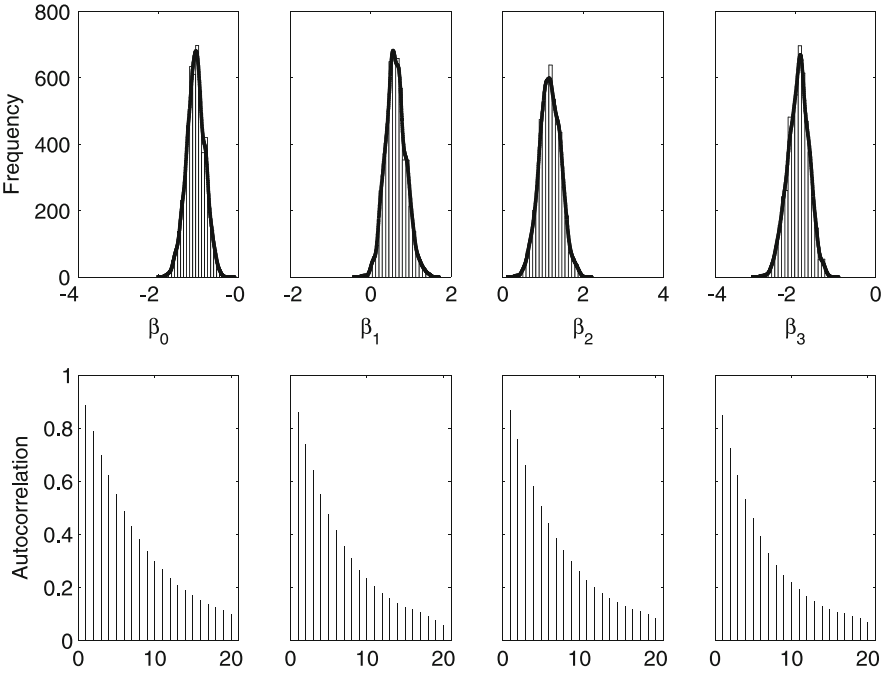

Fig. 4.2 Caesarean data with random-walk M–H algorithm: Marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

reported in the top panel of Fig. 4.2. These are virtually identical to those computed using the random-walk M–H algorithm. The most notable difference is in the serial correlation plots which decline much more quickly to zero indicating that the algorithm is mixing well. The same information is revealed by the inefficiency factors which are much closer to one than those from the previous algorithm.

The message from this analysis is that the two proposal densities produce similar results, with the differences appearing only in the autocorrelation plots (and inefficiency factors) of the sampled draws (Fig. 4.3).

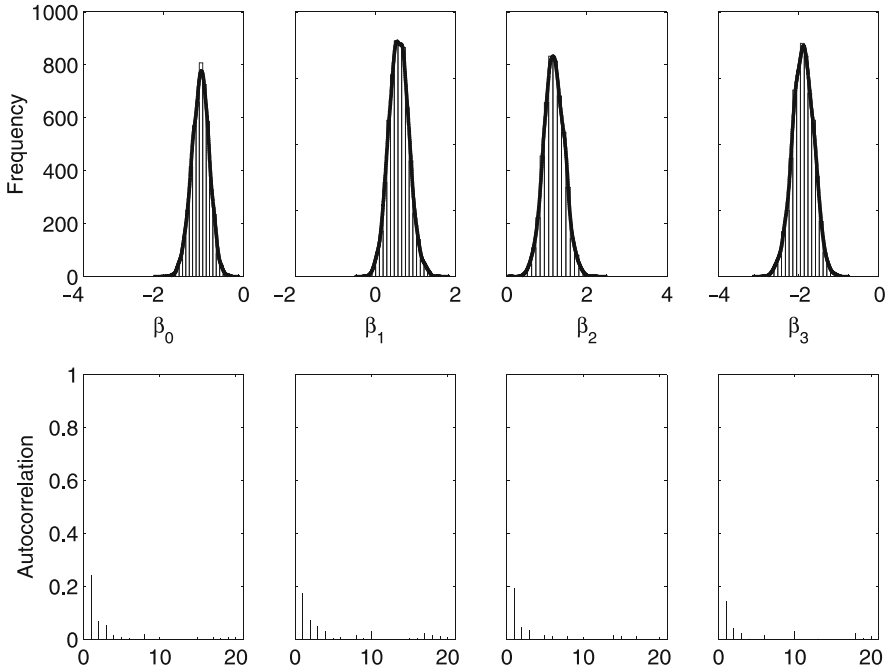

Fig. 4.3 Caesarean data with tailored M–H algorithm: Marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

4.3.3 Multiple-Block M–H Algorithm

In applications when the dimension of $\boldsymbol{\psi}$ is large, it can be difficult to construct a single block M–H algorithm that converges rapidly to the target density. In such cases, it is helpful to break up the variate space into smaller blocks and to then construct a Markov chain with these smaller blocks. Suppose, for illustration, that $\boldsymbol{\psi}$ is split into two vector blocks $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$. For example, in a regression model, one block may consist of the regression coefficients and the other block may consist of the error variance. Next, for each block, let

$$q_1(\boldsymbol{\psi}_1, \boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\ ; \quad q_2(\boldsymbol{\psi}_2, \boldsymbol{\psi}_2'|\boldsymbol{\psi}_1)\ ,$$

denote the corresponding proposal density. Here each proposal density q_k is allowed to depend on the data and the current value of the remaining block. Also define (by analogy with the single-block case)

$$\alpha(\boldsymbol{\psi}_1, \boldsymbol{\psi}_1'|\boldsymbol{\psi}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)q_1(\boldsymbol{\psi}_1', \boldsymbol{\psi}_1|\boldsymbol{\psi}_2)}{\pi(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)q_1(\boldsymbol{\psi}_1, \boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)}\right\}\ , \tag{4.19}$$

and

$$\alpha(\boldsymbol{\psi}_2, \boldsymbol{\psi}_2' | \boldsymbol{y}, \boldsymbol{\psi}_1) = \min\left\{1, \frac{\pi(\boldsymbol{\psi}_2' | \boldsymbol{\psi}_1) q_2(\boldsymbol{\psi}_2', \boldsymbol{\psi}_2 | \boldsymbol{\psi}_1)}{\pi(\boldsymbol{\psi}_2 | \boldsymbol{\psi}_1) q_2(\boldsymbol{\psi}_2, \boldsymbol{\psi}_2' | \boldsymbol{\psi}_1)}\right\} , \tag{4.20}$$

as the probability of move for block $\boldsymbol{\psi}_k$ $(k = 1, 2)$ conditioned on the other block. The conditional densities

$$\pi(\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2) \quad \text{and} \quad \pi(\boldsymbol{\psi}_2 | \boldsymbol{\psi}_1)$$

that appear in these functions are called the *full conditional densities*. By Bayes theorem each is proportional to the joint density. For example,

$$\pi(\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2) \propto \pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) ,$$

and, therefore, the probabilities of move in (4.19) and (4.20) can be expressed equivalently in terms of the kernel of the joint posterior density $\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ because the normalizing constant of the full conditional density (the norming constant in the latter expression) cancels in forming the ratio.

With these inputs, one sweep of the multiple-block M–H algorithm is completed by updating each block, say sequentially in fixed order, using a M–H step with the above probabilities of move, given the most current value of the other block.

Algorithm 2 Multiple-Block Metropolis–Hastings

1. Specify an initial value $\boldsymbol{\psi}^{(0)} = \left(\boldsymbol{\psi}_1^{(0)}, \boldsymbol{\psi}_2^{(0)}\right)$:
2. Repeat for $j = 1, 2, \ldots, n_0 + M$.
   (a) Repeat for $k = 1, 2$
      I. Propose a value for the kth block, conditioned on the previous value of kth block, and the current value of the other block $\boldsymbol{\psi}_{-k}$:

$$\boldsymbol{\psi}_k' \sim q_k\left(\boldsymbol{\psi}_k^{(j-1)}, \cdot | \boldsymbol{\psi}_{-k}\right) .$$

      II. Calculate the probability of move

$$\alpha_k\left(\boldsymbol{\psi}_k^{(j-1)}, \boldsymbol{\psi}_k' | \boldsymbol{y}, \boldsymbol{\psi}_{-k}\right) = \min\left\{1, \frac{\pi\left(\boldsymbol{\psi}_k' | \boldsymbol{\psi}_{-k}\right) q_k\left(\boldsymbol{\psi}_k', \boldsymbol{\psi}_k^{(j-1)} | \boldsymbol{\psi}_{-k}\right)}{h\left(\boldsymbol{\psi}_k^{(j-1)} | \boldsymbol{\psi}_{-k}\right) q_k\left(\boldsymbol{\psi}_k^{(j-1)}, \boldsymbol{\psi}_k' | \boldsymbol{\psi}_{-k}\right)}\right\} .$$

      III. Update the kth block as

$$\boldsymbol{\psi}_k^{(j)} = \begin{cases} \boldsymbol{\psi}_k' & \text{with prob} \quad \alpha_k\left(\boldsymbol{\psi}_k^{(j-1)}, \boldsymbol{\psi}_k' | \boldsymbol{\psi}_{-k}\right) \\ \boldsymbol{\psi}_k^{(j-1)} & \text{with prob} \quad 1 - \alpha_k\left(\boldsymbol{\psi}_k^{(j-1)}, \boldsymbol{\psi}_k' | \boldsymbol{\psi}_{-k}\right) \end{cases} .$$

3. Return the values $\left\{\boldsymbol{\psi}^{(n_0+1)}, \boldsymbol{\psi}^{(n_0+2)}, \ldots, \boldsymbol{\psi}^{(n_0+M)}\right\}$.

The extension of this method to more than two blocks is straightforward.

The transition kernel of the resulting Markov chain is given by the product of transition kernels

$$P(\boldsymbol{\psi}, d\boldsymbol{\psi}') = \prod_{k=1}^{2} P_k\left(\boldsymbol{\psi}_k, d\boldsymbol{\psi}'_k | \boldsymbol{\psi}_{-k}\right) \tag{4.21}$$

This transition kernel is not reversible, as can be easily checked, because under fixed sequential updating of the blocks updating in the reverse order never occurs. The multiple-block M–H algorithm, however, satisfies the weaker condition of invariance. To show this, we utilize the fact that each sub-move satisfies local reversibility (Chib and Jeliazkov 2001) and therefore the transition kernel $P_1(\boldsymbol{\psi}_1, d\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)$ has $\pi^*_{1|2}(\cdot|\boldsymbol{\psi}_2)$ as its local invariant distribution with density $\pi^*_{1|2}(\cdot|\boldsymbol{\psi}_2)$, i.e.,

$$\pi^*_{1|2}(d\boldsymbol{\psi}_1|\boldsymbol{\psi}_2) = \int P_1(\boldsymbol{\psi}_1, d\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)\pi_{1|2}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_1 \ . \tag{4.22}$$

Similarly, the conditional transition kernel $P_2(\boldsymbol{\psi}_2, d\boldsymbol{\psi}_2|\boldsymbol{\psi}_1)$ has $\pi^*_{2|1}(\cdot|\boldsymbol{\psi}_1)$ as its invariant distribution, for a given value of $\boldsymbol{\psi}_1$. Then, the kernel formed by multiplying the conditional kernels is invariant for $\pi^*(\cdot,\cdot)$:

$$\begin{aligned}
&\iint P_1(\boldsymbol{\psi}_1, d\boldsymbol{\psi}'_1|\boldsymbol{\psi}_2) P_2(\boldsymbol{\psi}_2, d\boldsymbol{\psi}'_2|\boldsymbol{\psi}'_1)\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_1\, d\boldsymbol{\psi}_2 \\
&= \int P_2(\boldsymbol{\psi}_2, d\boldsymbol{\psi}'_2|\boldsymbol{\psi}'_1)\left[\int P_1(\boldsymbol{\psi}_1, d\boldsymbol{\psi}'_1|\boldsymbol{\psi}_2)\pi_{1|2}(\boldsymbol{\psi}_1|\boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_1\right]\pi_2(\boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_2 \\
&= \int P_2(\boldsymbol{\psi}_2, d\boldsymbol{\psi}'_2|\boldsymbol{\psi}'_1)\pi^*_{1|2}(d\boldsymbol{\psi}'_1|\boldsymbol{\psi}_2)\pi_2(\boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_2 \\
&= \int P_2(\boldsymbol{\psi}_2, d\boldsymbol{\psi}'_2|\boldsymbol{\psi}'_1)\frac{\pi_{2|1}(\boldsymbol{\psi}_2|\boldsymbol{\psi}'_1)\pi^*_1(d\boldsymbol{\psi}'_1)}{\pi_2(\boldsymbol{\psi}_2)}\pi_2(\boldsymbol{\psi}_2)\, d\boldsymbol{\psi}_2 \\
&= \pi^*_1(d\boldsymbol{\psi}'_1)\int P_2(\boldsymbol{\psi}_2, d\boldsymbol{\psi}'_2|\boldsymbol{\psi}'_1)\pi_{2|1}(\boldsymbol{\psi}_2|\boldsymbol{\psi}'_1)\, d\boldsymbol{\psi}_2 \\
&= \pi^*_1(d\boldsymbol{\psi}'_1)\pi^*_{2|1}(d\boldsymbol{\psi}'_2|\boldsymbol{\psi}'_1) \\
&= \pi^*(d\boldsymbol{\psi}'_1, d\boldsymbol{\psi}'_2)\ ,
\end{aligned}$$

where the third line follows from (4.22), the fourth from Bayes theorem, the sixth from assumed invariance of P_2, and the last from the law of total probability.

The implication of this result is that it allows us to take draws in succession from each of the kernels, instead of having to run each to convergence for every value of the conditioning variable.

Remark 1. Versions of either random-walk or tailored proposal densities can be used in this algorithm, analogous to the single-block case. For example, Chib and Greenberg (1995) determine the proposal densities q_k by tailoring to $\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})$ in which case the proposal density is not fixed but varies across iterations. An

important special case occurs if each proposal density is taken to be the full conditional density of that block. Specifically, if we set

$$q_1\left(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1' | \boldsymbol{\psi}_2\right) = \pi(\boldsymbol{\psi}_1' | \boldsymbol{\psi}_2) \ ,$$

and

$$q_2\left(\boldsymbol{\psi}_2^{(j-1)}, \boldsymbol{\psi}_2' | \boldsymbol{\psi}_1\right) = \pi(\boldsymbol{\psi}_2' | \boldsymbol{\psi}_1) \ ,$$

then an interesting simplification occurs. The probability of move (for the first block) becomes

$$\begin{aligned} \alpha_1\left(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1' | \boldsymbol{\psi}_2\right) &= \min\left\{1, \frac{\pi\left(\boldsymbol{\psi}_1' | \boldsymbol{\psi}_2\right) \pi\left(\boldsymbol{\psi}_1^{(j-1)} | \boldsymbol{\psi}_2\right)}{\pi\left(\boldsymbol{\psi}_1^{(j-1)} | \boldsymbol{\psi}_2\right) \pi\left(\boldsymbol{\psi}_1' | \boldsymbol{\psi}_2\right)}\right\} \\ &= 1 \ , \end{aligned}$$

and similarly for the second block, implying that if proposal values are drawn from their full conditional densities then the proposal values are accepted with probability one. This special case of the multiple-block M–H algorithm (in which *each* block is proposed using its full conditional distribution) is called the Gibbs sampling algorithm.

4.4 The Gibbs Sampling Algorithm

The Gibbs sampling algorithm is one of the simplest Markov chain Monte Carlo algorithms. It was introduced by Geman and Geman (1984) in the context of image processing and then discussed in the context of missing data problems by Tanner and Wong (1987). The paper by Gelfand and Smith (1990) helped to demonstrate the value of the Gibbs algorithm for a range of problems in Bayesian analysis.

4.4.1 The Algorithm

To define the Gibbs sampling algorithm, let the set of full conditional distributions be

$$\left\{\pi(\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_p); \ \pi(\boldsymbol{\psi}_2 | \boldsymbol{\psi}_1, \boldsymbol{\psi}_3, \ldots, \boldsymbol{\psi}_p); \ \ldots, \pi(\boldsymbol{\psi}_p | \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{d-1})\right\} \ .$$

Now one cycle of the Gibbs sampling algorithm is completed by simulating $\{\boldsymbol{\psi}_k\}_{k=1}^p$ from these distributions, recursively refreshing the conditioning variables. When $d = 2$ one obtains the two block Gibbs sampler that appears in Tanner and

Wong (1987). The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm 3 Gibbs Sampling

1. Specify an initial value $\boldsymbol{\psi}^{(0)} = \left(\boldsymbol{\psi}_1^{(0)}, \ldots, \boldsymbol{\psi}_p^{(0)}\right)$:
2. Repeat for $j = 1, 2, \ldots, M$.

   Generate $\boldsymbol{\psi}_1^{(j+1)}$ from $\pi\left(\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2^{(j)}, \boldsymbol{\psi}_3^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)}\right)$

   Generate $\boldsymbol{\psi}_2^{(j+1)}$ from $\pi\left(\boldsymbol{\psi}_2 | \boldsymbol{\psi}_1^{(j+1)}, \boldsymbol{\psi}_3^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)}\right)$

   $\vdots$

   Generate $\boldsymbol{\psi}_p^{(j+1)}$ from $\pi\left(\boldsymbol{\psi}_p | \boldsymbol{\psi}_1^{(j+1)}, \ldots, \boldsymbol{\psi}_{p-1}^{(j+1)}\right)$.

3. Return the values $\left\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\right\}$.

It follows that the transition density of moving from $\boldsymbol{\psi}_k^{(j)}$ to $\boldsymbol{\psi}_k^{(j+1)}$ is given by

$$\pi\left(\boldsymbol{\psi}_k | \boldsymbol{\psi}_1^{(j+1)}, \ldots, \boldsymbol{\psi}_{k-1}^{(j+1)}, \boldsymbol{\psi}_{k+1}^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)}\right)$$

since when the kth block is reached, the previous $(k-1)$ blocks have been updated. Thus, the transition density of the chain, under the maintained assumption that π is absolutely continuous, is given by the product of transition kernels for each block:

$$K(\boldsymbol{\psi}, \boldsymbol{\psi}') = \prod_{k=1}^{p} \pi\left(\boldsymbol{\psi}_k | \boldsymbol{\psi}_1^{(j+1)}, \ldots, \boldsymbol{\psi}_{k-1}^{(j+1)}, \boldsymbol{\psi}_{k+1}^{(j)}, \ldots, \boldsymbol{\psi}_p^{(j)}\right) . \tag{4.23}$$

To illustrate the manner in which the blocks are revised, we consider a two block case, each with a single component, and trace out in Fig. 4.4 a possible trajectory of the sampling algorithm. The contours in the plot represent the joint distribution of $\boldsymbol{\psi}$ and the labels "(0)", "(1)" etc., denote the simulated values. Note that one iteration of the algorithm is completed after both components are revised. Also notice that each component is revised along the direction of the coordinate axes. This feature can be a source of problems if the two components are highly correlated because then the contours get compressed and movements along the coordinate axes tend to produce small moves. We return to this issue below.

4.4.2 Invariance of the Gibbs Markov Chain

The Gibbs transition kernel is invariant by construction. This is a consequence of the fact that the Gibbs algorithm is a special case of the multiple-block M–H algorithm

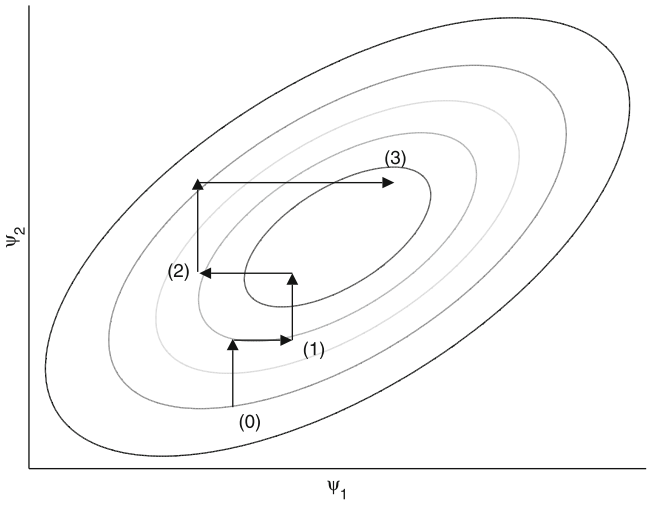

Fig. 4.4 Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations

which is invariant, as was established in the last section. Invariance can also be confirmed directly. Consider for simplicity a two block sampler with transition kernel density

$$K(\boldsymbol{\psi}, \boldsymbol{\psi}') = \pi(\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1') \, .$$

To check invariance we have to show that

$$\begin{aligned} &\int K(\boldsymbol{\psi}, \boldsymbol{\psi}')\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) d\boldsymbol{\psi}_1 d\boldsymbol{\psi}_2 \\ &\quad = \int \pi(\boldsymbol{\psi}_1'|\boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')\pi(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2) d\boldsymbol{\psi}_1 d\boldsymbol{\psi}_2 \end{aligned}$$

is equal to $\pi(\boldsymbol{\psi}_1', \boldsymbol{\psi}_2')$. This holds because $\pi(\boldsymbol{\psi}_2'|\boldsymbol{\psi}_1')$ comes out of the integral, and the integral over $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ produces $\pi(\boldsymbol{\psi}_1')$. This calculation can be extended to any number of blocks. It may be noted that the Gibbs Markov chain is not reversible. Reversible Gibbs samplers are discussed by Liu et al. (1995).

4.4.3 Sufficient Conditions for Convergence

Under rather general conditions, the Markov chain generated by the Gibbs sampling algorithm converges to the target density as the number of iterations become large. Formally, if we let $K(\boldsymbol{\psi}, \boldsymbol{\psi}')$ represent the transition density of the Gibbs algorithm and let $K^{(M)}(\boldsymbol{\psi}_0, \boldsymbol{\psi}')$ be the density of the draw $\boldsymbol{\psi}'$ after M iterations given the

starting value $\boldsymbol{\psi}_0$, then

$$\left\| K^{(M)}\left(\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}'\right) - \pi(\boldsymbol{\psi}')\right\| \to 0\,, \quad \text{as} \quad M \to \infty\,. \tag{4.24}$$

Roberts and Smith (1994) (see also Chan 1993) have shown that the conditions of Proposition 2 are satisfied under the following conditions: (1) $\pi(\boldsymbol{\psi}) > 0$ implies there exists an open neighborhood $N_{\boldsymbol{\psi}}$ containing $\boldsymbol{\psi}$ and $\epsilon > 0$ such that, for all $\boldsymbol{\psi}' \in N_{\boldsymbol{\psi}}$, $\pi(\boldsymbol{\psi}') \geq \epsilon > 0$; (2) $\int f(\boldsymbol{\psi})\, d\boldsymbol{\psi}_k$ is locally bounded for all k, where $\boldsymbol{\psi}_k$ is the kth block of parameters; and (3) the support of $\boldsymbol{\psi}$ is arc connected.

These conditions are satisfied in a wide range of problems.

4.4.4 Example: Simulating a Truncated Multivariate Normal

Consider the question of sampling a trivariate normal distribution truncated to the positive orthant. In particular, suppose that the target distribution is

$$\begin{aligned}\pi(\boldsymbol{\psi}) &= \frac{1}{\Pr(\boldsymbol{\psi} \in A)} f_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) I(\boldsymbol{\psi} \in A) \\ &\propto f_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) I(\boldsymbol{\psi} \in A)\end{aligned}$$

where $\boldsymbol{\mu} = (0.5, 1, 1.5)'$, $\boldsymbol{\Sigma}$ is in equi-correlated form with units on the diagonal and 0.7 on the off-diagonal, $A = (0, \infty) \times (0, \infty) \times (0, \infty)$ and $\Pr(\boldsymbol{\psi} \in A)$ is the normalizing constant which is difficult to compute. In this case, the Gibbs sampler is defined with the blocks ψ_1, ψ_2, ψ_3 and the full conditional distributions

$$\pi(\psi_1|\psi_2, \psi_3)\,; \quad \pi(\psi_2|\psi_1, \psi_3)\,; \quad \pi(\psi_3|\psi_1, \psi_2)\,,$$

where each of the these full conditional distributions is univariate truncated normal restricted to the interval $(0, \infty)$:

$$\begin{aligned}\pi(\psi_k|\boldsymbol{\psi}_{-k}) \propto f_N\left(\psi_k|\mu_k + \boldsymbol{C}_k'\boldsymbol{\Sigma}_{-k}^{-1}\left(\boldsymbol{\psi}_{-k} - \boldsymbol{\mu}_{-k}\right), \Sigma_k\right. \\ \left. -\boldsymbol{C}_k'\boldsymbol{\Sigma}_{-k}^{-1}\boldsymbol{C}_k\right) I(\psi_k \in (0, \infty))\,,\end{aligned} \tag{4.25}$$

$\boldsymbol{C}_k = Cov(\psi_k, \boldsymbol{\psi}_{-k})$, $\boldsymbol{\Sigma}_{-k} = Var(\boldsymbol{\psi}_{-k})$ and $\boldsymbol{\mu}_{-k} = E(\boldsymbol{\psi}_{-k})$. Figure 4.5 gives the marginal distribution of each component of ψ_k from a Gibbs sampling run of $M = 10000$ iterations with a burn-in of 100 cycles. The figures includes both the histograms of the sampled values and the Rao–Blackwellized estimates of the marginal densities (see Sect. 4.6 below) based on the averaging of (4.25) over the simulated values of $\boldsymbol{\psi}_{-k}$. The agreement between the two density estimates is close. In the bottom panel of Fig. 4.5 we plot the autocorrelation function of the sampled draws. The rapid decline in the autocorrelations for higher lags indicates that the sampler is mixing well.

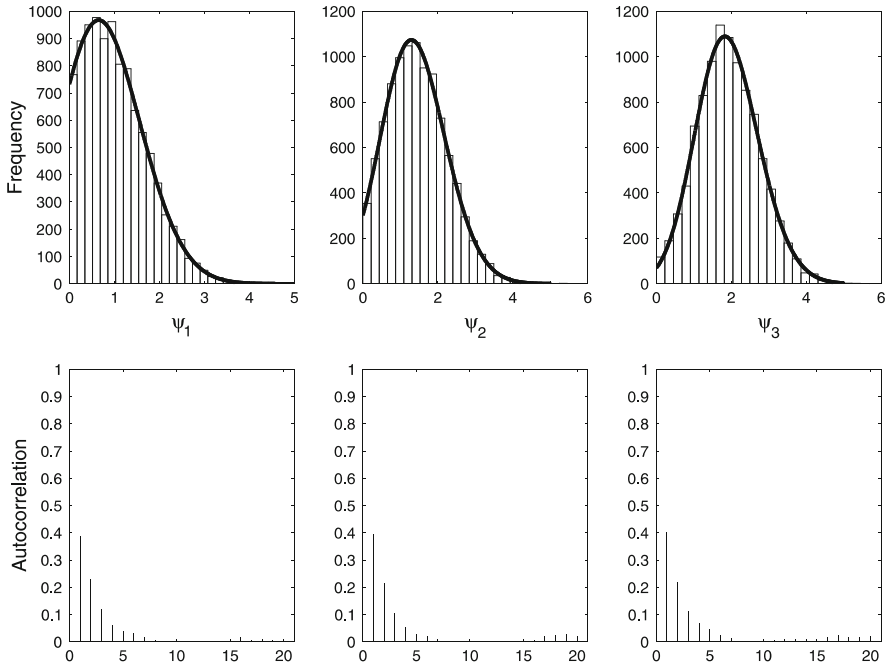

Fig. 4.5 Marginal distributions of ψ in truncated multivariate normal example (*top panel*). Histograms of the sampled values and Rao–Blackwellized estimates of the densities are shown. Autocorrelation plots of the Gibbs MCMC chain are in the *bottom panel*. Graphs are based on 10,000 iterations following a burn-in of 500 cycles

4.5 MCMC Sampling with Latent Variables

In designing MCMC simulations, it is sometimes helpful to modify the target distribution by introducing latent variables or auxiliary variables into the sampling. This idea was called data augmentation by Tanner and Wong (1987) in the context of missing data problems. Slice sampling, which we do not discuss in this chapter, is a particular way of introducing auxiliary variables into the sampling, for example see Damien et al. (1999).

To fix notations, suppose that z denotes a vector of latent variables and let the modified target distribution be $\pi(\boldsymbol{\psi}, z)$. If the latent variables are tactically introduced, the conditional distribution of $\boldsymbol{\psi}$ (or sub components of $\boldsymbol{\psi}$) given z may be easy to derive. Then, a multiple-block M–H simulation is conducted with the blocks $\boldsymbol{\psi}$ and z leading to the sample

$$\left(\boldsymbol{\psi}^{(n_0+1)}, z^{(n_0+1)}\right), \ldots, \left(\boldsymbol{\psi}^{(n_0+M)}, z^{(n_0+M)}\right) \sim \pi(\boldsymbol{\psi}, z) \ ,$$

where the draws on $\boldsymbol{\psi}$, ignoring those on the latent data, are from $\pi(\boldsymbol{\psi})$, as required.

To demonstrate this technique in action, we return to the probit regression example discussed in Sect. 4.3.2 to show how a MCMC sampler can be developed with the help of latent variables. The approach, introduced by Albert and Chib (1993), capitalizes on the simplifications afforded by introducing latent or auxiliary data into the sampling.

The model is rewritten as

$$\begin{aligned} z_i|\boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)\ , \\ y_i &= I[z_i > 0]\ , \quad i \le n\ , \\ \boldsymbol{\beta} &\sim \mathrm{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0)\ . \end{aligned} \tag{4.26}$$

This specification is equivalent to the probit regression model since

$$\Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \Pr(z_i > 0|\boldsymbol{x}_i, \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})\ .$$

Now the Albert–Chib algorithm proceeds with the sampling of the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \{z_i\}\ ; \quad \{z_i\}|\boldsymbol{y}, \boldsymbol{\beta}\ ,$$

where both these distributions are tractable (i.e., requiring no M–H steps). In particular, the distribution of $\boldsymbol{\beta}$ conditioned on the latent data becomes independent of the observed data and has the same form as in the Gaussian linear regression model with the response data given by $\{z_i\}$ and is multivariate normal with mean $\hat{\boldsymbol{\beta}} = \boldsymbol{B}(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^n \boldsymbol{x}_i z_i)$ and variance matrix $\boldsymbol{B} = (\boldsymbol{B}_0^{-1} + \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}$. Next, the distribution of the latent data conditioned on the data and the parameters factor into a set of n independent distributions with each depending on the data through y_i:

$$\{z_i\}|\boldsymbol{y}, \boldsymbol{\beta} \stackrel{d}{=} \prod_{i=1}^n z_i|y_i, \boldsymbol{\beta}\ ,$$

where the distribution $z_i|y_i, \boldsymbol{\beta}$ is the normal distribution $z_i|\boldsymbol{\beta}$ truncated by the knowledge of y_i; if $y_i = 0$, then $z_i \le 0$ and if $y_i = 1$, then $z_i > 0$. Thus, one samples z_i from $\mathcal{TN}_{(-\infty,0)}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ if $y_i = 0$ and from $\mathcal{TN}_{(0,\infty)}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ if $y_i = 1$, where $\mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes the $\mathcal{N}(\mu, \sigma^2)$ distribution truncated to the region (a, b).

The results, based on 5,000 MCMC draws beyond a burn-in of a 100 iterations, are reported in Fig. 4.4. The results are close to those presented above, especially to the ones from the tailored M–H chain (Fig. 4.6).

4.6 Estimation of Density Ordinates

We mention that if the full conditional densities are available, whether in the context of the multiple-block M–H algorithm or that of the Gibbs sampler, then the MCMC output can be used to estimate posterior marginal density functions (Gelfand and

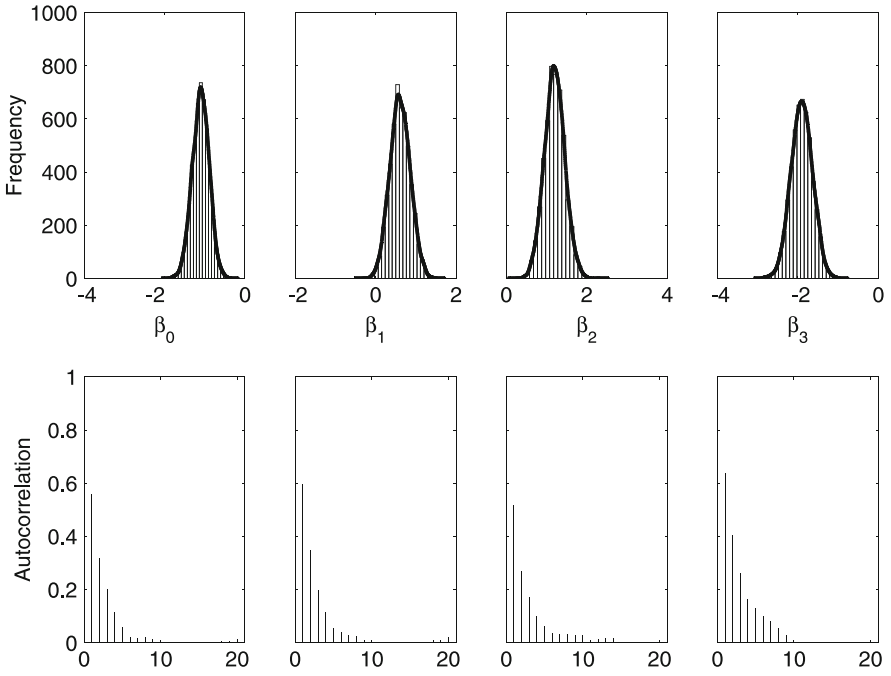

Fig. 4.6 Caesarean data with Albert–Chib algorithm: Marginal posterior densities (*top panel*) and autocorrelation plot (*bottom panel*)

Smith 1990; Tanner and Wong 1987). We exploit the fact that the marginal density of $\boldsymbol{\psi}_k$ at the point $\boldsymbol{\psi}_k^*$ is

$$\pi(\boldsymbol{\psi}_k^*) = \int \pi(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k})\, \pi(\boldsymbol{\psi}_{-k}) d\boldsymbol{\psi}_{-k} \ ,$$

where as before $\boldsymbol{\psi}_{-k} = \boldsymbol{\psi} \backslash \boldsymbol{\psi}_k$. Provided the normalizing constant of $\pi\left(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k}\right)$ is known, an estimate of the marginal density is available as an average of the full conditional density over the simulated values of $\boldsymbol{\psi}_{-k}$:

$$\hat{\pi}(\boldsymbol{\psi}_k^*) = M^{-1} \sum_{j=1}^{M} \pi\left(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k}^{(j)}\right) \ .$$

Under the assumptions of Proposition 1,

$$M^{-1} \sum_{j=1}^{M} \pi\left(\boldsymbol{\psi}_k^* | \boldsymbol{\psi}_{-k}^{(j)}\right) \to \pi(\boldsymbol{\psi}_k^*) \ , \quad \text{as} \quad M \to \infty \ .$$

Gelfand and Smith (1990) refer to this approach as Rao–Blackwellization because of the connections with the Rao–Blackwell theorem in classical statistics. That connection is more clearly seen in the context of estimating (say) the mean of ψ_k, $E(\psi_k) = \int \psi_k \pi(\psi_k) d\psi_k$. By the law of the iterated expectation,

$$E(\psi_k) = E\{E(\psi_k|\psi_{-k})\}$$

and therefore the estimates

$$M^{-1}\sum_{j=1}^{M}\psi_k^j$$

and

$$M^{-1}\sum_{j=1}^{M}E\left(\psi_k|\psi_{-k}^{(j)}\right)$$

both converge to $E(\psi_k)$ as $M \to \infty$. Under iid sampling, and under Markov sampling provided some conditions are satisfied – see Liu et al. (1994), Casella and Robert (1996) and Robert and Casella (1999), it can be shown that the variance of the latter estimate is smaller than that of the former. Thus, it can help to average the conditional mean $E(\psi_k|\psi_{-k})$, if that were available, rather than average the draws directly. Gelfand and Smith (1990) appeal to this analogy to argue that the Rao–Blackwellized estimate of the density is preferable to that based on the method of kernel smoothing. Chib (1995) extends the Rao–Blackwellization approach to estimate reduced conditional ordinates defined as the density of ψ_k conditioned on one or more of the remaining blocks. Finally, Chen (1994) provides an importance weighted estimate of the marginal density for cases where the conditional posterior density does not have a known normalizing constant. Chen's estimator is based on the identity

$$\pi(\psi_k^*) = \int w(\psi_k|\psi_{-k})\frac{\pi(\psi_k^*,\psi_{-k})}{\pi(\psi_k,\psi_{-k})}\pi(\psi)d\psi \ ,$$

where $w(\psi_k|\psi_{-k})$ is a completely known conditional density whose support is equal to the support of the full conditional density $\pi(\psi_k|\psi_{-k})$. In this form, the normalizing constant of the full conditional density is not required and given a sample of draws $\{\psi^{(1)},\ldots,\psi^{(M)}\}$ from $\pi(\psi)$, a Monte Carlo estimate of the marginal density is given by

$$\hat{\pi}(\psi_k^*) = M^{-1}\sum_{j=1}^{M} w\left(\psi_k^{(j)}\middle|\psi_{-k}^{(j)}\right)\frac{\pi\left(\psi_k^*,\psi_{-k}^{(j)}\right)}{\pi\left(\psi_k^{(j)},\psi_{-k}^{(j)}\right)} \ .$$

Chen (1994) discusses the choice of the conditional density w. Since it depends on ψ_{-k}, the choice of w will vary from one sampled draw to the next.

4.7 Sampler Performance and Diagnostics

In implementing a MCMC method it is important to assess the performance of the sampling algorithm to determine the rate of mixing and the size of the burn-in, both having implications for the number of iterations required to get reliable answers. A large literature has emerged on these issues, for example, Robert (1995), Tanner (1996, Sect. 6.3), Cowles and Carlin (1996), Gammermann (1997, Sect. 5.4) and Robert and Casella (1999), but the ideas, although related in many ways, have not coalesced into a single prescription.

One approach for determining sampler performance and the size of the burn-in time is to employ analytical methods to the specified Markov chain, prior to sampling. This approach is exemplified in the work of, for example, Polson (1996), Roberts and Tweedie (1996) and Rosenthal (1995). Two factors have inhibited the growth and application of these methods. The first is that the calculations are difficult and problem-specific and, second, the upper bounds for the burn-in that emerge from such calculations are usually conservative.

At this time the more popular approach is to utilize the sampled draws to assess both the performance of the algorithm and its approach to the invariant distribution. Several such relatively informal methods are available. Gelfand and Smith (1990) recommend monitoring the evolution of the quantiles as the sampling proceeds. Another useful diagnostic, one that is perhaps the most direct, are autocorrelation plots (and autocorrelation times) of the sampled output. Slowly decaying correlations indicate problems with the mixing of the chain. It is also useful in connection with M–H Markov chains to monitor the acceptance rate of the proposal values with low rates implying "stickiness" in the sampled values and thus a slower approach to the invariant distribution.

Somewhat more formal sample-based diagnostics are summarized in the CODA routines provided by Best et al. (1995). Although these diagnostics often go under the name "convergence diagnostics" they are in principle approaches that detect lack of convergence. Detection of convergence based entirely on the sampled output, without analysis of the target distribution, is perhaps impossible. Cowles and Carlin (1996) discuss and evaluate thirteen such diagnostics (for example, those proposed by Geweke 1992; Raftery and Lewis 1992; Ritter and Tanner 1992; Gelman and Rubin 1992; Gelman and Rubin 1992; and Zellner and Min 1995, amongst others) without arriving at a consensus. Difficulties in evaluating these methods stem from the fact that some of these methods apply only to Gibbs Markov chains (for example, those of Ritter and Tanner 1992; and Zellner and Min 1995) while others are based on the output not just of a single chain but on that of multiple chains specifically run from "disparate starting values" as in the method of Gelman and Rubin (1992). Finally, some methods assess the behavior of univariate moment estimates (as in the approach of Geweke 1992; and Gelman and Rubin 1992) while others are concerned with the behavior of the entire transition kernel (as in Ritter and Tanner 1992; and Zellner and Min 1995).

4.8 Strategies for Improving Mixing

In practice, while implementing MCMC methods it is important to construct samplers that mix well, where mixing is measured by the autocorrelation time, because such samplers can be expected to converge more quickly to the invariant distribution. Over the years a number of different recipes for designing samplers with low autocorrelation times have been proposed although it may sometimes be difficult, because of the complexity of the problem, to apply any of these recipes.

4.8.1 Choice of Blocking

As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution.

Blocks can be combined by the method of composition. For example, suppose that $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ denote three blocks and that the distribution $\boldsymbol{\psi}_1|\boldsymbol{\psi}_3$ is tractable (i.e., can be sampled directly). Then, the blocks $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ can be collapsed by first sampling $\boldsymbol{\psi}_1$ from $\boldsymbol{\psi}_1|\boldsymbol{\psi}_3$ followed by $\boldsymbol{\psi}_2$ from $\boldsymbol{\psi}_2|\boldsymbol{\psi}_1, \boldsymbol{\psi}_3$. This amounts to a two block MCMC algorithm. In addition, if it is possible to sample $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ marginalized over $\boldsymbol{\psi}_3$ then the number of blocks is reduced to one. Liu et al. (1994) discuss the value of these strategies in the context of a three-block Gibbs MCMC chains. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains used to sample multivariate normal target distributions.

4.8.2 Tuning the Proposal Density

As mentioned above, the proposal density in a M–H algorithm has an important bearing on the mixing of the MCMC chain. Fortunately, one has great flexibility in the choice of candidate generating density and it is possible to adapt the choice to the given problem. For example, Chib et al. (1998) develop and compare four different choices in longitudinal random effects models for count data. In this problem, each cluster (or individual) has its own random effects and each of these has to be sampled from an intractable target distribution. If one lets n denote the number of clusters, where n is typically large, say in excess of a thousand, then the number of blocks in the MCMC implementation is $n + 3$ (n for each of the random effect distributions, two for the fixed effects and one for the variance components matrix). For this problem, the multiple-block M–H algorithm requires $n + 1$ M–H steps within one iteration of the algorithm. Tailored proposal densities are therefore

computationally expensive but one can use a mixture of proposal densities where a less demanding proposal, for example a random walk proposal, is combined with the tailored proposal to sample each of the n random effect target distributions. Further discussion of mixture proposal densities is contained in Tierney (1994).

4.8.3 Other Strategies

Other approaches have also been discussed in the literature. Marinari and Parsi (1992) develop the simulated tempering method whereas Geyer and Thompson (1995) develop a related technique that they call the Metropolis-coupled MCMC method. Both these approaches rely on a series of transition kernels $\{K_1, \ldots, K_m\}$ where only K_1 has π^* as the stationary distribution. The other kernels have equilibrium distributions π_i, which Geyer and Thompson (1995) take to be $\pi_i(\boldsymbol{\psi}) = \pi(\boldsymbol{\psi})^{1/i}$, $i = 2, \ldots, m$. This specification produces a set of target distributions that have higher variance than π^*. Once the transition kernels and equilibrium distributions are specified then the Metropolis-coupled MCMC method requires that each of the m kernels be used in parallel. At each iteration, after the m draws have been obtained, one randomly selects two chains to see if the states should be swapped. The probability of swap is based on the M–H acceptance condition. At the conclusion of the sampling, inference is based on the sequence of draws that correspond to the distribution π^*. These methods promote rapid mixing because draws from the various "flatter" target densities have a chance of being swapped with the draws from the base kernel K_1. Thus, variates that are unlikely under the transition K_1 have a chance of being included in the chain, leading to more rapid exploration of the parameter space.

4.9 Concluding Remarks

In this survey we have provided an outline of Markov chain Monte Carlo methods. These methods provide a set of general recipes for sampling intractable multivariate distributions and have proved vital in the recent virtually revolutionary evolution and growth of Bayesian statistics. Refinements and extensions of these methods continue to occur. Two recent developments are the slice sampling method discussed by Mira and Tierney (2002), Damien et al. (1999) and Roberts and Rosenthal (1999) and the perfect sampling method proposed by Propp and Wilson (1996). The slice sampling method is based on the introduction of auxiliary uniform random variables to simplify the sampling and improve mixing while the perfect sampling method uses Markov chain coupling to generate an exact draw from the target distribution.

References

Albert, J., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)

Besag, J.: Spatial interaction and the statistical analysis of lattice systems (withdiscussion). J. Roy. Stat. Soc. B **36**, 192–236 (1974)

Besag, J., Green, E., Higdon, D., Mengersen, K.L.: Bayesian computation and stochastic systems (with discussion). Stat. Sci. **10**, 3–66 (1995)

Best, N.G., Cowles, M.K., Vines, S.K.: CODA: Convergence diagnostics and output analysis software for Gibbs sampling. Technical report, Cambridge MRC Biostatistics Unit (1995)

Carlin, B.P., Louis, T.: Bayes and Empirical Bayes Methods for Data Analysis, (2nd edn.), Chapman and Hall, London (2000)

Casella, G., Robert, C.P.: Rao–Blackwellization of sampling schemes. Biometrika **83**, 81–94 (1996)

Chan, K.S.: Asymptotic behavior of the Gibbs sampler. J. Am. Stat. Assoc. **88**, 320–326 (1993)

Chan, K.S., Ledolter, J.: Monte Carlo EM estimation for time series models involving counts. J. Am. Stat. Assoc. **90**, 242–252 (1995)

Chen, M-H.: Importance-weighted marginal Bayesian posterior density estimation. J. Am. Stat. Assoc. **89**, 818–824 (1994)

Chen, M-H., Shao, Qi-M., Ibrahim, J.G.: Monte Carlo Methods in Bayesian Computation. Springer, New York (2000)

Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**, 1313–1321 (1995)

Chib, S.: Markov Chain Monte Carlo Methods: Computation and Inference. In: Heckman, J.J., Leamer, E. (eds.) Handbook of Econometrics, Vol. 5, pp. 3569–3649. North Holland, Amsterdam (2001)

Chib, S., Greenberg, E.: Bayes inference for regression models with ARMA(p, q) errors. J. Econometrics **64**, 183–206 (1994)

Chib, S., Greenberg, E.: Understanding the Metropolis–Hastings algorithm. Am. Stat. **49**, 327–335 (1995)

Chib, S., Greenberg, E.: Markov chain Monte Carlo simulation methods in econometrics. Economet. Theor. **12**, 409–431 (1996)

Chib, S., Greenberg, E., Winklemann, R.: Posterior simulation and Bayes factors in panel count data models. J. Econometrics **86**, 33–54 (1998)

Chib, S., Jeliazkov, I.: Marginal likelihood from the Metropolis–Hastings output. J. Am. Stat. Assoc. **96**, 270–281 (2001)

Congdon, P.: Bayesian Statistical Modeling. Wiley, Chicester (20011)

Cowles, M.K., Carlin, B.: Markov chain Monte Carlo convergence diagnostics: A comparative review. J. Am. Stat. Assoc. **91**, 883–904 (1996)

Damien, P., Wakefield, J., Walker, S.: Gibbs Sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. J. Roy. Stat. Soc. B **61**, 331–344 (1999)

Gammerman, D.: Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall, London (1997)

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

Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **4**, 457–472 (1992)

Gelman, A., Meng, X.L., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, (2nd edn.), Chapman and Hall, London (2003)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 609–628 (1984)

Geweke, J.: Bayesian inference in econometric models using Monte Carlo integration. Econometrica **57**, 1317–1340 (1989)

Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics, pp. 169–193, Oxford University Press, New York (1992)
Geyer, C., Thompson, E.A.: Annealing markov chain monte carlo with applications to ancestral inference. J. Am. Stat. Assoc. **90**, 909–920 (1995)
Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice, Chapman and Hall, London (1996)
Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
Liu, J.S.: Monte Carlo Strategies in Scientific Computing, Springer, New York (2001)
Liu, J.S., Wong, W.H., Kong, A.: Covariance structure of the gibbs sampler with applications to the comparisons of estimators and data augmentation schemes. Biometrika **81**, 27–40 (1994)
Liu, J.S., Wong, W.H., Kong, A.: Covariance structure and convergence rate of the Gibbs sampler with various scans. J. Roy. Stat. Soc. B **57**, 157–169 (1995)
Marinari, E., Parsi, G.: Simulated tempering: A new Monte Carlo scheme. Europhys. Lett. **19**, 451–458 (1992)
Mengersen, K.L., Tweedie, R.L.: Rates of convergence of the Hastings and Metropolis algorithms. Ann. Stat. **24**, 101–121 (1996)
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
Meyn, S.P., Tweedie, R.L.: Markov Chains and Stochastic Stability. Springer, London (1993)
Mira, A., Tierney, L.: Efficiency and convergence properties of slice samplers. Scand. J. Stat. **29**, 1–12 (2002)
Nummelin, E.: General Irreducible Markov Chains and Non-negative Operators. Cambridge University Press, Cambridge (1984)
Polson, N.G.: Convergence of Markov Chain Monte Carlo algorithms. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Proceedings of the Fifth Valencia International Conference on Bayesian Statistics. pp. 297–323. Oxford University Press, Oxford (1996)
Propp, J.G., Wilson, D.B.: Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Struct. Algorithm. **9**, 223–252 (1996)
Raftery, A.E., Lewis, S.M.: How many iterations in the Gibbs sampler? In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Proceedings of the Fourth Valencia International Conference on Bayesian Statistics. pp. 763–774. Oxford University Press, New York (1992)
Ripley, B.: Stochastic Simulation. Wiley, New York (1987)
Ritter, C, Tanner, M.A.: Facilitating the gibbs sampler: The gibbs stopper and the Griddy-Gibbs sampler. J. Am. Stat. Assoc. **87**, 861–868 (1992)
Robert C.P.: Convergence control methods for Markov chain Monte Carlo algorithms. Stat. Sci. **10**, 231–253 (1995)
Robert, C.P.: Bayesian Choice. (2nd ed.), Springer, New York (2001)
Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (1999)
Roberts, G.O., Rosenthal, J.S.: Convergence of slice sampler Markov chains. J. Roy. Stat. Soc. B **61**, 643–660 (1999)
Roberts, G.O., Sahu, S.K.: Updating schemes, correlation structure, blocking, and parametization for the Gibbs sampler. J. Roy. Stat. Soc. B **59**, 291–317 (1997)
Roberts, G.O., Smith, A.F.M.: Some simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms. Stochas. Process. Appls. **49**, 207–216 (1994)
Roberts, G.O., Tweedie, R.L.: Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika **83**, 95–110 (1996)
Rosenthal, J.S.: Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Am. Stat. Assoc. **90**, 558–566 (1995)
Smith, A.F.M., Roberts, G.O.: Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. Roy. Stat. Soc. B **55**, 3–24 (1993)
Tanner, M.A.: Tools for Statistical Inference, (3rd edn.), Springer, New York (1996)

Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528–549 (1987)
Tierney, L.: Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762 (1994)
Zellner, A., Min, C.: Gibbs sampler convergence criteria. J. Am. Stat. Assoc. **90**, 921–927 (1995)