5 Introduction to Simulation and MCMC Methods
Siddhartha Chib

https://doi.org/10.1093/oxfordhb/9780199559084.013.0006  Pages 183–217
Published: 21 November 2012

Abstract

The purpose of this article is to provide an overview of Monte Carlo methods for generating variates from a target probability distribution that are based on Markov chains. These methods, called Markov chain Monte Carlo (MCMC) methods, are widely used to summarize complicated posterior distributions in Bayesian statistics and econometrics. This article begins with an intuitive explanation of the ideas and concepts that underlie popular algorithms such as the Metropolis–Hastings algorithm and multi-block algorithm. It provides the concept of a source or proposal density, which is used to supply a randomization step or an acceptance condition to determine if the candidate draw should be accepted. It is important to assess the performance of the sampling algorithm to determine the rate of mixing. Finally, this article offers an extensive discussion of marginal likelihood calculation using posterior simulator output.

Keywords: Monte Carlo methods, Markov chains, Metropolis-Hastings algorithm, multi-block algorithm, randomization step

Subject: Econometric and Statistical Methods and Methodology: General, Economics

Series: Oxford Handbooks

1 Introduction

The purpose of this chapter is to provide an overview of a class of Monte Carlo methods of generating variates from a target probability distribution which are based on Markov chains whose stationary distribution is the probability distribution of interest. These methods, popularly called Markov chain Monte Carlo (MCMC) methods, are widely used to summarize complicated posterior distributions in Bayesian statistics and econometrics.

Let \( p(y|\theta) \) denote the sampling density and \( \pi(\theta) \) the prior density, where \( y \) is a vector of observations and \( \theta = (\theta_1, \ldots, \theta_d) \in \Theta \subseteq \mathbb{R}^d \) is an unknown parameter. Then, the posterior density is given by
which in most practical problems is a complex distribution of unrecognizable form. The norming constant, or the marginal likelihood,

\[ m(y) = \int_\Theta p(y | \theta) \pi(\theta) \, d\theta \]

is also difficult to evaluate. As a result, analysis of the posterior distribution by direct analytical methods, or by numerical quadrature or even classical Monte Carlo methods, is generally not possible. These difficulties, however, can in principle be resolved by MCMC methods. These methods involve the simulation of a Markov sequence\( \{ \theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(g)}, \ldots \} \) on the state space \( \Theta \) such that the stationary distribution of the Markov chain is the target distribution of interest. A general method of constructing such chains is by the Metropolis–Hastings method that is due to Metropolis, et al. (1953) and Hastings (1970). The M–H method, as it is called, is quite general and flexible and forms the basis for almost all MCMC methods, including the Gibbs sampling method of Geman and Geman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990). It has been extensively studied and discussed, starting with Tierney (1994) and Chib and Greenberg (1995). Liu (2001), Chib (2001), Roberts and Rosenthal (2004), and Robert and Casella (2004) summarize the theory and provide extensions.

The defining feature of Markov chains is the fact that the future evolution of the chain, depends only on the current value of the chain. Thus, the distribution of \( \theta^{(g+1)} \), the \((g + 1)\)st element of the sequence, given the current and past values of the chain, depends only on \( \theta^{(g)} \), and is represented by the one-step transition kernel

\[ K(\theta^{(g)}; A | y) = \Pr(\theta^{(g+1)} \in A | y, \theta^{(g)}) \]

where \( A \) is any measurable set under \( \pi(\theta | y) \). The idea behind MCMC methods is to construct \( K(\theta, A | y) \) such that for any starting point \( \theta^{(0)} \), the probability that the chain will be in the set \( A \) after \( g \) steps, given by the \( g \)-step transition kernel \( K^{(g)}(\theta^{(0)}; A | y) \) and defined as

\[ K^{(g)}(\theta^{(0)}; A | y) = \Pr(\theta^{(g)} \in A | y, \theta^{(0)}) \]

converges to the probability of \( A \) under the target, as \( g \) becomes large. Once such a \( K(\theta, A | y) \) has been constructed, a sequence of variates can be obtained by recursively sampling the transition density in the following way:
\[ \theta^{(1)} \sim K(\theta^{(0)}, \cdot | y) \]
\[ \theta^{(2)} \sim K(\theta^{(1)}, \cdot | y) \]
\[ \vdots \]
\[ \theta^{(g+1)} \sim K(\theta^{(g)}, \cdot | y) \]

Because the chain was constructed to converge to the posterior density, the \( G \) values beyond the first \( n_0 \) iterations

\[
\left\{ \theta^{(n_0+1)}, \theta^{(n_0+2)}, \ldots, \theta^{(n_0+G)} \right\}
\]

(for suitably large \( n_0 \)) can be taken as draws from \( \pi(\theta | y) \). The initial draws up to \( n_0 \), constituting the "burn-in" period, are discarded to allow the effect of the starting value to wear off.

Provided \( G \) is large, the sample \( \left\{ \theta^{(n_0+1)}, \theta^{(n_0+2)}, \ldots, \theta^{(n_0+G)} \right\} \) can be used as a surrogate for the posterior density. One can summarize the target with the help of this sample. For instance, one can estimate the expectation of a real-valued function \( h(\theta) \) that is integrable under \( \pi(\theta | y) \) from the sample average

\[
\hat{h}_G = G^{-1} \sum_{g=1}^{G} h(\theta^{(g)}),
\]

(5.1)

as in the case of random samples (here and later the iterates used in the calculations are those beyond the burn-in period). By suitable laws of large numbers for Markov chains, one can show that

\[
\hat{h}_G \to \int \theta h(\theta) \pi(\theta | y) d\theta,
\]

almost surely as the simulation sample size \( G \) becomes large. Of course, the sample can be used to construct many other summaries of the target distribution. For example, the sample of draws

\[
\theta_{l}^{(n_0+1)}, \ldots, \theta_{l}^{(n_0+G)}
\]

on the \( l \)th component of \( \theta \) can be used to summarize the marginal distribution of \( \theta_l \)

\[
\pi(\theta_l | y) = \int \pi(\theta | y) d\theta_{-l},
\]

where
\[ \theta_{-l} = (\theta_1, \ldots, \theta_{l-1}, \theta_{l+1}, \ldots, \theta_d) \]

denotes the components of \( \theta \) excluding \( \theta_l \). This is because of the result that if a collection of variates is from a joint distribution, then the components of that sample are from the marginal distributions.

Despite the enormous power of the MCMC approach, the actual application of these methods requires considerable care and expertise. In practice, it is possible to construct several different transition densities \( K(\theta, \cdot \mid y) \) for the same target distribution. Some, or all, of these transition densities can produce draws that are heavily correlated, thus providing a poor exploration of the target even with large samples. These problems are particularly acute in high-dimensional problems with parameter constraints, ridges, fat areas, and other complications. Such complications are the norm in the dynamic stochastic general equilibrium (DSGE) models that are discussed in Marco and Schorfheide (Chapter 7, this volume). MCMC procedures have to be carefully designed in state space models (Giordani et al, Chapter 3 this volume), in financial time series models (Jacquier and Polson, Chapter 9, this volume), in hierarchical models (Rossi and Allenby, Chapter 8, this volume), nonparametric problems (Griffin et al, Chapter 4, this volume), in categorical response models (Li and Tobias, Chapter 6, this volume), and essentially in almost any problem where the model is relatively complex in relation to the available data.

1.1 Organization

The rest of the chapter is organized as follows. In Section 2 we summarize two classical Monte Carlo sampling methods, the accept-reject and importance sampling methods, that provide background for the newer MCMC methods. In Section 3 we discuss the M-H method and include the relevant Markov chain theory that justifies simulation by the M-H method. Section 4 deals with some special topics and Section 5 with the calculation of the marginal likelihood. Section 6 has several examples to illustrate the techniques. The last section has some summary comments.

2 Two classical sampling methods

We begin by briefly presenting two classical Monte Carlo methods. These methods share with MCMC methods the concept of a source or proposal density which is used to supply candidate draws and a randomization step or an acceptance condition to determine if the candidate draw should be accepted. As in the case of MCMC methods, as we will see later, the reliability or efficiency of the methods depends vitally on the match between the proposal and the target. Unlike MCMC methods, however, the two methods we present next produce independent samples from the target density (unless correlation is deliberately introduced as a variance reduction device).

2.1 Accept-reject method

The accept-reject method is the basis for many of the well-known univariate random number generators that are provided in software programs. This method is characterized by a source density \( q(\theta \mid y) \) which is used to supply candidate values and a constant \( c \) such that for all \( \theta \in \Theta \)

\[ q(\theta \mid y) \leq cy(\theta \mid y). \]
Thus, $c = \sup_{\theta \in \Theta} \{\pi(\theta | y)/q(\theta | y)\}$. Note that the accept-reject method does not require knowledge of the normalizing constant of $\pi$ because that constant can be absorbed in $c$. Then, in the accept-reject method, one draws a variate from $q$, accepting it with probability $\pi(\theta | y) / \{cq(\theta | y)\}$. If the particular proposal is rejected, a new one is drawn and the process continued until one is accepted. The accepted draws constitute an i.i.d (independent and identically distributed) sample from $\pi$. The efficiency of this method depends on $c$ which essentially is equal to the expected number of draws from $q$ before one is accepted.

**Algorithm: Accept-reject**

**Step 1:** In each iteration $g, g = 1, \ldots, G$,

- **Propose**
  
  $\theta^* \sim q(\theta | y)$; and independently $U \sim \text{Unif}(0, 1)$

- **Accept-reject** Let $\theta^{(g)} = \theta^*$ if
  
  $$U \leq \frac{\pi(\theta^{(g)} | y)}{cq(\theta^{(g)} | y)}$$

  otherwise go to Propose

**Step 2:** Return the values $\{\theta^{(g)}\}$.

The idea behind this algorithm may be explained quite simply using Figure 5.1. Imagine drawing random bivariate points in the region bounded above by the function $cq(\theta | y)$ and below by the $x$-axis. A point in this region may be drawn by first drawing $\theta^*$ from $q(\theta | y)$, which fixes the $x$-coordinate of the point, and then drawing the $y$-coordinate of the point as $U \sim \text{spec}(\theta^*)$. Now, if $U \sim \text{spec}(\theta^*) \leq \pi(\theta^* | y)$, the point lies below $\pi$ and is accepted; but the latter is simply the acceptance condition of the AR method, which completes the justification.

**Example:** The accept-reject method is sometimes applied to sampling a target density that is truncated to a non-standard region. In that case, one draws from the untruncated distribution. If the draw lies in the truncated region it is accepted; otherwise the process is repeated until a value that satisfies the truncation is found. As a simple example of this approach, consider sampling a standard normal distribution that is truncated to $(0, \infty)$. This is the half-normal distribution. If the proposal distribution is taken to be untruncated distribution, as in Figure 5.2, it is clear that the value $c = 2$ times the proposal bounds the target. The problem, however, is that although $2N(\theta, 1)$ perfectly bounds the target on $(0, \infty)$, it is quite badly matched to the target on the interval $(-\infty, 0)$. This is the generic problem with this approach to sampling a truncated distribution.
Figure 5.1

Accept-reject method

Notes: The x-coordinate of the points below the target density are accepted.

Figure 5.2

Half-normal target and the source density for the accept-reject method

2.2 Importance Sampling

Suppose that one is interested in calculating the value of the integral

\[ I = \int_{\Theta} h(\theta) \pi(\theta | y) d\theta \]

and suppose that there is a source density \( q(\theta | y) \) that is easy to sample from and which is a close match to \( \pi(\theta | y) \). Write

\[
I = \frac{\int_{\Theta} h(\theta) p(\theta | y) \pi(\theta) d\theta}{\int_{\Theta} p(\theta | y) \pi(\theta) d\theta}.
\]

Then in the method of importance sampling, we re-express \( I \) as a ratio of expectations with respect to \( q(\theta | y) \):
then take a large number of draws \( \{ \theta^{(1)}, \ldots, \theta^{(G)} \} \) from \( q(\theta | y) \) and estimate \( I \) as

\[
\hat{I} = \frac{1}{G} \sum_{g=1}^{G} h(\theta^{(g)}) w(\theta^{(g)}, y)
\]

where

\[
w(\theta^{(g)}, y) = \frac{p(y | \theta^{(g)}) \pi(\theta^{(g)})}{q(\theta^{(g)} | y)}
\]

are the so-called importance weights. For this method to work (in particular for \( \hat{I} \) to have finite Monte Carlo variance) these importance weights must be bounded as a function of \( \theta \). Importance sampling was first discussed in an econometric setting by Kloek and Van Dijk (1978). Further details about the importance sampling approach for evaluating integrals and assessing the approximation error are given in Geweke (1989 and 2005).

**Sampling importance re-sampling** Over the last decade or so, the scope of importance sampling methods has broadened to include the sampling of the target density itself. Now since

\[
\pi(\theta | y) = \frac{\pi(\theta | y)}{q(\theta | y)} q(\theta | y)
\]

it follows that if \( \{ \theta^{(1)}, \ldots, \theta^{(G)} \} \) are draws from \( q(\theta | y) \), then the target can be expressed as the discrete distribution

\[
\hat{\pi}(\theta | y) = w(\theta^{(g)}, y) \delta_{\theta^{(g)} | \theta}
\]

where \( w(\theta^{(g)}, y) \), which is defined above, is the probability mass attached to the point \( \theta^{(g)} \) and \( \delta_{\theta^{(g)} | \theta} \) is the indicator function which takes the value 1 if \( \theta = \theta^{(g)} \) and the value zero otherwise. The variates \( \theta^{(g)} \) are called particles.

Given this representation, it follows that to get the particles \( \{ \theta^{*^{(1)}}, \ldots, \theta^{*^{(L)}} \} \) from \( \pi(\theta | y) \) (where \( L \) is likely much smaller than \( G \)), we sample \( \pi(\theta | y) \). This amounts to re-sampling \( \{ \theta^{(1)}, \ldots, \theta^{(G)} \} \) with replacement with probabilities \( \{ w(\theta^{(g)}, y) \} \).

That this method works is easily checked. Under \( \pi(\theta | y) \), and for any measurable set \( A \),
as $L$ and $G$ both increase with $L/G$ going to zero and provided the expectations are bounded. Rubin (1988) has called this the sampling importance re-sampling or SIR method.

Figure 5.3

Notes: Incoming particles are reweighted (resampled) according to probabilities proportional to the ratio of the target to the proposal density.

We illustrate this method in Figure 5.3 which shows how the particle $\theta^{(g)}$ gets weights according to the importance of the ratio of the target to the proposal.

The SIR approach is now also heavily used in nonlinear state-space models where it is referred to as particle filtering, a method that was introduced into econometrics in Kim et al. (1998). Particle filtering is discussed in several recent papers, for example, Del Moral et al. (2006) and is extensively discussed in Giordani et al. (Chapter 3, this volume).

3 Metropolis-Hastings Algorithm

Suppose that we are interested in sampling the target density $\pi(\theta | y)$, where $\theta$ is the parameter vector and $\pi(\theta | y)$ is a continuous density. The idea behind the M-H algorithm is to simulate a convenient transition density $q(\theta, \theta' | y)$, where $(\theta, \theta')$ are any two points, and then to modify the transition density to ensure that the modified Markov chain has the correct limiting distribution. The source density $q(\theta, \theta' | y)$ is called the candidate generating density or proposal density.

To define the M-H algorithm, let $\theta^{(g)}$ be the current value. Then the next value $\theta^{(g+1)}$ is produced by a two-step process consisting of a "proposal step" and a "move step."

Algorithm: Metropolis-Hastings

Step 1: Initialize $\theta^{(0)}$
3.1 Derivation of the M-H Algorithm

Step 2: In each iteration \( g, g = 1, \ldots, n_0 + G \),

- **Propose:** Sample a proposal value \( \theta^* \) from \( q(\theta^{(g)}, \theta | y) \) and calculate the quantity (the acceptance probability or the probability of move)

\[
\alpha(\theta^{(g)}, \theta^* | y) = \min \left\{ 1, \frac{\pi(\theta^* | y) q(\theta^{(g)}, \theta^* | y)}{\pi(\theta^{(g)} | y) q(\theta^{(g)}, \theta^{*} | y)} \right\}
\]  
(5.2)

- **Move:** Set

\[
\theta^{(g+1)} = \begin{cases} 
\theta^* & \text{with prob } \alpha(\theta^{(g)}, \theta^* | y) \\
\theta^{(g)} & \text{with prob } 1 - \alpha(\theta^{(g)}, \theta^* | y)
\end{cases}
\]

Step 3: Discard the draws from the first \( n_0 \) iterations and save the subsequent \( G \) draws \( \theta^{(n_0+1)}, \ldots, \theta^{(n_0+G)} \)

Note that the probability \( \alpha(\theta^{(g)}, \theta^* | y) \) does not require knowledge of the norming constant of the posterior density \( \pi(\theta | y) \). Also if the proposal density satisfies the condition \( q(\theta, \theta^* | y) = q(\theta^*, \theta | y) \), then \( \alpha(\theta^{(g)}, \theta^* | y) \) reduces to

\[
\alpha(\theta^{(g)}, \theta^* | y) = \min \left\{ 1, \frac{\pi(\theta^* | y)}{\pi(\theta^{(g)} | y)} \right\}.
\]

This simplification is a feature of the random-walk version of the M-H algorithm that is discussed below.

### 3.1 Derivation of the M-H Algorithm

The M-H algorithm has been derived by Chib and Greenberg (1995) from the logic of reversible Markov chains. Their derivation is helpful in understanding the M-H algorithm.

To start, we need some definitions. A Markov transition density \( K(\theta, \theta^* | y) \) is reversible for \( \pi(\theta | y) \) if for every \( (\theta, \theta^*) \) in the support of the target distribution

\[
\pi(\theta | y) K(\theta, \theta^* | y) = \pi(\theta^* | y) K(\theta^*, \theta | y).
\]  
(5.3)

If a chain is reversible it is also invariant. Invariance refers to the property that

\[
\pi(\theta^* | y) = \int K(\theta, \theta^* | y) \pi(\theta | y) d\theta
\]  
(5.4)
which means that once convergence is achieved, a subsequent value $\theta'$ drawn from the transition density is also from the target density. To see that reversibility implies invariance one simply integrates both sides of (5.3) over $\theta$. This leads to the invariance condition since $\int K(\theta',\theta\mid y)\,d\theta = 1$ by virtue of being a transition density.

Now consider the Markov chain induced by the proposal density $q(\theta,\theta'\mid y)$. Suppose that for a pair of points $(\theta,\theta')$ it is true that

$$p(\theta\mid y)q(\theta,\theta'\mid y) = p(\theta'\mid y)q(\theta',\theta\mid y),$$  \hspace{1cm} (5.5)

which means informally that the process moves from $\theta$ to $\theta'$ too frequently and too rarely in the reverse direction. We can correct this situation by reducing the flow from $\theta$ to $\theta'$ by introducing probabilities $\alpha(\theta,\theta'\mid y)$ and $\alpha(\theta',\theta\mid y)$ of making the moves in either direction so that

$$p(\theta\mid y)q(\theta,\theta'\mid y)\alpha(\theta,\theta'\mid y) = p(\theta'\mid y)q(\theta',\theta\mid y)\alpha(\theta',\theta\mid y).$$  \hspace{1cm} (5.6)

We now set $\alpha(\theta,\theta'\mid y)$ to be as high as possible, namely equal to one. Solving for $\alpha(\theta,\theta'\mid y)$, we then get that

$$\alpha(\theta,\theta'\mid y) = \frac{p(\theta'\mid y)q(\theta',\theta\mid y)}{p(\theta\mid y)q(\theta,\theta'\mid y)}.$$

This quantity is less than one because we started from (5.5). On the other hand, if we reverse the inequality in (5.5), a similar argumentation leads to the conclusion that $\alpha(\theta,\theta'\mid y) = 1$, which produces the expression given in (5.2).

### 3.2 Transition density of the M-H chain

Because of the fact that the M-H chain can repeat values, the transition density of the M-H chain $\pi_{\text{MH}}(\theta,\theta'\mid y)$ has two components—one for the move away from $\theta$ given by

$$\alpha(\theta,\theta'\mid y)q(\theta,\theta'\mid y)$$

and one for the probability of staying at $\theta$ given by

$$r(\theta\mid y) = 1 - \int \alpha(\theta,\theta'\mid y)q(\theta,\theta'\mid y)\,d\theta'.$$
3.3 Convergence properties

Theoretical properties of the M-H algorithm (in particular, the ergodic behavior of the chain from an arbitrarily specified initial value) require assumptions about the properties of the Markov chain. The main results are from Tierney (1994), which also has the definition of the terms involved.

Theorem 1. Suppose that the Markov chain \( \{ \theta^{(g)} \} \) is \( \pi \)-irreducible and has invariant distribution \( \pi(\theta \mid y) \). Then \( \pi(\theta \mid y) \) is the unique invariant distribution. If the chain is \( \pi \)-irreducible, aperiodic, and the invariant distribution is proper, then for \( \pi \)-every \( \theta^{(0)} \) and all measurable sets \( A \)

\[
\| \Pr(\theta^{(g)} \in A \mid y, \theta^{(0)}) - \int_A \pi(\theta \mid y) \, d\theta \| \rightarrow 0
\]

as \( g \rightarrow \infty \), where \( \cdot \mid \cdot \) denotes the total variation distance. If the chain is ergodic (\( \pi \)-irreducible, aperiodic, and Harris recurrent), then for all functions \( h(\theta) \) such that \( \int_{\Theta} |h(\theta)| \pi(\theta \mid y) \, d\theta \) \( \langle \infty \) and any initial distribution,

\[
\hat{h}_G = G^{-1} \sum_{g=1}^G h(\theta^{(g)}) \rightarrow \int_{\Theta} h(\theta) \pi(\theta \mid y) \, d\theta \quad \text{as} \quad G \rightarrow \infty, \ a.s.
\]

These powerful results hold under relatively weak conditions (for example, as discussed in Tierney (1994), \( \pi \)-irreducibility of the chain is satisfied if the proposal density is everywhere positive in the support of the posterior density; it is Harris recurrent if it is \( \pi \)-irreducible, has \( \pi \) as it is unique invariant distribution, and the transition kernel is absolutely continuous with respect to \( \pi \)). These provide the basis for MCMC methods. For a given target distribution and MCMC transition density, they allow us to utilize the simulated sample path of that transition kernel to develop simulation-consistent estimates of posterior moments, posterior probabilities, and other summaries of the target.

A central limit theorem for sample-path averages requires a further strengthening of the conditions. One now requires that the chain is also uniformly ergodic. An ergodic chain with invariant distribution \( \pi \) is
uniformly ergodic if there exists a non-negative bounded real-valued function $C(\cdot)$ and a positive constant $r$ (1 such that

$$\|k^{(\theta)}(\theta^{(0)}, A \mid y) - \int_{\Theta} p(\theta \mid y) d\theta \| \leq C(\theta^{(0)})$$

for every $\theta^{(0)}$, $g$, and measurable sets $A$. Then we have the following ergodic limit theorem for the sample average $h_G$.

**Theorem 2.** Suppose that the Markov chain $\{\theta^{(i)}\}$ is uniformly ergodic and has invariant distribution $\pi(\theta \mid y)$. Then for functions $h(\theta)$ such that $\int_\Theta h(\theta)^2 \pi(\theta \mid y) d\theta < \infty$, and any initial distribution, the sample average $h_G$ satisfies the ergodic limit theorem

$$\sqrt{G} \left( h_G - E_h \right) \stackrel{d}{\rightarrow} N(0, \sigma_h^2)$$

where

$$E_h = \int_\Theta h(\theta) \pi(\theta \mid y) d\theta$$

$$\sigma_h^2 = \text{Var}_x \left\{ h(\theta^{(1)}), h(\theta^{(1)}) \right\} + 2 \sum_{k=2}^\infty \text{Cov}_x \left\{ h(\theta^{(1)}), h(\theta^{(k)}) \right\}$$

and the subscript $\pi$ indicates that the expectations are calculated under the invariant distribution.

### 3.4 Numerical standard error and inefficiency factor

In order to understand how accurately $h_G$ has estimated $E_h$, it is necessary to calculate $\text{Var}(h_G) = \sigma_h^2 / G$, where $\sigma_h^2$ is the variance that appears in Theorem 2. The square root of $\text{Var}(h_G)$ is called the numerical standard error. One effective way of estimating $\text{Var}(h_G)$ is by the method of batch means (Ripley 1987). First, we let $Z_g = h(\theta^{(i)})$, $g = 1, 2, \ldots, G$. Next, we divide the data $\{Z_1, Z_2, \ldots, Z_G\}$ into $k$ non-overlapping batches of length $m$ with means

$$B_i = m^{-1}(Z_{(i-1)m+1} + \ldots + Z_{im}), i = 1, 2, \ldots, k$$

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

$$\bar{B} = \frac{1}{k} \sum_{i=1}^k B_i$$

is of course $h_g$ and the estimate of the sample variance of $B$ by standard calculations is
3.5 Choice of proposal density

\[
\text{Var}(\tilde{b}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{y}_i - \bar{y})^2.
\]

In the batch means method, this variance estimate is taken to be the estimate of \(\text{Var}(h_g)\). Jones et al. (2006) show that it is a consistent estimate of \(\sigma^2_h / G\) if \(k\) and \(m\) both increase with \(G\).

**Inefficiency factor** Once we have an estimate of \(\text{Var}(h_G)\), a useful quantity to calculate and report is the _inefficiency factor_ defined as

\[
\text{Ineff}(\hat{h}_G) = \frac{\text{Var}(\hat{h})}{\sigma^2 / G},
\]

(5.7)

where \(s^2\) is just the sample variance of \(\{Z_g\}\). This quantity is the ratio of variance of \(h_G\) to the variance of \(h_G\) relevant for the case of independent draws. One way to interpret this quantity is in terms of the _effective sample size_, or ESS, defined as

\[
\text{ESS}(\hat{h}_G) = \frac{G}{\text{Ineff}(\hat{h}_G)}.
\]

(5.8)

With independent sampling, the inefficiency factor is theoretically equal to one, and the effective sample size is \(G\). When the inefficiency factor is greater than one (the typical case with MCMC sampling), the effective sample size is smaller than \(G\).

### 3.5 Choice of proposal density

There are many ways of formulating the proposal density. We consider two that are popular in practice.

**Random walk proposals** Given the current value \(\theta\), the proposal is drawn as

\[
\theta^\dagger = \theta + z,
\]

where \(z\) follows some symmetric distribution \(q\) such as the multivariate normal with mean of zero and some covariance matrix \(\mathcal{N}\), which is adjusted in trial runs to reach some desired acceptance rate (given by the proportion of proposed values that are accepted). Because of the symmetry of the increment distribution it follows that \(q(\theta^\dagger | y) = q(\theta | \rho y)\) and hence, due to the cancellation of the \(q\) terms, the M-H probability of move takes the simplified form
\[ \alpha(\theta, \theta^\dagger | y) = \min \left\{ 1, \frac{\pi(\theta^\dagger | y)}{\pi(\theta | y)} \right\} \]

as a function solely of the target density, as illustrated in Figure 5.4.

Although the random walk M-H proposal is quite popular in applications, it can be difficult to tune, especially when the dimension of \( \theta \) is large. In such cases, it is difficult to generate reasonable acceptance rates and large enough moves to ensure a full exploration of the posterior surface.

**Independent proposals** Another common strategy is to set \( q(\theta, \theta^\dagger | y) = q(\theta \mid y) \), an independence M-H chain in the terminology of Tierney (1994). In this case,

\[ \alpha(\theta, \theta^\dagger | y) = \min \left\{ 1, \frac{\pi(\theta^\dagger | y)}{\pi(\theta \mid y)} \frac{q(\theta \mid y)}{q(\theta^\dagger \mid y)} \right\} , \]

which involves the ratio of targets and the ratio of proposal densities, as shown in Figure 5.5.

**Figure 5.4**

Random-walk M-H: the two points that determine the probability of move

One way to implement such chains is by tailoring 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 (Chib and Greenberg 1994). Specifically, one can let \( q(\theta \mid y) = p(\theta \mid m, V) \), where \( p \) is some multivariate density and the parameters of the proposal density are taken to be

\[ m = \max_{\theta} \log \pi(\theta \mid y) \quad \text{and} \quad V = \tau - \frac{\partial^2 \log \pi(\theta \mid y)}{\partial \theta \partial \theta'}^{-1} \theta_{\text{current}} - \hat{\theta} \]

\[ (5.9) \]
3.6 Multiple-block sampling

When the dimension of $\theta$ is large, sampling of the target in one block (as described in the preceding section) tends not to be efficient. It becomes necessary then to divide the parameters into smaller groups or blocks and then to sample these blocks in turn. To explain this idea, suppose that $\theta$ is split as $(\theta_1, \theta_2)$, with $\theta_k \in \Omega_k \subseteq \mathbb{R}^{d_k}$. In many problems, this sort of grouping is suggested by the model structure itself. For example, in a regression model, one block may consist of the regression coefficients and the other block may consist of the error variance.

Now let

$$q_1(\theta_1, \theta_1^* | y, \theta_2); q_2(\theta_2, \theta_2^* | y, \theta_1)$$

denote the proposal densities, one for each block $\theta_k$, where the proposal density $q_k$ may depend on the current value of the remaining block. Versions of either the random-walk or tailored proposal densities are possible, analogous to the single-block case. Also define

$$a_1(\theta_1, \theta_1^* | y, \theta_2) = \min \left\{ \frac{\pi(\theta_1^* | y, \theta_2)q_1(\theta_1^*, \theta_1 | y, \theta_2)}{\pi(\theta_1 | y, \theta_2)q_1(\theta_1, \theta_1^* | y, \theta_2)}, 1 \right\}$$

(5.10)

and
Algorithm: Multiple-block M-H

\( \alpha_2(\theta_2, \theta^*_2 \mid y, \theta_1) = \min \left\{ 1, \frac{\pi(\theta^*_2 \mid y, \theta_1)q_2(\theta_2, \theta^*_2 \mid y, \theta_1)}{\pi(\theta_2 \mid y, \theta_1)q_2(\theta^*_2, \theta_2 \mid y, \theta_1)} \right\} \)

(5.11)

as the probability of move for block \( \theta_k \) (\( k = 1, 2 \)) conditioned on the other block. The conditional posterior densities

\[ \pi(\theta_k \mid y, \theta_{\overline{k}}) \]

that appear in these functions are called the full conditional densities and by Bayes’ theorem are just proportional to the joint posterior density. For example,

\[ \pi(\theta_1 \mid y, \theta_2) \propto \pi(\theta_1, \theta_2 \mid y), \]

and, therefore, the probabilities of move in (5.10) and (5.11) can be expressed equivalently in terms of the kernel of the joint posterior density \( \pi(\theta_1, \theta_2 \mid y) \) because the normalizing constant of the full conditional density (the norming constant in the latter expression) cancels in forming the ratio. Then in the multiple-block M-H algorithm, one sweep of the algorithm is completed by updating each block, say sequentially in fixed order, using an M-H step with the above probabilities of move, given the most current value of the other block. Thus, at the \( g \)th iteration, given the value \( \theta_2^{(g)} \) one sweep is completed by the following steps.

**FIGURE 5.6**

Notes: Left panel has the target and proposal densities of the first block and the four points that determine the probability of move; right panel has the same information for the second block.

**Algorithm: Multiple-block M-H**

Step 1: Propose

\( \theta_1^* \sim q_1(\theta_1^{(g)} \mid y, \theta_2^{(g)}) \)
and move with probability

\[ a_1(\theta^{(g)}_1, \theta^*| y, \theta_2^{(g)}) \]

(otherwise stay at the current value) to produce the value \( \theta_1^{(g+1)} \).

**Step 2:** Given this updated value of the first block, propose

\[ \theta^*_2 \sim q_2(\theta^{(g+1)}_2, \theta^*_2| y, \theta_1^{(g)}) \]

and move with probability

\[ a_2(\theta^{(g)}_2, \theta^*_2| y, \theta_1^{(g)}) \]

(otherwise stay at the current value) to produce the value \( \theta_2^{(g+1)} \).

The ingredients involved in these two steps are illustrated in Figure 5.6.

It is important to note that the Markov chain generated by this multiple-block M-H satisfies the invariance condition but not reversibility. It does, however, satisfy the condition of local reversibility, a concept that is introduced in Chib and Jeliazkov (2001). The assumptions under which such an algorithm is Harris-recurrent are taken up in Roberts and Rosenthal (2006).

**Gibbs sampling** In many situations, it is the case that each full conditional distribution is of recognizable form and easily sampled. We can use these as the proposal distributions in the multiple-block M-H algorithm. Specifically, let

\[ q_1(\theta^{(g)}_1, \theta^*_1| y, \theta_2^{(g)}) = \pi(\theta^*_1| y, \theta_2^{(g)}) \]

and

\[ q_2(\theta^{(g+1)}_2, \theta^*_2| y, \theta_1^{(g+1)}) = \pi(\theta^*_2| y, \theta_1^{(g+1)}) \]

If we insert these choices into the probabilities of move, an interesting simplification occurs. For instance, for the first block, the probability of move reduces to
and similarly for the second block, implying that if the 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 MH algorithm (in which each block is proposed using its full conditional distribution) is called the Gibbs sampling algorithm.

**Algorithm**: Gibbs sampling

In each iteration $g,g' = 1,...,n_0 + G$,

- Generate $\theta_1^{(g+1)}$ from $\pi(\theta_1 | y, \theta_2^{(g)})$
- Generate $\theta_2^{(g+1)}$ from $\pi(\theta_2 | y, \theta_1^{(g+1)})$

Return the values $\{\theta(0+1), \theta(n_0+2), ..., \theta(n_0+G)\}$.

**Metropolis-within-Gibbs** In some problems, it is the case that one or more of the full conditional distributions are of recognizable form, but at least one full conditional distribution is not. In that case, the blocks with the tractable full conditional distributions can be sampled directly and the blocks with the intractable full conditional distributions $\psi_i$ by an M-H step. Such an algorithm is sometimes called the Metropolis-within-Gibbs algorithm though the multiple-block M-H terminology is sufficient to cover this special case.

**General case** The extension of the multiple block method to more than two blocks is straightforward in principle. 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. It is also possible in some cases to reduce the number of blocks by the method of composition. For example, suppose that $\theta_1, \theta_2, \text{and} \theta_3$ denote three blocks and that the distribution $\theta_1 y, \theta_3$ is tractable (i.e. can be sampled directly). Then, the blocks $(\theta_1, \theta_2)$ can be collapsed by first sampling $\theta_1$ from $\theta_1 y, \theta_3$ followed by $\theta_2$ from $\theta_2 y, \theta_1, \theta_3$. This amounts to a two block MCMC algorithm. In addition, if it is possible to sample $(\theta_1, \theta_2)$ marginalized over $\theta_3$ then the number of blocks is reduced to one.

Tailored Randomized Block M-H Despite these general precepts, the question of the number of blocks and the composition of the blocks is not always easy to address especially in nonlinear models (such as DSGE models and arbitrage-free term structure models) where no natural grouping of the parameters is suggested by the model structure. Because incorrect grouping can compromise the mixing of the Markov chain, Chib and Ramamurthy (2010) have explored an extension of the above multiple-block algorithm in which the number of blocks and the composition of the blocks is randomized in each iteration. The proposal density of each of the blocks in each iteration is determined by tailoring. They call the resulting algorithm the TaRB-MH algorithm (for Tailored Randomized Block M-H). The algorithm has the following form.

\[
a_i(\theta_1^{(g)}, \theta_1^{(g)} | y, \theta_2^{(g)}) = \min \left\{ \frac{x(\theta_1^{(g)} | y, \theta_2^{(g)}) q_1(\theta_1^{(g)} | y, \theta_2^{(g)})}{x(\theta_2^{(g)} | y, \theta_2^{(g)}) q_1(\theta_1^{(g)} | y, \theta_2^{(g)})}, 1 \right\}
\]

where $q_1(\theta_1^{(g)} | y, \theta_2^{(g)})$ is the proposal density for the $i$th block, and similarly for the second block, implying that if the 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 MH algorithm (in which each block is proposed using its full conditional distribution) is called the Gibbs sampling algorithm.
Algorithm 1. TaRB-MH algorithm

**Step 1:** Initialize $\theta^{(0)}$

**Step 2:** In each iteration $g$, $g = 1,..., n_0 + G$

- **Randomize:** Randomly generate blocks $(\theta_{g, 1},, \theta_{g, 2}, ..., \theta_{g, p_g})$
- **M-H:** Sample each block $\theta_{g, l}$, $l = 1,..., p_g$ by the multiple-block M-H algorithm with the proposal density of each block found by tailoring to the target density of that block

**Step 3:** Discard the draws from the first $n_0$ iterations and save the subsequent $G$ draws $\theta^{(n_0+1)}, ..., \theta^{(n_0+G)}$

We illustrate both the fixed blocks and randomized blocks M-H algorithms in the examples.

### 4 Special topics

#### 4.1 MCMC sampling with latent variables

In sampling a given target distribution, 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) and Mira and Tierney (2002).

To fix notations, suppose that $z$ denotes a vector of latent variables and let the modified target distribution be $\pi(\theta, z | y)$. Then, in many cases (see, for example, Li and Tobias, Chapter 6, this volume), the conditional distribution of $\theta$ (or subcomponents of $\theta$) given $z$ are easy to derive. A multiple-block M-H simulation over $\theta$ and $z$ would lead to the sample

$$
\begin{pmatrix}
\theta^{(n_0+1)}, z^{(n_0+1)} \\
\vdots \\
\theta^{(n_0+M)}, z^{(n_0+M)}
\end{pmatrix} \sim \pi(\theta, z | y).
$$

At the end of this process, the sampled draws on $\theta$ are from $\pi(\theta | y)$, which was the original objective.

Early use of this technique, which formed the basis for many subsequent developments, is to be found in Chib (1992) for the Tobit censored regression model, Albert and Chib (1993b) for binary, ordinal, and categorical outcomes, Albert and Chib (1993a) and Chib (1996) for hidden Markov models, and Carlin et al. (1992), Carter and Kohn (1994), and Frühwirth-Schnatter (1994) for state space models.

**Example:** In the binary probit case, the model is cast in the form

$$
z_i \sim \mathcal{N}(x_i \beta, 1),
y_i = I[z_i > 0], i \leq n,
\beta \sim \mathcal{N}(\beta_0, B_0).$$

(5.12)
Then, the Albert-Chib (1993) algorithm proceeds with the sampling of the full conditional distributions

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

where both these distributions are tractable. In particular, the distribution of $\beta$ 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 $\{z_i\}$. It is multivariate normal with mean $\hat{\beta} = B_0^{-1}\beta_0 + \sum_{i=1}^n x_i^T x_i^{-1}$ and variance matrix $B_0^{-1} + \sum_{i=1}^n x_i^T x_i^{-1}$. 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 $y_i$:

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

where the distribution $z_i | y_i, \beta$ is the distribution $z_i | \beta$ truncated by the knowledge of $y_i$; if $y_i = 0$, then $z_i \leq 0$ and if $y_i = 1$, then $z_i > 0$. Thus, one samples $z_i$ from $\text{TN}(-\infty, 0) \{x_i^T \beta, 1\}$ if $y_i = 0$ and from $\text{TN}(0, \infty) \{x_i^T \beta, 1\}$ if $y_i = 1$, where $\text{TN}(a,b)(\mu, \sigma^2)$ denotes the $\text{N}(\mu, \sigma^2)$ distribution truncated to the region $(a, b)$.

5 Estimation Of Density Ordinates

If the full conditional densities are available, then the MCMC output can be used to estimate the posterior marginal density functions (Gelfand and Smith 1990; Tanner and Wong 1987). By definition, the marginal density of $\theta_k$ at the point $\theta_k^*$ is

$$
\pi(\theta_k^* | y) = \int \pi(\theta_k^* | y, \theta_{-k}) \pi(\theta_{-k} | y) d\theta_{-k},
$$

where as before $\theta_{-k}$ is $\theta$ excluding $\theta_k$. Provided the normalizing constant of $\pi(\theta_k^* | y, \theta_{-k})$ is known, the marginal density can be estimated by the sample average

$$
\hat{\pi}(\theta_k^* | y) = G^{-1} \sum_{g=1}^G \pi(\theta_k^* | y, \theta_k^{(g)}).
$$

Gelfand and Smith (1990) refer to this as the Rao-Blackwell method because of the connections with the Rao-Blackwell theorem in classical statistics. Chib (1995) extends this method for estimating the posterior density of $\theta_k$, conditioned on one or more of the remaining blocks.
In implementing an 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. A large literature is available on this topic, for example, Cowles and Rosenthal (1998), Gamerman and Lopes (2006), Robert and Casella (2004), Fan et al. (2006), and Diaconis et al. (2008).

In practice, convergence (or more properly, lack of convergence) is assessed by informal methods based on the sampled output. For example, one can monitor the autocorrelation plots and the inefficiency factors. 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.

The somewhat different issue from assessing non-convergence, that of detecting coding errors, can also be based on the sampled output (Geweke, 2004).

### 7 Marginal Likelihood Computation

Computation of the marginal likelihood is of considerable importance in Bayesian statistics because the marginal likelihood is needed for the Bayesian comparison of models (Carlin and Louis, 2008; Congdon, 2006; Geweke, 2005; Robert, 2001).

Consider the situation in which there are $K$ possible models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ for the observed data defined by the sampling densities $\{p(y \mid \theta_k, \mathcal{M}_k)\}$ and proper prior densities $\{\pi(\theta_k \mid \mathcal{M}_k)\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach, this question is answered by placing prior probabilities $\Pr(\mathcal{M}_k)$ on each of the $K$ models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathcal{M}_1 \mid y), \ldots, \Pr(\mathcal{M}_K \mid y)\}$ conditioned on the data but marginalized over the unknowns $\theta_k$.

Specifically, the posterior probability of $\mathcal{M}_k$ is given by the expression

$$
\Pr(\mathcal{M}_k \mid y) = \frac{\Pr(\mathcal{M}_k) m(y \mid \mathcal{M}_k)}{\sum_{l=1}^K \Pr(\mathcal{M}_l) m(y \mid \mathcal{M}_l)} \propto \Pr(\mathcal{M}_k) m(y \mid \mathcal{M}_k), \quad (k \leq K)
$$

where

$$
m(y \mid \mathcal{M}_k) = \int p(y \mid \theta_k, \mathcal{M}_k) \pi(\theta_k \mid \mathcal{M}_k) d\theta_k
$$

(5.13)

is the marginal density of the data and is called the marginal likelihood of $\mathcal{M}_k$. In words, the posterior probability of $\mathcal{M}_k$ is proportional to the prior probability of $\mathcal{M}_k$ times the marginal likelihood of $\mathcal{M}_k$. The evidence provided by the data about the models under consideration is summarized by the posterior probability of each model.

Often the posterior probabilities are summarized in terms of the posterior odds.
Therefore, as part of any complete Bayesian study, it is necessary to get an estimate of the marginal likelihood. The obvious strategy of estimating the marginal likelihood by sampling the prior and averaging the likelihood $p(y|\theta_i, M_i)$ tends to be highly inefficient because the prior is rarely concentrated enough in the region where the likelihood has mass. This has led to methods that estimate the marginal likelihood by other means. Green (1995) and Carlin and Chib (1995) first developed methods that involve the joint sampling of parameters and models. Recent applications and developments of these methods include Holmes and Mallick (2003), Dellaportas et al. (2006) and Jasra et al. (2007). These model space methods are particularly useful for the problem of variable selection (Cottet et al., 2008; Lamnisos et al., 2009), but beyond that setting they can be difficult to implement. Here we focus on the method of Chib (1995), which is both general and easy to implement.

For notational simplicity, suppress the model index $k$. By virtue of the fact that $m(y)$ is the normalizing constant of the posterior density, we can write

$$m(y) = \frac{p(y|\theta^*)\pi(\theta^*)}{\hat{\pi}(\theta^*|y)},$$

for any given point $\theta^*$ (generally taken to be a high density point such as the posterior mode or mean). This implies that we can estimate the marginal likelihood on the log scale as

$$\log m(y) = \log p(y|\theta^*) + \log \pi(\theta^*) - \log \hat{\pi}(\theta^*|y),$$

where the first two terms on the right-hand side are generally available in closed form and $\hat{\pi}(\theta^*|y)$ is an estimate of the posterior ordinate. Chib (1995) shows how such an estimate can be found.

Suppose that the MCMC simulation is run with $B$ blocks. Let $\theta_i = (\theta_1, \ldots, \theta_i)$ and $\theta_j = (\theta_i, \ldots, \theta_B)$ denote the list of blocks up to $i$ and the set of blocks from $i$ to $B$, respectively, and let $z$ denote any latent data that is included in the sampling. Then, we can write

$$\pi(\theta^*|y) = \pi(\theta_1^*|y) \times \ldots \times \pi(\theta_i^*|y, \Theta_{i-1}) \times \ldots \times \pi(\theta_B^*|y, \Theta_{B-1}),$$

where the typical term is of the form

$$\frac{\Pr(M_i|y)}{\Pr(M_j|y)} = \frac{\Pr(M_i)}{\Pr(M_j)} \frac{m(y|M_i)}{m(y|M_j)},$$

where the first term is the prior odds and the second the Bayes factor of $M_i$ versus $M_j$. 

\[ p. 204 \]
\[ \pi(\theta^*_i | y, \Theta^*_{i-1}) = \int \pi(y, \Theta^*_{i-1}, \theta^{i+1}, z) \pi(\theta^{i+1}, z | y, \Theta^*_{i-1}) d\theta^{i+1} dz. \]

This is the reduced conditional ordinate since one is integrating only over \((\theta^{i+1}, z)\) and the measure is conditioned on \(\Theta^*_{i-1}\). One can estimate each of the ordinates in the marginal-conditional decomposition from the output of the full MCMC and suitably designed reduced MCMC runs.

**Case 1** Consider first the case where the normalizing constant of each full conditional density is known and MCMC sampling is by the Gibbs algorithm. Then, the first term of (5.16) can be estimated by the Rao-Blackwell method. To estimate the typical reduced conditional ordinate, a reduced MCMC run is set up, consisting of the full conditional distributions

\[ \left\{ \pi(\theta_1 | y, \Theta^*_{i-1}, \Theta^{i+1}, z); \ldots; \pi(\theta_p | y, \Theta^*_{i-1}, \Theta_B, \ldots, \Theta_{B-1}, z) ; \pi(z | y, \Theta^*_{i-1}, \Theta^B) \right\}, \]

(5.17)

where the blocks in \(\Theta_{i-1}\) are set equal to \(\Theta^*_{i-1}\). By MCMC theory, the draws on \((\theta^{i+1}, z)\) from this run are from the distribution \(\pi(\theta^{i+1}, z | y, \Theta^*_{i-1})\) and so the reduced conditional ordinate can be estimated as the average

\[ \hat{\pi}(\theta_i | y, \Theta^*_{i-1}) = G^{-1} \sum_{g=1}^G \pi(\theta_i | y, \Theta^*_{i-1}, \Theta^{i+1}(g), z(g)) \]

over the simulated values of \(\theta^{i+1}\) and \(z\) from the reduced run. Each subsequent reduced conditional ordinate that appears in the decomposition (5.16) is estimated in the same way though, conveniently, with fewer and fewer distributions appearing in the reduced runs. Given the marginal and reduced conditional ordinates, the marginal likelihood on the log scale is available as

\[ \log \pi(y | \theta^*_{i-1}) = \log \pi(y | \theta^*) + \log \pi(\theta^*) - \sum_{i=1}^B \log \hat{\pi}(\theta_i | y, \Theta^*_{i-1}), \]

(5.18)

where \(p(y|\theta^*)\) is the density of the data marginalized over the latent data \(z\).

**Case 2** Consider next the case where the normalizing constant of one or more of the full conditional densities is not known and sampling is by the M-H algorithm. In that case, the posterior ordinate can be estimated by a modified method developed by Chib and Jeliazkov (2001). If sampling is conducted in one block by the M-H algorithm with proposal density \(q(\theta^{(\theta)}, \theta^{(\theta+1)})\) and probability of move

\[ a(\theta^{(\theta)}, \theta^{(\theta+1)} | y) = \min \left\{ 1, \frac{\pi(y | \theta^{(\theta)}, \theta^{(\theta+1)}) q(\theta^{(\theta+1)}, \theta^{(\theta)})}{\pi(y | \theta^{(\theta)}, \theta^{(\theta+1)}) q(\theta^{(\theta)}, \theta^{(\theta+1)})} \right\}, \]

where
then it can be shown that the posterior ordinate is given by

\[ \pi(\theta^* | y) = \frac{E_1 \left\{ a(\theta, \theta^* | y) q(\theta^* | y) \right\}}{E_2 \left\{ a(\theta^* | y) \right\}}, \]

where the numerator expectation \( E_1 \) is with respect to the distribution \( \pi(\theta | y) \) and the denominator expectation \( E_2 \) is with respect to the proposal density \( q(\theta^* | y) \). This leads to the simulation consistent estimate

\[ \hat{\pi}(\theta^* | y) = \frac{G^{-1} \sum_{g=1}^{G} a(\theta^*, \theta_g | y) q(\theta_g | y)}{J^{-1} \sum_{j=1}^{J} a(\theta^*, \theta_j | y)}, \]

(5.19)

where \( \theta^{(g)} \) are the given draws from the posterior distribution, while the draws \( \theta^{(j)} \) in the denominator are from \( q(\theta^*, \theta | y) \), given the fixed value \( \theta^* \).

In general, when sampling is done with \( B \) blocks, the typical reduced conditional ordinate is given by

\[ \pi(\theta^* | y, \Theta_{-i}) = \frac{E_1 \left\{ a(\theta_i, \theta^*_{i-1}, \theta^*_{i+1}, x) q_i(\theta_i, \theta^*_{i-1}, \theta^*_{i+1} | y, \Theta_{-i}, x) \right\}}{E_2 \left\{ a(\theta^*, \theta_i | y, \Theta_{-i}, x) \right\}}, \]

(5.20)

where \( E_1 \) is the expectation with respect to \( \pi(\theta^*_i, x | y, \Theta_{-i}) \) and \( E_2 \) that with respect to the product measure \( \pi(\Theta_{i+1} | y, \Theta_{i-1}, \theta^*_i, \theta^*_{-i}) \). The quantity \( a(\theta_i, \theta^*_i | y, \Theta_{-i}, \theta^*_{i+1}, x) \) is the M-H probability of move. The two expectations are estimated from the output of the reduced runs in an obvious way.

### 8 Examples

#### 8.1 Jump-diffusion model

To illustrate use of MCMC methods in a practical problem, we consider an example that is drawn from Chan and Wong (2006). The fitting of this model highlights several interesting MCMC aspects:

- multiple block MCMC sampling in which each block is updated by proposing from its (tractable) full conditional distribution;
- marginalization to improve the efficiency of the sampling;
- treatment of mixture distributions;
- involvement of latent variables to simplify sampling.

In the model we consider, it is assumed that returns (in continuous time) are given by
\begin{align*}
d\log r_t &= \mu dt + J_t dN_t + \sigma dW_t,
\end{align*}

where

\[ J_t \mid \theta \sim \mathcal{N}(k, s^2) \]

is the jump component, \( N_t \) is a Poisson process with intensity \( \lambda \), \( W_t \) is the Wiener process, and

\[ \theta = (\mu, k, \sigma^2, s^2, \lambda) \]

the parameters.

Now suppose that this process is observed at the \( n \) equi-spaced time points \( t_1, \ldots, t_n \) over the interval \( (0, T) \), where \( t_{i+1} = t_i + \Delta \) (\( i = 0, 1, \ldots, n \)), \( t_0 = 0 \), \( t_n = T \), and \( \Delta = T/n \). Then, taking the Euler discretization, we have the model

\[ \log r_{t_{i+1}} - \log r_{t_i} = \mu \Delta + J_i (N_{t_{i+1}} - N_i) + \sigma (W_{t_{i+1}} - W_i), \]

which we can denote as

\[ y_t = \mu \Delta + J_i \Delta N_i + \varepsilon_t, \]

where from the properties of the Poisson and Wiener processes

\[ \Delta N_i \mid \theta \sim \mathcal{B}(q), \]

a Bernoulli random variable with probability of success \( q = \lambda \Delta \) and

\[ \varepsilon_t \mid \theta \sim \mathcal{N}(0, \sigma^2 \Delta). \]

Under these assumptions, it is obvious that
\[ p(y_t | J_t, \Delta N_t, \theta) = \mathcal{N}(y_t | \mu \Delta + J_t \Delta N_t, \sigma^2 \Delta). \]

Now given the form of the model in (5.22), it is clear that the prior-posterior analysis will be aided by including \( \{J_t\} \) and \( \{\Delta N_t\} \) in the sampling. But one has to be careful, because \( J_t \) and \( \Delta N_t \) appear as a product and neither one is observed. As a result, sampling one conditioned on the other is likely to produce highly correlated output. This point is made by Chib et al. (2006) in the context of factor models. The solution is to sample \( \Delta N_t \) marginalized over \( J_t \). Marginalization over \( J_t \) also helps in the sampling of \( \mu \), as noted by Chib and Carlin (1999) in the context of panel models where the equivalent of \( J_t \) is the random effect. With this in mind, the distribution of \( y_t \) marginalized over \( J_t \) can be calculated as

\[ p(y_t | \Delta N_t, \theta) = \mathcal{N}(y_t | \mu \Delta, \sigma^2 \Delta). \]

Here we made use of the fact that \( (\Delta N_t)^2 = \Delta N_t \). A final integration over the distribution of \( \Delta N_t \) shows that the density of \( y_t \) is a two-component mixture of normal distributions

\[ p(y_t | \theta) = qp(y_t | \Delta N_t = 1, \theta) + (1 - q)p(y_t | \Delta N_t = 0, \theta) = q\mathcal{N}(y_t | \mu \Delta + k, \sigma^2 \Delta + s^2) + (1 - q)\mathcal{N}(y_t | \mu \Delta, \sigma^2 \Delta). \]

Therefore, given independently distributed outcomes \( y = (y_1, \ldots, y_n) \), the joint distribution of the outcomes is

\[ p(y | \theta) = \prod_{t=1}^n \left\{ q\mathcal{N}(y_t | \mu \Delta + k, \sigma^2 \Delta + s^2) + (1 - q)\mathcal{N}(y_t | \mu \Delta, \sigma^2 \Delta) \right\}. \]

Even though this is a mixture model, the usual problem of "label-switching" does not arise here because the first component, which is unambiguously different from the second, occurs with probability \( q \) that can be expected to be substantially smaller than \( 1 - q \). Thus, the first component, in which \( \Delta N_t = 1 \), cannot be switched with the second component, in which \( \Delta N_t = 0 \), without changing the probability distribution of the data.

To show how this model can be estimated, we simulate data from this model under the assumption that \( T = 10 \) (for ten years), \( \Delta = 1/250 \) (corresponding to daily data measured in years assuming 250 trading days), \( \mu = 0.08 \) (corresponding to an 8% annual rate of return), \( k = 0, s^2 = 0.15, \lambda = 5 \) (corresponding to five jumps on...
average every year) and \( \sigma = 0.3 \) (corresponding to an annual volatility of 30%). The \( n = 2,500 \) observations that we have simulated are given in Figure 5.7.

For doing inference, suppose that the following prior distribution is appropriate:

\[
\beta = (\mu, k) \sim \mathcal{N}(\beta_0, B_0) \\
\sigma^2 \sim \mathcal{IG}(v_0, \delta_0) \\
s^2 \sim \mathcal{IG}(v_{00}, \delta_{00}) \\
q \sim \text{Beta}(a_0, b_0),
\]

where \( \beta_0 = (0.05, 0) \), \( B_0 = \text{diag}(0.01, 0.01) \), \( v_0 = 12 \), \( \delta_0 = 1 \), \( v_{00} = 4.889 \), \( \delta_{00} = 0.058 \), \( a_0 = 0.9 \), and \( b_0 = 12.1 \). Our MCMC algorithm is defined in terms of four blocks as follows.

**Figure 5.7**

Simulated data from the jump-diffusion model

*Notes: Observations that include the jump component are circled.*

(1) Sample \((\beta, q)\) from \(\pi(\beta, q|\{y_t\}, \{\Delta N_t\}, \sigma^2, s^2)\). On combining the model in (5.23) with the prior of \(\beta\) and \(q\), one sees that \(\beta\) and \(q\) are a posteriori independent and that

\[
\pi(\beta, q|\{y_t\}, \{\Delta N_t\}, \sigma^2, s^2) = \pi(\beta|\{y_t\}, \{\Delta N_t\}, \sigma^2) \pi(q|\{\Delta N_t\}) \\
= \mathcal{N}(\hat{\beta}, B^0) \text{Beta}(q|a_0 + n_1, b_0 + (n - n_1)),
\]

where

\[
B = \left( B_0^{-1} + \sum_{t=1}^{n} x_t V_t^{-1} x_t^T \right)^{-1} \\
\hat{\beta} = B \left( B_0^{-1} \beta_0 + \sum_{t=1}^{n} y_t V_t^{-1} x_t \right)
\]
and \( n_1 \) is the sum of \( \Delta N_t \) (i.e. the number of jumps in the sample).

(2) Sample \( \Delta N_t \) from \( \Delta N_t | y_t, \theta \) for \( t = 1, \ldots, n \). Since \( \Delta N_t \) takes the values 1 or 0, with prior probabilities \( q \) and \( 1 - q \) respectively, it follows by Bayes theorem that the updated probabilities are

\[
\Pr(\Delta N_t = 1 | y_t, \theta) \propto q \Pr(y_t | \Delta N_t = 1, \theta) \propto q \mathcal{N}(\mu \Delta + k, \sigma^2 \Delta + s^2)
\]

\[
\Pr(\Delta N_t = 0 | y_t, \theta) \propto (1 - q) \Pr(y_t | \Delta N_t = 0, \theta) \propto (1 - q) \mathcal{N}(\mu \Delta, \sigma^2 \Delta).
\]

(3) Sample \( J_t \) (whenever \( \Delta N_t = 1 \)) from \( J_t | \Delta N_t = 1, \Delta N_t = 1, \theta \). This distribution can be derived as a regression update from the model

\[
y_t - \mu \Delta = J_t + \epsilon_t.
\]

A simple calculation shows that this distribution is

\[
\mathcal{N}\left(J_t, Q_t\right),
\]

where

\[
Q_t = \left(s^{-2} + \left(\sigma^2 \Delta\right)^{-1}\right)^{-1}
\]

and

\[
J_t = Q_t \left(s^{-2} k + \left(\sigma^2 \Delta\right)^{-1} (y_t - \mu \Delta)\right).
\]

(4) Sample \( \sigma^2 \) and \( s^2 \) from \((\sigma^2, s^2)|\{y_t\},\{\Delta N_t\},\{J_t\}, \beta, q\). Again by well-known Bayesian calculations it is easily checked that \( \sigma^2 \) and \( s^2 \) are independent and that

\[
\pi(\sigma^2, s^2 | \{y_t\}, \{\Delta N_t\}, \{J_t\}, \beta, q) = \pi(\sigma^2 | \{y_t\}, \{\Delta N_t\}, \{J_t\}, \beta) \pi(s^2 | \{J_t\}, \beta)
\]

\[
= \mathcal{G}(\sigma^2 | \frac{v_{0\sigma} + n}{2}, \frac{\delta_{0\sigma} + \frac{1}{2} \sum (y_t - \mu \Delta - J_t)^2}{2}) \mathcal{G}(s^2 | \frac{v_{0s} + n}{2}, \frac{\delta_{0s} + \frac{1}{2} \sum (y_t - J_t)^2}{2}).
\]

We run our MCMC algorithm for 21,000 iterations, and drop the first 1,000 as part of the burn-in. We summarize the posterior distribution by calculating the sample mean, standard deviation, and 0.025 and 0.975 quantiles from the sampled draws. These are the MCMC estimates of the corresponding posterior parameters. We also calculate the inefficiency factors from the sampled output. These results, along with the prior mean and prior standard deviations, are given in Table 5.1. These results show that in this problem
it is difficult to estimate the parameter $\mu$ given the noise in the sample, but that the parameter estimates are still close to the true values that were used to generate the data. The inefficiency factors in the last column of the table are small and show that the sampler has mixed almost as well as a hypothetical independence sample. We also use the sampled output to calculate the marginal posterior densities of the parameters. These are calculated by kernel smoothing and are reported in the top panel of Figure 5.8. The bottom panel of the figure has the autocorrelation functions from the sampled draws. These decline quickly, as one would expect given the small values of the inefficiency factors.

Table 5.1: Posterior summary: jump-diffusion model

<table>
<thead>
<tr>
<th></th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.050</td>
<td>0.100</td>
</tr>
<tr>
<td>$k$</td>
<td>0.000</td>
<td>0.100</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.100</td>
<td>0.050</td>
</tr>
<tr>
<td>$s^2$</td>
<td>0.040</td>
<td>0.020</td>
</tr>
<tr>
<td>$q$</td>
<td>0.050</td>
<td>0.050</td>
</tr>
</tbody>
</table>

8.2 DSGE model

As another interesting illustration, we consider the application of MCMC methods in the analysis of DSGE models (Del Negro and Schorfheide, Chapter 7, this volume). An interesting feature of DSGE models is that because of the complex process that is used to solve the model, the structural parameters appear in the reduced form model in a highly nonlinear way that cannot be described in closed form. This nonlinearity poses a serious challenge for inference. The method of maximum likelihood does not always deliver parameter estimates that are reasonable on a priori grounds. Interestingly, the Bayesian approach provides a useful alternative to frequentist approaches because the prior distribution can be used to concentrate attention on regions of the parameter space that are economically meaningful.

To demonstrate the issues involved, we consider the relatively small-scale DSGE model in Ireland (2004). The linearized model, derived by log-linearizing the original non-linear model around its deterministic steady state (Ireland, 2004), is expressed by the following system of equations

$$
\begin{align*}
\hat{x}_t &= \alpha \hat{x}_{t-1} + (1 - \alpha) E \hat{x}_{t+1} - (\hat{r}_t - E \hat{\pi}_{t+1}) + (1 - \omega) (1 - \rho_a) \hat{a}_t \\
\hat{r}_t &= \beta_0 \hat{\pi}_{t-1} + \beta (1 - \alpha) E \hat{\pi}_{t+1} + \psi \hat{a}_t - \hat{\pi}_t \\
\hat{\pi}_t &= \hat{y}_t - \hat{y}_{t-1} + \hat{\epsilon}_t \\
\hat{a}_t &= \hat{y}_t - \omega \hat{a}_t \\
\hat{\pi}_t &= \rho_1 \hat{\pi}_{t-1} + \rho_2 \hat{\pi}_{t-1} + \rho_3 \hat{\pi}_t + \rho_4 \hat{\pi}_1 + \epsilon_{\pi,t} \\
\end{align*}
$$

(5.25)
where $x_t$, $g_t$, $\pi_t$, $r_t$, and $x_t$ denote output gap, output growth, inflation, nominal interest rate, and stochastically detrended output, respectively, and the hats denote log-deviation of the variables from their steady-state or average values; $E_t$ denotes the expectation of the relevant variables by the agents in this economy, formed under rational expectations; finally, $\hat{a}_t$, $\hat{e}_t$, and $\hat{z}_t$ capture exogenous shifts in preferences, costs of production, and technology, respectively. These equations, in descending order, represent a forward-looking IS curve, a new Keynesian Phillips curve, growth rate of output, growth rate of output gap, and the modified Taylor rule (1993). Further, it is assumed that the exogenous driving processes $\hat{a}_t$, $\hat{e}_t$, and $\hat{z}_t$ evolve independently of one another as

\begin{align}
\hat{a}_t &= \rho_a \hat{a}_{t-1} + \epsilon_{a,t} \\
\hat{e}_t &= \rho_e \hat{e}_{t-1} + \epsilon_{e,t} \\
\hat{z}_t &= \epsilon_{z,t} \\
\end{align}

(5.26)

The innovations in the model $\epsilon_t = [\epsilon_{a,t}, \epsilon_{e,t}, \epsilon_{z,t}, \epsilon_{r,t}]'$ are assumed to be distributed as multivariate normal

$$\epsilon_t \sim \mathcal{N}(0, \Omega).$$

where $\Omega = \text{diag}(\sigma_a^2, \sigma_e^2, \sigma_z^2, \sigma_r^2)$. There are two other parameters $z$ and $\pi$ in the nonlinear model (that do not appear in the linearized form) that determine the steady-state values of output growth and inflation, respectively. In addition, $\beta$ determines the steady-state value of the short-term nominal interest rate through the relation $r = z \pi / \beta$. Following Ireland (2004), the values of $z$, $\pi$, and $\beta$ are set to the average levels of output growth, inflation, and interest rates in the data. Also, $\gamma$ and $\rho_r$ are fixed at 0.10 and 1.00, respectively. We collect the remaining 12 parameters of interest in the vector $\theta$

$$\theta = (\omega, \alpha_x, \alpha_\pi, \rho_\pi, \rho_g, \rho_x, \rho_a, \rho_e, \sigma_a, \sigma_e, \sigma_z, \sigma_r).$$
The parameters \((\omega, \alpha_x, \alpha_\pi, \rho_\pi, \rho_g, \rho_x)\), where \(\omega, \alpha_x, \alpha_\pi\) are each assumed to be between 0 and 1, and each of \(\rho_\pi, \rho_g, \rho_x\) are greater than 0, may be called the structural parameters. It is also assumed that \((\rho_\pi, \rho_\varphi)\) each lie between 0 and 1. Let \(S_p\) denote the subset of \(\mathbb{R}^2\) satisfying these linear constraints. In addition, one is interested in restricting the parameters to the determinacy region of the solution space. Denote this constraint set by \(S_D\). Finally, the variance parameters \(\sigma^2_i\) lie in the region \(S_\Omega\) that satisfies the usual positivity and positive definiteness constraints.

It is now assumed that for a given value of \(\theta\), the model is solved subject to the determinacy constraint to produce a Markov process in the endogenous variables of the model

\[
s_t = D(\theta) s_{t-1} + F(\theta) \varepsilon_t,
\]

where

\[
s_t = [\hat{y}_t, \hat{r}_t, \hat{\pi}_t, \hat{g}_t, \hat{x}_t, \hat{a}_t, \hat{e}_t, \hat{z}_t, E_t \hat{\pi}_{t+1}, E_t \hat{x}_{t+1}]'
\]

and \(\varepsilon_t\) is as defined above. Moreover, the matrices \(D(\theta)\) and \(F(\theta)\) are awkward implicit functions of the model parameters, obtained (numerically) from the solution.

The data for the fitting are the series of demeaned log-deviations of output growth \(\hat{g}_t\), inflation \(\hat{\pi}_t\), and the short-term nominal interest rate \(\hat{r}_t\) from their steady-state or average values for the period 1980:1 to 2003:1, with all three observables measured in decimal units. The resulting measurement equation has the straightforward form

\[
\begin{bmatrix}
\hat{y}_t \\
\hat{r}_t \\
\hat{\pi}_t
\end{bmatrix} =
\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
S_r
\end{bmatrix}
\]

Given this state space representation, the likelihood function is calculated by the usual Kalman filtering recursions. To complete the model, one needs to specify a prior distribution for the parameters, which, as mentioned above, can play an important role in the estimation of DSGE models. We construct our prior by reasoning in terms of the implied distribution of the data. Our prior is summarized in Table 5.2. This prior implies a quarterly deviation, as measured by the 90% interval, of roughly 4.5% for output and the rate of interest, and around 5% in the case of inflation.
Table 5.2: Posterior sampling results using the TaRB-MH algorithm for the Ireland (2004) model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>ω</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td>α_x</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>α_n</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>ρ_n</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>ρ_g</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>ρ_x</td>
<td>0.25</td>
<td>0.0625</td>
</tr>
<tr>
<td>ρ_a</td>
<td>0.85</td>
<td>0.10</td>
</tr>
<tr>
<td>ρ_e</td>
<td>0.85</td>
<td>0.10</td>
</tr>
<tr>
<td>10000σ_a²</td>
<td>30.00</td>
<td>30.00</td>
</tr>
<tr>
<td>10000σ_e²</td>
<td>0.08</td>
<td>1.00</td>
</tr>
<tr>
<td>10000σ_r²</td>
<td>5.00</td>
<td>15.00</td>
</tr>
<tr>
<td>10000σ_r²</td>
<td>0.50</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Note: The results reported in this table are based on the prior mean as the starting value. However, the results are insensitive to this choice.

Our interest lies in the posterior distribution resulting from the combination of the likelihood function and the prior distribution. We summarize the posterior distribution with the help of the TaRB-MH algorithm of Chib and Ramamurthy (2010). As discussed above, in this algorithm, the parameters are updated in blocks by the M-H algorithm, but the number and composition of the blocks is randomized in every MCMC iteration. The M-H proposal density of each block is in turn obtained by tailoring to the target, as described earlier in the context of the choice of proposal density in Section 3. That is, the proposal density takes the form of a multivariate student-t with 15 degrees of freedom and with location and dispersion that are the mode and negative inverse Hessian of the target. In the current problem, the latter quantities are found from the output of simulated annealing, a powerful stochastic optimization method that proves helpful when standard gradient-based optimizers are difficult to implement (for a detailed discussion on the
implementation of simulated annealing, the reader is referred to Chib and Ramamurthy, 2010). The TaRB-MH chain is initialized at the prior mean and run for 11,000 iterations. The first 1,000 draws are discarded as part of the burn-in.

The posterior summary from this algorithm is given in Table 5.2. As can be seen from the table, the inefficiency factors are all mostly in the single digits, indicating a well-mixing chain. Figure 5.9 gives the prior-posterior plots and autocorrelation functions for the structural and autoregressive parameters of the model. As can be observed, for most parameters, the likelihood function carries information beyond that contained in the prior. These plots also show that the serial correlations among the draws decay quickly to zero. The success of the MCMC sampling is further revealed by the fact that the results are virtually identical, regardless of where the chain is initialized.

**Figure 5.9**

Sampling results for the Ireland (2004) model using the TaRB-MH algorithm: marginal prior-posterior plots and autocorrelation functions for the structural and autoregressive parameters

**9 Concluding remarks**

It should be clear from this survey that the importance of MCMC methods for Bayesian statistics and econometrics cannot be overstated. The remarkable growth of Bayesian thinking over the last 20 years has been made possible largely by the innovative use of MCMC methods, as the chapters in this book so amply confirm.
References


