CHAPTER 12

Markov Chain Monte Carlo Methods with Applications

Advances in computing facilities and computational methods have dramatically increased our ability to solve complicated problems. The advances also extend the applicability of many existing econometric and statistical methods. Examples of such achievements in statistics include the Markov chain Monte Carlo (MCMC) method and data augmentation. These techniques enable us to make some statistical inference that was not feasible just a few years ago. In this chapter, we introduce the ideas of MCMC methods and data augmentation that are widely applicable in finance. In particular, we discuss Bayesian inference via Gibbs sampling and demonstrate various applications of MCMC methods. Rapid developments in the MCMC methodology make it impossible to cover all the new methods available in the literature. Interested readers are referred to some recent books on Bayesian and empirical Bayesian statistics (e.g., Carlin and Louis, 2000; Gelman, Carlin, Stern, and Rubin, 2003).

For applications, we focus on issues related to financial econometrics. The demonstrations shown in this chapter represent only a small fraction of all possible applications of the techniques in finance. As a matter of fact, it is fair to say that Bayesian inference and the MCMC methods discussed here are applicable to most, if not all, of the studies in financial econometrics.

We begin the chapter by reviewing the concept of a *Markov process*. Consider a stochastic process $\{X_t\}$, where each X_t assumes a value in the space $\mathbf{\Theta}$. The process $\{X_t\}$ is a Markov process if it has the property that, given the value of X_t, the values of X_h, $h > t$, do not depend on the values X_s, $s < t$. In other words, $\{X_t\}$ is a Markov process if its conditional distribution function satisfies

$$P(X_h|X_s, s \le t) = P(X_h|X_t), \qquad h > t.$$

Analysis of Financial Time Series, Third Edition, By Ruey S. Tsay

If $\{X_t\}$ is a discrete-time stochastic process, then the prior property becomes

$$P(X_h|X_t, X_{t-1}, \ldots) = P(X_h|X_t), \qquad h > t.$$

Let $\boldsymbol{A}$ be a subset of $\mathbf{\Theta}$. The function

$$P_t(\boldsymbol{\theta}, h, \boldsymbol{A}) = P(X_h \in \boldsymbol{A}|X_t = \boldsymbol{\theta}), \qquad h > t$$

is called the transition probability function of the Markov process. If the transition probability depends on $h - t$, but not on t, then the process has a stationary transition distribution.

12.1 MARKOV CHAIN SIMULATION

Consider an inference problem with parameter vector $\boldsymbol{\theta}$ and data $\boldsymbol{X}$, where $\boldsymbol{\theta} \in \mathbf{\Theta}$. To make inference, we need to know the distribution $P(\boldsymbol{\theta}|\boldsymbol{X})$. The idea of Markov chain simulation is to simulate a Markov process on $\mathbf{\Theta}$, which converges to a stationary transition distribution that is $P(\boldsymbol{\theta}|\boldsymbol{X})$.

The key to Markov chain simulation is to create a Markov process whose stationary transition distribution is a specified $P(\boldsymbol{\theta}|\boldsymbol{X})$ and run the simulation sufficiently long so that the distribution of the current values of the process is close enough to the stationary transition distribution. It turns out that, for a given $P(\boldsymbol{\theta}|\boldsymbol{X})$, many Markov chains with the desired property can be constructed. We refer to methods that use Markov chain simulation to obtain the distribution $P(\boldsymbol{\theta}|\boldsymbol{X})$ as MCMC methods.

The development of MCMC methods took place in various forms in the statistical literature. Consider the problem of "missing value" in data analysis. Most statistical methods discussed in this book were developed under the assumption of "complete data" (i.e., there is no missing value). For example, in modeling daily volatility of an asset return, we assume that the return data are available for all trading days in the sample period. What should we do if there is a missing value?

Dempster, Laird, and Rubin (1977) suggest an iterative method called the Expectation-Maximization (EM) algorithm to solve the problem. The method consists of two steps. First, if the missing value were available, then we could use methods of complete-data analysis to build a volatility model. Second, given the available data and the fitted model, we can derive the statistical distribution of the missing value. A simple way to fill in the missing value is to use the conditional expectation of the derived distribution of the missing value. In practice, one can start the method with an arbitrary value for the missing value and iterate the procedure for many many times until convergence. The first step of the prior procedure involves performing the maximum-likelihood estimation of a specified model and is called the M-step. The second step is to compute the conditional expectation of the missing value and is called the E-step.

Tanner and Wong (1987) generalize the EM algorithm in two ways. First, they introduce the idea of iterative simulation. For instance, instead of using the conditional expectation, one can simply replace the missing value by a random draw

from its derived conditional distribution. Second, they extend the applicability of the EM algorithm by using the concept of data augmentation. By data augmentation, we mean adding auxiliary variables to the problem under study. It turns out that many of the simulation methods can often be simplified or speeded up by data augmentation; see the application sections of this chapter.

12.2 GIBBS SAMPLING

Gibbs sampling (or Gibbs sampler) of Geman and Geman (1984) and Gelfand and Smith (1990) is perhaps the most popular MCMC method. We introduce the idea of Gibbs sampling by using a simple problem with three parameters. Here the word *parameter* is used in a very general sense. A missing data point can be regarded as a parameter under the MCMC framework. Similarly, an unobservable variable such as the "true" price of an asset can be regarded as N parameters when there are N transaction prices available. This concept of parameter is related to data augmentation and becomes apparent when we discuss applications of the MCMC methods.

Denote the three parameters by θ_1, θ_2, and θ_3. Let $\boldsymbol{X}$ be the collection of available data and M the entertained model. The goal here is to estimate the parameters so that the fitted model can be used to make inference. Suppose that the likelihood function of the model is hard to obtain, but the three conditional distributions of a single parameter given the others are available. In other words, we assume that the following three conditional distributions are known:

$$f_1(\theta_1|\theta_2, \theta_3, \boldsymbol{X}, M), \qquad f_2(\theta_2|\theta_3, \theta_1, \boldsymbol{X}, M), \qquad f_3(\theta_3|\theta_1, \theta_2, \boldsymbol{X}, M), \tag{12.1}$$

where $f_i(\theta_i|\theta_{j\neq i}, \boldsymbol{X}, M)$ denotes the conditional distribution of the parameter θ_i given the data, the model, and the other two parameters. In application, we do not need to know the exact forms of the conditional distributions. What is needed is the ability to draw a random number from each of the three conditional distributions.

Let $\theta_{2,0}$ and $\theta_{3,0}$ be two arbitrary starting values of θ_2 and θ_3. The Gibbs sampler proceeds as follows:

1. Draw a random sample from $f_1(\theta_1|\theta_{2,0}, \theta_{3,0}, \boldsymbol{X}, M)$. Denote the random draw by $\theta_{1,1}$.
2. Draw a random sample from $f_2(\theta_2|\theta_{3,0}, \theta_{1,1}, \boldsymbol{X}, M)$. Denote the random draw by $\theta_{2,1}$.
3. Draw a random sample from $f_3(\theta_3|\theta_{1,1}, \theta_{2,1}, \boldsymbol{X}, M)$. Denote the random draw by $\theta_{3,1}$.

This completes a Gibbs iteration and the parameters become $\theta_{1.1}$, $\theta_{2,1}$, and $\theta_{3,1}$.

Next, using the new parameters as starting values and repeating the prior iteration of random draws, we complete another Gibbs iteration to obtain the updated

parameters $\theta_{1,2}$, $\theta_{2,2}$, and $\theta_{3,2}$. We can repeat the previous iterations for m times to obtain a sequence of random draws:

$$(\theta_{1,1}, \theta_{2,1}, \theta_{3,1}), \ldots, (\theta_{1,m}, \theta_{2,m}, \theta_{3,m}).$$

Under some regularity conditions, it can be shown that, for a sufficiently large m, $(\theta_{1,m}, \theta_{2,m}, \theta_{3,m})$ is approximately equivalent to a random draw from the joint distribution $f(\theta_1, \theta_2, \theta_3|\boldsymbol{X}, M)$ of the three parameters. The regularity conditions are weak; they essentially require that for an arbitrary starting value $(\theta_{1,0}, \theta_{2,0}, \theta_{3,0})$, the prior Gibbs iterations have a chance to visit the full parameter space. The actual convergence theorem involves using the Markov chain theory; see Tierney (1994).

In practice, we use a sufficiently large n and discard the first m random draws of the Gibbs iterations to form a Gibbs sample, say,

$$(\theta_{1,m+1}, \theta_{2,m+1}, \theta_{3,m+1}), \ldots, (\theta_{1,n}, \theta_{2,n}, \theta_{3,n}). \tag{12.2}$$

Since the previous realizations form a random sample from the joint distribution $f(\theta_1, \theta_2, \theta_3|\boldsymbol{X}, M)$, they can be used to make inference. For example, a point estimate of θ_i and its variance are

$$\widehat{\theta}_i = \frac{1}{n-m}\sum_{j=m+1}^{n} \theta_{i,j}, \qquad \widehat{\sigma}_i^2 = \frac{1}{n-m-1}\sum_{j=m+1}^{n} (\theta_{i,j} - \widehat{\theta}_i)^2. \tag{12.3}$$

The Gibbs sample in Eq. (12.2) can be used in many ways. For example, if we are interested in testing the null hypothesis $H_0 : \theta_1 = \theta_2$ versus the alternative hypothesis $H_a : \theta_1 \neq \theta_2$, then we can simply obtain the point estimate of $\theta = \theta_1 - \theta_2$ and its variance as

$$\widehat{\theta} = \frac{1}{n-m}\sum_{j=m+1}^{n} (\theta_{1,j} - \theta_{2,j}), \qquad \widehat{\sigma}^2 = \frac{1}{n-m-1}\sum_{j=m+1}^{n} (\theta_{1,j} - \theta_{2,j} - \widehat{\theta})^2.$$

The null hypothesis can then be tested by using the conventional t-ratio statistic $t = \widehat{\theta}/\widehat{\sigma}$.

Remark. The first m random draws of a Gibbs sampling, which are discarded, are commonly referred to as the *burn-in* sample. The burn-ins are used to ensure that the Gibbs sample in Eq. (12.2) is indeed close enough to a random sample from the joint distribution $f(\theta_1, \theta_2, \theta_3|\boldsymbol{X}, M)$. □

Remark. The method discussed before consists of running a single long chain and keeping all random draws after the burn-ins to obtain a Gibbs sample. Alternatively, one can run many relatively short chains using different starting values and a relatively small n. The random draw of the last Gibbs iteration in each chain is then used to form a Gibbs sample. □

From the prior introduction, Gibbs sampling has the advantage of decomposing a high-dimensional estimation problem into several lower dimensional ones via full conditional distributions of the parameters. At the extreme, a high-dimensional problem with N parameters can be solved iteratively by using N univariate conditional distributions. This property makes the Gibbs sampling simple and widely applicable. However, it is often not efficient to reduce all the Gibbs draws into a univariate problem. When parameters are highly correlated, it pays to draw them jointly. Consider the three-parameter illustrative example. If θ_1 and θ_2 are highly correlated, then one should employ the conditional distributions $f(\theta_1, \theta_2|\theta_3, \boldsymbol{X}, M)$ and $f_3(\theta_3|\theta_1, \theta_2, \boldsymbol{X}, M)$ whenever possible. A Gibbs iteration then consists of (a) drawing jointly (θ_1, θ_2) given θ_3, and (b) drawing θ_3 given (θ_1, θ_2). For more information on the impact of parameter correlations on the convergence rate of a Gibbs sampler, see Liu, Wong, and Kong (1994).

In practice, convergence of a Gibbs sample is an important issue. The theory only states that the convergence occurs when the number of iterations m is sufficiently large. It provides no specific guidance for choosing m. Many methods have been devised in the literature for checking the convergence of a Gibbs sample. But there is no consensus on which method performs best. In fact, none of the available methods can guarantee 100% that the Gibbs sample under study has converged for all applications. Performance of a checking method often depends on the problem at hand. Care must be exercised in a real application to ensure that there is no obvious violation of the convergence requirement; see Carlin and Louis (2000) and Gelman et al. (2003) for convergence checking methods. In application, it is important to repeat the Gibbs sampling several times with different starting values to ensure that the algorithm has converged.

12.3 BAYESIAN INFERENCE

Conditional distributions play a key role in Gibbs sampling. In the statistical literature, these conditional distributions are referred to as *conditional posterior distributions* because they are distributions of parameters given the data, other parameter values, and the entertained model. In this section, we review some well-known posterior distributions that are useful in using MCMC methods.

12.3.1 Posterior Distributions

There are two approaches to statistical inference. The first approach is the classical approach based on the maximum-likelihood principle. Here a model is estimated by maximizing the likelihood function of the data, and the fitted model is used to make inference. The other approach is Bayesian inference that combines prior belief with data to obtain posterior distributions on which statistical inference is based. Historically, there were heated debates between the two schools of statistical inference. Yet both approaches have proved to be useful and are now widely accepted. The methods discussed so far in this book belong to the classical approach. However,

Bayesian solutions exist for all of the problems considered. This is particularly so in recent years with the advances in MCMC methods, which greatly improve the feasibility of Bayesian analysis. Readers can revisit the previous chapters and derive MCMC solutions for the problems considered. In most cases, the Bayesian solutions are similar to what we had before. In some cases, the Bayesian solutions might be advantageous. For example, consider the calculation of value at risk in Chapter 7. A Bayesian solution can easily take into consideration the parameter uncertainty in VaR calculation. However, the approach requires intensive computation.

Let $\boldsymbol{\theta}$ be the vector of unknown parameters of an entertained model and $\boldsymbol{X}$ be the data. Bayesian analysis seeks to combine knowledge about the parameters with the data to make inference. Knowledge of the parameters is expressed by specifying a *prior* distribution for the parameters, which is denoted by $P(\boldsymbol{\theta})$. For a given model, denote the likelihood function of the data by $f(\boldsymbol{X}|\boldsymbol{\theta})$. Then by the definition of conditional probability,

$$f(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{X})}{f(\boldsymbol{X})} = \frac{f(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{f(\boldsymbol{X})}, \tag{12.4}$$

where the marginal distribution $f(\boldsymbol{X})$ can be obtained by

$$f(\boldsymbol{X}) = \int f(\boldsymbol{X}, \boldsymbol{\theta})\, d\boldsymbol{\theta} = \int f(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

The distribution $f(\boldsymbol{\theta}|\boldsymbol{X})$ in Eq. (12.4) is called the *posterior distribution* of $\boldsymbol{\theta}$. In general, we can use Bayes's rule to obtain

$$f(\boldsymbol{\theta}|\boldsymbol{X}) \propto f(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{12.5}$$

where $P(\boldsymbol{\theta})$ is the prior distribution and $f(\boldsymbol{X}|\boldsymbol{\theta})$ is the likelihood function. From Eq. (12.5), making statistical inference based on the likelihood function $f(\boldsymbol{X}|\boldsymbol{\theta})$ amounts to using a Bayesian approach with a constant prior distribution.

12.3.2 Conjugate Prior Distributions

Obtaining the posterior distribution in Eq. (12.4) is not simple in general, but there are cases in which the prior and posterior distributions belong to the same family of distributions. Such a prior distribution is called a *conjugate* prior distribution. For MCMC methods, use of conjugate priors means that a closed-form solution for the conditional posterior distributions is available. Random draws of the Gibbs sampler can then be obtained by using the commonly available computer routines of probability distributions. In what follows, we review some well-known conjugate priors. For more information, readers are referred to textbooks on Bayesian statistics (e.g., DeGroot 1970, Chapter 9).

Result 12.1. Suppose that $x_1, \ldots, x_n$ form a random sample from a normal distribution with mean μ, which is unknown, and variance σ^2, which is known

and positive. Suppose that the prior distribution of μ is a normal distribution with mean μ_o and variance σ_o^2. Then the posterior distribution of μ given the data and prior is normal with mean μ_* and variance σ_*^2 given by

$$\mu_* = \frac{\sigma^2\mu_o + n\sigma_o^2\bar{x}}{\sigma^2 + n\sigma_o^2} \qquad \text{and} \qquad \sigma_*^2 = \frac{\sigma^2\sigma_o^2}{\sigma^2 + n\sigma_o^2},$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ is the sample mean.

In Bayesian analysis, it is often convenient to use the *precision* parameter $\eta = 1/\sigma^2$ (i.e., the inverse of the variance σ^2). Denote the precision parameter of the prior distribution by $\eta_o = 1/\sigma_o^2$ and that of the posterior distribution by $\eta_* = 1/\sigma_*^2$. Then Result 12.1 can be rewritten as

$$\eta_* = \eta_o + n\eta \qquad \text{and} \qquad \mu_* = \frac{\eta_o}{\eta_*} \times \mu_o + \frac{n\eta}{\eta_*} \times \bar{x}.$$

For the normal random sample considered, data information about μ is contained in the sample mean $\bar{x}$, which is the sufficient statistic of μ. The precision of $\bar{x}$ is $n/\sigma^2 = n\eta$. Consequently, Result 12.1 says that (a) precision of the posterior distribution is the sum of the precisions of the prior and the data, and (b) the posterior mean is a weighted average of the prior mean and sample mean with weight proportional to the precision. The two formulas also show that the contribution of the prior distribution is diminishing as the sample size n increases.

A multivariate version of Result 12.1 is particularly useful in MCMC methods when linear regression models are involved; see Box and Tiao (1973).

Result 12.1a. Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ form a random sample from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and a known covariance matrix $\boldsymbol{\Sigma}$. Suppose also that the prior distribution of $\boldsymbol{\mu}$ is multivariate normal with mean vector $\boldsymbol{\mu}_o$ and covariance matrix $\boldsymbol{\Sigma}_o$. Then the posterior distribution of $\boldsymbol{\mu}$ is also multivariate normal with mean vector $\boldsymbol{\mu}_*$ and covariance matrix $\boldsymbol{\Sigma}_*$, where

$$\boldsymbol{\Sigma}_*^{-1} = \boldsymbol{\Sigma}_o^{-1} + n\boldsymbol{\Sigma}^{-1} \qquad \text{and} \qquad \boldsymbol{\mu}_* = \boldsymbol{\Sigma}_*(\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}),$$

where $\bar{\boldsymbol{x}} = \sum_{i=1}^n \boldsymbol{x}_i/n$ is the sample mean, which is distributed as a multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}/n$. Note that $n\boldsymbol{\Sigma}^{-1}$ is the precision matrix of $\bar{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_o^{-1}$ is the precision matrix of the prior distribution.

A random variable η has a gamma distribution with positive parameters α and β if its probability density function is

$$f(\eta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\eta^{\alpha-1}e^{-\beta\eta}, \qquad \eta > 0,$$

where $\Gamma(\alpha)$ is a gamma function. For this distribution, $E(\eta) = \alpha/\beta$ and $\text{Var}(\eta) = \alpha/\beta^2$.

Result 12.2. Suppose that $x_1, \ldots, x_n$ form a random sample from a normal distribution with a given mean μ and an unknown precision η. If the prior distribution of η is a gamma distribution with positive parameters α and β, then the posterior distribution of η is a gamma distribution with parameters $\alpha + (n/2)$ and $\beta + \sum_{i=1}^n (x_i - \mu)^2/2$.

A random variable θ has a beta distribution with positive parameters α and β if its probability density function is

$$f(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad 0 < \theta < 1.$$

The mean and variance of θ are $E(\theta) = \alpha/(\alpha+\beta)$ and $\text{Var}(\theta) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$.

Result 12.3. Suppose that $x_1, \ldots, x_n$ form a random sample from a Bernoulli distribution with parameter θ. If the prior distribution of θ is a beta distribution with given positive parameters α and β, then the posterior of θ is a beta distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$.

Result 12.4. Suppose that $x_1, \ldots, x_n$ form a random sample from a Poisson distribution with parameter λ. Suppose also that the prior distribution of λ is a gamma distribution with given positive parameters α and β. Then the posterior distribution of λ is a gamma distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n$.

Result 12.5. Suppose that $x_1, \ldots, x_n$ form a random sample from an exponential distribution with parameter λ. If the prior distribution of λ is a gamma distribution with given positive parameters α and β, then the posterior distribution of λ is a gamma distribution with parameters $\alpha + n$ and $\beta + \sum_{i=1}^n x_i$.

A random variable X has a negative binomial distribution with parameters m and λ, where $m > 0$ and $0 < \lambda < 1$, if X has a probability mass function

$$p(n|m, \lambda) = \begin{cases} \binom{m+n-1}{n}\lambda^m(1-\lambda)^n & \text{if } n = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

A simple example of negative binomial distribution in finance is how many MBA graduates a firm must interview before finding exactly m "right candidates" for its m openings, assuming that the applicants are independent and each applicant has a probability λ of being a perfect fit. Denote the total number of interviews by Y. Then $X = Y - m$ is distributed as a negative binomial with parameters m and λ.

Result 12.6. Suppose that $x_1, \ldots, x_n$ form a random sample from a negative binomial distribution with parameters m and λ, where m is positive and fixed. If

the prior distribution of λ is a beta distribution with positive parameters α and β, then the posterior distribution of λ is a beta distribution with parameters $\alpha + mn$ and $\beta + \sum_{i=1}^{n} x_i$.

Next we consider the case of a normal distribution with an unknown mean μ and an unknown precision η. The two-dimensional prior distribution is partitioned as $P(\mu, \eta) = P(\mu|\eta)P(\eta)$.

Result 12.7. Suppose that $x_1, \ldots, x_n$ form a random sample from a normal distribution with an unknown mean μ and an unknown precision η. Suppose also that the conditional distribution of μ given $\eta = \eta_o$ is a normal distribution with mean μ_o and precision $\tau_o\eta_o$ and the marginal distribution of η is a gamma distribution with positive parameters α and β. Then the conditional posterior distribution of μ given $\eta = \eta_o$ is a normal distribution with mean μ_* and precision η_*,

$$\mu_* = \frac{\tau_o\mu_o + n\bar{x}}{\tau_o + n} \qquad \text{and} \qquad \eta_* = (\tau_o + n)\eta_o,$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ is the sample mean, and the marginal posterior distribution of η is a gamma distribution with parameters $\alpha + (n/2)$ and β_*, where

$$\beta_* = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\tau_o n(\bar{x} - \mu_o)^2}{2(\tau_o + n)}.$$

When the conditional variance of a random variable is of interest, an inverted chi-squared distribution (or inverse chi-squared) is often used. A random variable Y has an inverted chi-squared distribution with v degrees of freedom if $1/Y$ follows a chi-squared distribution with the same degrees of freedom. The probability density function of Y is

$$f(y|v) = \frac{2^{-v/2}}{\Gamma(v/2)} y^{-(v/2+1)} e^{-1/(2y)}, \qquad y > 0.$$

For this distribution, we have $E(Y) = 1/(v-2)$ if $v > 2$ and $\text{Var}(Y) = 2/[(v-2)^2(v-4)]$ if $v > 4$.

Result 12.8. Suppose that $a_1, \ldots, a_n$ form a random sample from a normal distribution with mean zero and variance σ^2. Suppose also that the prior distribution of σ^2 is an inverted chi-squared distribution with v degrees of freedom [i.e., $(v\lambda)/\sigma^2 \sim \chi_v^2$, where $\lambda > 0$]. Then the posterior distribution of σ^2 is also an inverted chi-squared distribution with $v + n$ degrees of freedom—that is, $(v\lambda + \sum_{i=1}^{n} a_i^2)/\sigma^2 \sim \chi_{v+n}^2$.

12.4 ALTERNATIVE ALGORITHMS

In many applications, there are no closed-form solutions for the conditional posterior distributions. But many clever alternative algorithms have been devised in the statistical literature to overcome this difficulty. In this section, we discuss some of these algorithms.

12.4.1 Metropolis Algorithm

This algorithm is applicable when the conditional posterior distribution is known except for a normalization constant; see Metropolis and Ulam (1949) and Metropolis et al. (1953). Suppose that we want to draw a random sample from the distribution $f(\boldsymbol{\theta}|\boldsymbol{X})$, which contains a complicated normalization constant so that a direct draw is either too time-consuming or infeasible. But there exists an approximate distribution for which random draws are easily available. The Metropolis algorithm generates a sequence of random draws from the approximate distribution whose distributions converge to $f(\boldsymbol{\theta}|\boldsymbol{X})$. The algorithm proceeds as follows:

1. Draw a random starting value $\boldsymbol{\theta}_0$ such that $f(\boldsymbol{\theta}_0|\boldsymbol{X}) > 0$.
2. For $t = 1, 2, \ldots,$
   a. Draw a candidate sample $\boldsymbol{\theta}_*$ from a *known* distribution at iteration t given the previous draw $\boldsymbol{\theta}_{t-1}$. Denote the known distribution by $J_t(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$, which is called a *jumping distribution* in Gelman et al. (2003). It is also referred to as a *proposal distribution*. The jumping distribution must be symmetric—that is, $J_t(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j) = J_t(\boldsymbol{\theta}_j|\boldsymbol{\theta}_i)$ for all $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_j$, and t.
   b. Calculate the ratio
   $$r = \frac{f(\boldsymbol{\theta}_*|\boldsymbol{X})}{f(\boldsymbol{\theta}_{t-1}|\boldsymbol{X})}.$$
   c. Set
   $$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_* & \text{with probability } \min(r, 1), \\ \boldsymbol{\theta}_{t-1} & \text{otherwise.} \end{cases}$$

Under some regularity conditions, the sequence $\{\boldsymbol{\theta}_t\}$ converges in distribution to $f(\boldsymbol{\theta}|\boldsymbol{X})$; see Gelman et al. (2003).

Implementation of the algorithm requires the ability to calculate the ratio r for all $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}_{t-1}$, to draw $\boldsymbol{\theta}_*$ from the jumping distribution, and to draw a random realization from a uniform distribution to determine the acceptance or rejection of $\boldsymbol{\theta}_*$. The normalization constant of $f(\boldsymbol{\theta}|\boldsymbol{X})$ is not needed because only a ratio is used.

The acceptance and rejection rule of the algorithm can be stated as follows: (i) if the jump from $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_*$ increases the conditional posterior density, then accept $\boldsymbol{\theta}_*$ as $\boldsymbol{\theta}_t$; (ii) if the jump decreases the posterior density, then set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_*$

with probability equal to the density ratio r, and set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ otherwise. Such a procedure seems reasonable.

Examples of symmetric jumping distributions include the normal and Student-t distributions for the mean parameter. For a given covariance matrix, we have $f(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j) = f(\boldsymbol{\theta}_j|\boldsymbol{\theta}_i)$, where $f(\boldsymbol{\theta}|\boldsymbol{\theta}_o)$ denotes a multivariate normal density function with mean vector $\boldsymbol{\theta}_o$.

12.4.2 Metropolis–Hasting Algorithm

Hasting (1970) generalizes the Metropolis algorithm in two ways. First, the jumping distribution does not have to be symmetric. Second, the jumping rule is modified to

$$r = \frac{f(\boldsymbol{\theta}_*|\boldsymbol{X})/J_t(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1})}{f(\boldsymbol{\theta}_{t-1}|\boldsymbol{X})/J_t(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_*)} = \frac{f(\boldsymbol{\theta}_*|\boldsymbol{X})J_t(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_*)}{f(\boldsymbol{\theta}_{t-1}|\boldsymbol{X})J_t(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1})}.$$

This modified algorithm is referred to as the Metropolis–Hasting algorithm. Tierney (1994) discusses methods to improve computational efficiency of the algorithm.

12.4.3 Griddy Gibbs

In financial applications, an entertained model may contain some nonlinear parameters (e.g., the moving-average parameters in an ARMA model or the GARCH parameters in a volatility model). Since conditional posterior distributions of nonlinear parameters do not have a closed-form expression, implementing a Gibbs sampler in this situation may become complicated even with the Metropolis–Hasting algorithm. Tanner (1996) describes a simple procedure to obtain random draws in a Gibbs sampling when the conditional posterior distribution is univariate. The method is called the *Griddy Gibbs sampler* and is widely applicable. However, the method could be inefficient in a real application.

Let θ_i be a scalar parameter with conditional posterior distribution $f(\theta_i|\boldsymbol{X}, \boldsymbol{\theta}_{-i})$, where $\boldsymbol{\theta}_{-i}$ is the parameter vector after removing θ_i. For instance, if $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$, then $\boldsymbol{\theta}_{-1} = (\theta_2, \theta_3)'$. The Griddy Gibbs proceeds as follows:

1. Select a grid of points from a properly selected interval of θ_i, say, $\theta_{i1} \leq \theta_{i2} \leq \cdots \leq \theta_{im}$. Evaluate the conditional posterior density function to obtain $w_j = f(\theta_{ij}|\boldsymbol{X}, \boldsymbol{\theta}_{-i})$ for $j = 1, \ldots, m$.
2. Use $w_1, \ldots, w_m$ to obtain an approximation to the inverse cumulative distribution function (CDF) of $f(\theta_i|\boldsymbol{X}, \boldsymbol{\theta}_{-i})$.
3. Draw a uniform (0,1) random variate and transform the observation via the approximate inverse CDF to obtain a random draw for θ_i.

Some remarks on the Griddy Gibbs are in order. First, the normalization constant of the conditional posterior distribution $f(\theta_i|\boldsymbol{X}, \boldsymbol{\theta}_{-i})$ is not needed because the inverse CDF can be obtained from $\{w_j\}_{j=1}^m$ directly. Second, a simple approximation to the inverse CDF is a discrete distribution for $\{\theta_{ij}\}_{j=1}^m$ with probability $p(\theta_{ij}) = w_j / \sum_{v=1}^m w_v$. Third, in a real application, selection of the interval

$[\theta_{i1}, \theta_{im}]$ for the parameter θ_i must be checked carefully. A simple checking procedure is to consider the histogram of the Gibbs draws of θ_i. If the histogram indicates substantial probability around θ_{i1} or θ_{im}, then the interval must be expanded. However, if the histogram shows a concentration of probability inside the interval $[\theta_{i1}, \theta_{im}]$, then the interval is too wide and can be shortened. If the interval is too wide, then the Griddy Gibbs becomes inefficient because most of w_j would be zero. Finally, the Griddy Gibbs or Metropolis–Hasting algorithm can be used in a Gibbs sampling to obtain random draws of some parameters.

12.5 LINEAR REGRESSION WITH TIME SERIES ERRORS

We are ready to consider some specific applications of MCMC methods. Examples discussed in the next few sections are for illustrative purposes only. The goal here is to highlight the applicability and usefulness of the methods. Understanding these examples can help readers gain insights into applications of MCMC methods in finance.

The first example is to estimate a regression model with serially correlated errors. This is a topic discussed in Chapter 2, where we use SCA to perform the estimation. A simple version of the model is

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + z_t, \\ z_t &= \phi z_{t-1} + a_t, \end{aligned}$$

where y_t is the dependent variable, x_{it} are explanatory variables that may contain lagged values of y_t, and z_t follows a simple AR(1) model with $\{a_t\}$ being a sequence of independent and identically distributed normal random variables with mean zero and variance σ^2. Denote the parameters of the model by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi, \sigma^2)'$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, and let $\boldsymbol{x}_t = (1, x_{1t}, \ldots, x_{kt})'$ be the vector of all regressors at time t, including a constant of unity. The model becomes

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + z_t, \qquad z_t = \phi z_{t-1} + a_t, \qquad t = 1, \ldots, n, \tag{12.6}$$

where n is the sample size.

A natural way to implement Gibbs sampling in this case is to iterate between regression estimation and time series estimation. If the time series model is known, we can estimate the regression model easily by using the least-squares method. However, if the regression model is known, we can obtain the time series z_t by using $z_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$ and use the series to estimate the AR(1) model. Therefore, we need the following conditional posterior distributions:

$$f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma^2), \qquad f(\phi|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2), \qquad f(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \phi),$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{X}$ denotes the collection of all observations of explanatory variables.

We use conjugate prior distributions to obtain closed-form expressions for the conditional posterior distributions. The prior distributions are

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o), \qquad \phi \sim N(\phi_o, \sigma_o^2), \qquad \frac{\upsilon\lambda}{\sigma^2} \sim \chi_\upsilon^2, \tag{12.7}$$

where again $\sim$ denotes distribution, and $\boldsymbol{\beta}_o$, $\boldsymbol{\Sigma}_o$, λ, υ, ϕ_o, and σ_o^2 are known quantities. These quantities are referred to as hyperparameters in Bayesian inference. Their exact values depend on the problem at hand. Typically, we assume that $\boldsymbol{\beta}_o = \mathbf{0}$, $\phi_o = 0$, and $\boldsymbol{\Sigma}_o$ is a diagonal matrix with large diagonal elements. The prior distributions in Eq. (12.7) are assumed to be independent of each other. Thus, we use independent priors based on the partition of the parameter vector $\boldsymbol{\theta}$.

The conditional posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma^2)$ can be obtained by using Result 12.1a of Section 12.3. Specifically, given ϕ, we define

$$y_{o,t} = y_t - \phi y_{t-1}, \qquad \boldsymbol{x}_{o,t} = \boldsymbol{x}_t - \phi \boldsymbol{x}_{t-1}.$$

Using Eq. (12.6), we have

$$y_{o,t} = \boldsymbol{\beta}'\boldsymbol{x}_{o,t} + a_t, \qquad t = 2, \ldots, n. \tag{12.8}$$

Under the assumption of $\{a_t\}$, Eq. (12.8) is a multiple linear regression. Therefore, information of the data about the parameter vector $\boldsymbol{\beta}$ is contained in its least-squares estimate

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=2}^{n} \boldsymbol{x}_{o,t}\boldsymbol{x}_{o,t}'\right)^{-1}\left(\sum_{t=2}^{n} \boldsymbol{x}_{o,t} y_{o,t}\right),$$

which has a multivariate normal distribution

$$\widehat{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, \quad \sigma^2\left(\sum_{t=2}^{n} \boldsymbol{x}_{o,t}\boldsymbol{x}_{o,t}'\right)^{-1}\right].$$

Using Result 12.1a, the posterior distribution of $\boldsymbol{\beta}$, given the data, ϕ, and σ^2, is multivariate normal. We write the result as

$$(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma) \sim N(\boldsymbol{\beta}_*, \boldsymbol{\Sigma}_*), \tag{12.9}$$

where the parameters are given by

$$\boldsymbol{\Sigma}_*^{-1} = \frac{\sum_{t=2}^{n} \boldsymbol{x}_{o,t}\boldsymbol{x}_{o,t}'}{\sigma^2} + \boldsymbol{\Sigma}_o^{-1}, \qquad \boldsymbol{\beta}_* = \boldsymbol{\Sigma}_*\left(\frac{\sum_{t=2}^{n} \boldsymbol{x}_{o,t}\boldsymbol{x}_{o,t}'}{\sigma^2}\widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_o^{-1}\boldsymbol{\beta}_o\right).$$

Next, consider the conditional posterior distribution of ϕ given $\boldsymbol{\beta}$, σ^2, and the data. Because $\boldsymbol{\beta}$ is given, we can calculate $z_t = y_t - \boldsymbol{\beta}'\boldsymbol{x}_t$ for all t and consider the AR(1) model

$$z_t = \phi z_{t-1} + a_t, \qquad t = 2, \ldots, n.$$

The information of the likelihood function about ϕ is contained in the least-squares estimate

$$\widehat{\phi} = \left(\sum_{t=2}^{n} z_{t-1}^2\right)^{-1}\left(\sum_{t=2}^{n} z_{t-1} z_t\right),$$

which is normally distributed with mean ϕ and variance $\sigma^2(\sum_{t=2}^{n} z_{t-1}^2)^{-1}$. Based on Result 12.1, the posterior distribution of ϕ is also normal with mean ϕ_* and variance σ_*^2, where

$$\sigma_*^{-2} = \frac{\sum_{t=2}^{n} z_{t-1}^2}{\sigma^2} + \sigma_o^{-2}, \qquad \phi_* = \sigma_*^2\left(\frac{\sum_{t=2}^{n} z_{t-1}^2}{\sigma^2}\widehat{\phi} + \sigma_o^{-2}\phi_o\right). \tag{12.10}$$

Finally, turn to the posterior distribution of σ^2 given $\boldsymbol{\beta}$, ϕ, and the data. Because $\boldsymbol{\beta}$ and ϕ are known, we can calculate

$$a_t = z_t - \phi z_{t-1}, \qquad z_t = y_t - \boldsymbol{\beta}'\boldsymbol{x}_t, \qquad t = 2, \ldots, n.$$

By Result 12.8, the posterior distribution of σ^2 is an inverted chi-squared distribution—that is,

$$\frac{\upsilon\lambda + \sum_{t=2}^{n} a_t^2}{\sigma^2} \sim \chi_{\upsilon+(n-1)}^2, \tag{12.11}$$

where χ_k^2 denotes a chi-squared distribution with k degrees of freedom.

Using the three conditional posterior distributions in Eqs. (12.9)–(12.11), we can estimate Eq. (12.6) via Gibbs sampling as follows:

1. Specify the hyperparameter values of the priors in Eq. (12.7).
2. Specify arbitrary starting values for $\boldsymbol{\beta}$, ϕ, and σ^2 (e.g., the ordinary least-squares estimate of $\boldsymbol{\beta}$ without time series errors).
3. Use the multivariate normal distribution in Eq. (12.9) to draw a random realization for $\boldsymbol{\beta}$.
4. Use the univariate normal distribution in Eq. (12.10) to draw a random realization for ϕ.
5. Use the chi-squared distribution in Eq. (12.11) to draw a random realization for σ^2.

Repeat steps 3–5 for many iterations to obtain a Gibbs sample. The sample means are then used as point estimates of the parameters of model (12.6).

Example 12.1. As an illustration, we revisit the example of U.S. weekly interest rates of Chapter 2. The data are the 1-year and 3-year Treasury constant maturity rates from January 5, 1962, to April 10, 2009, and are obtained from the Federal Reserve Bank of St. Louis. Because of unit-root nonstationarity, the dependent and independent variables are

1. $c_{3t} = r_{3t} - r_{3,t-1}$, which is the weekly change in 3-year maturity rate,
2. $c_{1t} = r_{1t} - r_{1,t-1}$, which is the weekly change in 1-year maturity rate,

where the original interest rates r_{it} are measured in percentages. In Chapter 2, we employed a linear regression model with an MA(1) error for the data. Here we consider an AR(2) model for the error process. Using the traditional approach in R, we obtain the model

$$c_{3t} = 0.782c_{1t} + z_t, \qquad z_t = 0.183z_{t-1} - 0.036z_{t-2} + a_t, \tag{12.12}$$

where $\widehat{\sigma}_a = 0.068$. Standard errors of the coefficient estimates of Eq. (12.12) are 0.0075, 0.0201, and 0.0201, respectively. Except for a marginally significant residual ACF at lags 4 and 6, the prior model seems adequate.

Writing the model as

$$c_{3t} = \beta c_{1t} + z_t, \qquad z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t, \tag{12.13}$$

where $\{a_t\}$ is an independent sequence of $N(0, \sigma^2)$ random variables, we estimate the parameters by Gibbs sampling. The prior distributions used are

$$\boldsymbol{\beta} \sim N(0, 4), \quad \boldsymbol{\phi} \sim N[\mathbf{0}, \text{diag}(0.25, 0.16)], \quad (v\lambda)/\sigma^2 = (10 \times 0.05)/\sigma^2 \sim \chi^2_{10}.$$

The initial parameter estimates are obtained by the ordinary least-squares method [i.e., by using a two-step procedure of fitting the linear regression model first, then fitting an AR(2) model to the regression residuals]. Since the sample size 2466 is large, the initial estimates are close to those given in Eq. (12.12). We iterated the Gibbs sampling for 2100 iterations but discard results of the first 100 iterations. Table 12.1 gives the posterior means and standard errors of the parameters. From the table, the posterior mean of σ is approximately 0.069. Figure 12.1 shows the time plots of the 2000 Gibbs draws of the parameters. The plots show that the draws are stable. Figure 12.2 gives the histogram of the marginal posterior distribution of each parameter.

We repeated the Gibbs sampling with different initial values but obtained similar results. The Gibbs sampling appears to have converged. From Table 12.1, the posterior means are close to the estimates of Eq. (12.12).This is expected as the sample size is large and the model is relatively simple.

TABLE 12.1 Posterior Means and Standard Errors of Model (12.13) Estimated by Gibbs Sampling with 2100 Iterations[a]

Parameter	β	ϕ_1	ϕ_2	σ^2
Mean	0.793	0.184	−0.036	0.00479
Standard error	0.008	0.019	0.021	0.00013

[a]The results are based on the last 2000 iterations, and the prior distributions are given in the text.

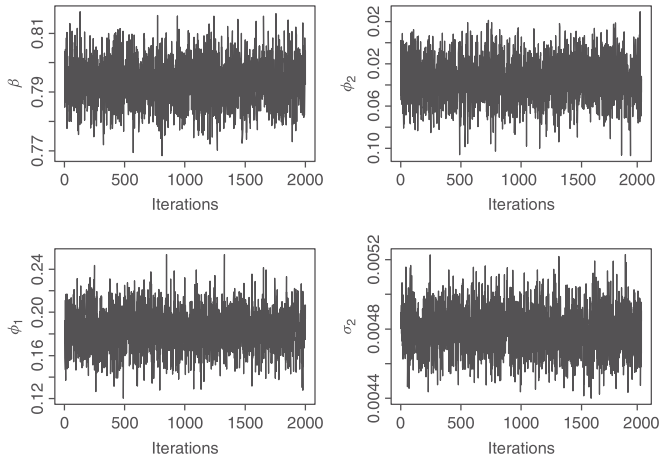

Figure 12.1 Time plots of Gibbs draws for the model in Eq. (12.13) with 2100 iterations. Results are based on last 2000 draws. Prior distributions and starting parameter values are given in text.

12.6 MISSING VALUES AND OUTLIERS

In this section, we discuss MCMC methods for handling missing values and detecting additive outliers. Let $\{y_t\}_{t=1}^n$ be an observed time series. A data point y_h is an additive outlier if

$$y_t = \begin{cases} x_h + \omega & \text{if } t = h, \\ x_t & \text{otherwise,} \end{cases} \tag{12.14}$$

where ω is the magnitude of the outlier and x_t is an outlier-free time series. Examples of additive outliers include recording errors (e.g., typos and measurement errors). Outliers can seriously affect time series analysis because they may induce substantial biases in parameter estimation and lead to model misspecification.

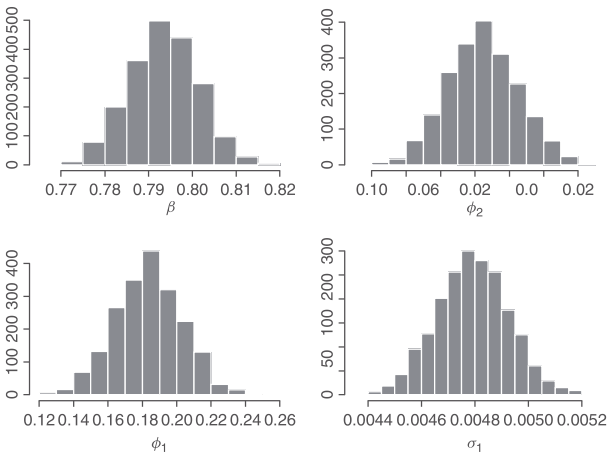

Figure 12.2 Histograms of Gibbs draws for model in Eq. (12.13) with 2100 iterations. Results are based on last 2000 draws. Prior distributions and starting parameter values are given in text.

Consider a time series x_t and a fixed time index h. We can learn a lot about x_h by treating it as a missing value. If the model of x_t were known, then we could derive the conditional distribution of x_h given the other values of the series. By comparing the observed value y_h with the derived distribution of x_h, we can determine whether y_h can be classified as an additive outlier. Specifically, if y_h is a value that is likely to occur under the derived distribution, then y_h is not an additive outlier. However, if the chance to observe y_h is very small under the derived distribution, then y_h can be classified as an additive outlier. Therefore, detection of additive outliers and treatment of missing values in time series analysis are based on the same idea.

In the literature, missing values in a time series can be handled by using either the Kalman filter or MCMC methods; see Jones (1980), Chapter 11, and McCulloch and Tsay (1994a). Outlier detection has also been carefully investigated; see Chang, Tiao, and Chen (1988), Tsay (1988), Tsay, Peña, and Pankratz (2000), and the references therein. The outliers are classified into four categories depending on the nature of their impacts on the time series. Here we focus on additive outliers.

12.6.1 Missing Values

For ease in presentation, consider an AR(p) time series

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t, \tag{12.15}$$

where $\{a_t\}$ is a Gaussian white noise series with mean zero and variance σ^2. Suppose that the sampling period is from $t = 1$ to $t = n$, but the observation x_h is missing, where $1 < h < n$. Our goal is to estimate the model in the presence of a missing value.

In this particular instance, the parameters are $\boldsymbol{\theta} = (\boldsymbol{\phi}', x_h, \sigma^2)'$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. Thus, we treat the missing value x_h as an unknown parameter. If we assume that the prior distributions are

$$\boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{\Sigma}_o), \qquad x_h \sim N(\mu_o, \sigma_o^2), \qquad \frac{v\lambda}{\sigma^2} \sim \chi_v^2,$$

where the hyperparameters are known, then the conditional posterior distributions $f(\boldsymbol{\phi}|\boldsymbol{X}, x_h, \sigma^2)$ and $f(\sigma^2|\boldsymbol{X}, x_h, \boldsymbol{\phi})$ are exactly as those given in the previous section, where $\boldsymbol{X}$ denotes the observed data. The conditional posterior distribution $f(x_h|\boldsymbol{X}, \boldsymbol{\phi}, \sigma^2)$ is univariate normal with mean μ_* and variance σ_h^2. These two parameters can be obtained by using a linear regression model. Specifically, given the model and the data, x_h is only related to $\{x_{h-p}, \ldots, x_{h-1}, x_{h+1}, \ldots, x_{h+p}\}$. Keeping in mind that x_h is an unknown parameter, we can write the relationship as follows:

1. For $t = h$, the model says

   $$x_h = \phi_1 x_{h-1} + \cdots + \phi_p x_{h-p} + a_h.$$

   Letting $y_h = \phi_1 x_{h-1} + \cdots + \phi_p x_{h-p}$ and $b_h = -a_h$, the prior equation can be written as

   $$y_h = x_h + b_h = \phi_0 x_h + b_h,$$

   where $\phi_0 = 1$.
2. For $t = h + 1$, we have

   $$x_{h+1} = \phi_1 x_h + \phi_2 x_{h-1} + \cdots + \phi_p x_{h+1-p} + a_{h+1}.$$

   Letting $y_{h+1} = x_{h+1} - \phi_2 x_{h-1} - \cdots - \phi_p x_{h+1-p}$ and $b_{h+1} = a_{h+1}$, the prior equation can be written as

   $$y_{h+1} = \phi_1 x_h + b_{h+1}.$$

3. In general, for $t = h + j$ with $j = 1, \ldots, p$, we have

   $$x_{h+j} = \phi_1 x_{h+j-1} + \cdots + \phi_j x_h + \phi_{j+1} x_{h-1} + \cdots + \phi_p x_{h+j-p} + a_{h+j}.$$

   Let $y_{h+j} = x_{h+j} - \phi_1 x_{h+j-1} - \cdots - \phi_{j-1} x_{h+1} - \phi_{j+1} x_{h-1} - \cdots - \phi_p x_{h+j-p}$ and $b_{h+j} = a_{h+j}$. The prior equation reduces to

   $$y_{h+j} = \phi_j x_h + b_{h+j}.$$

Consequently, for an AR(p) model, the missing value x_h is related to the model, and the data in $p+1$ equations

$$y_{h+j} = \phi_j x_h + b_{h+j}, \qquad j = 0, \ldots, p, \tag{12.16}$$

where $\phi_0 = 1$. Since a normal distribution is symmetric with respect to its mean, a_h and $-a_h$ have the same distribution. Consequently, Eq. (12.16) is a special simple linear regression model with $p+1$ data points. The least-squares estimate of x_h and its variance are

$$\widehat{x}_h = \frac{\sum_{j=0}^{p} \phi_j y_{h+j}}{\sum_{j=0}^{p} \phi_j^2}, \qquad \text{Var}(\widehat{x}_h) = \frac{\sigma^2}{\sum_{j=0}^{p} \phi_j^2}.$$

For instance, when $p = 1$, we have $\widehat{x}_h = [\phi_1/(1+\phi_1^2)](x_{h-1} + x_{h+1})$, which is referred to as the filtered value of x_h. Because a Gaussian AR(1) model is time reversible, equal weights are applied to the two neighboring observations of x_h to obtain the filtered value.

Finally, using Result 12.1, we obtain that the posterior distribution of x_h is normal with mean μ_* and variance σ_*^2, where

$$\mu_* = \frac{\sigma^2 \mu_o + \sigma_o^2 (\sum_{j=0}^{p} \phi_j^2) \widehat{x}_h}{\sigma^2 + \sigma_o^2 (\sum_{j=0}^{p} \phi_j^2)}, \qquad \sigma_*^2 = \frac{\sigma^2 \sigma_o^2}{\sigma^2 + \sigma_o^2 \sum_{j=0}^{p} \phi_j^2}. \tag{12.17}$$

Missing values may occur in patches, resulting in the situation of multiple consecutive missing values. These missing values can be handled in two ways. First, we can generalize the prior method directly to obtain a solution for multiple filtered values. Consider, for instance, the case that x_h and x_{h+1} are missing. These missing values are related to $\{x_{h-p}, \ldots, x_{h-1}; x_{h+2}, \ldots, x_{h+p+1}\}$. We can define a dependent variable y_{h+j} in a similar manner as before to set up a multiple linear regression with parameters x_h and x_{h+1}. The least-squares method is then used to obtain estimates of x_h and x_{h+1}. Combining with the specified prior distributions, we have a bivariate normal posterior distribution for $(x_h, x_{h+1})'$. In Gibbs sampling, this approach draws the consecutive missing values jointly. Second, we can apply the result of a single missing value in Eq. (12.17) multiple times within a Gibbs iteration. Again consider the case of missing x_h and x_{h+1}. We can employ the conditional posterior distributions $f(x_h|\boldsymbol{X}, x_{h+1}, \boldsymbol{\phi}, \sigma^2)$ and $f(x_{h+1}|\boldsymbol{X}, x_h, \boldsymbol{\phi}, \sigma^2)$ separately. In Gibbs sampling, this means that we draw the missing value one at a time.

Because x_h and x_{h+1} are correlated in a time series, drawing them jointly is preferred in a Gibbs sampling. This is particularly so if the number of consecutive missing values is large. Drawing one missing value at a time works well if the number of missing values is small.

Remark. In the previous discussion, we assumed $h - p \geq 1$ and $h + p \leq n$. If h is close to the end points of the sample period, the number of data points available in the linear regression model must be adjusted. □