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

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

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 $\boldsymbol{\Theta}$. 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 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. 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 $\boldsymbol{\Theta}$, 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:

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

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

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

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,

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

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.

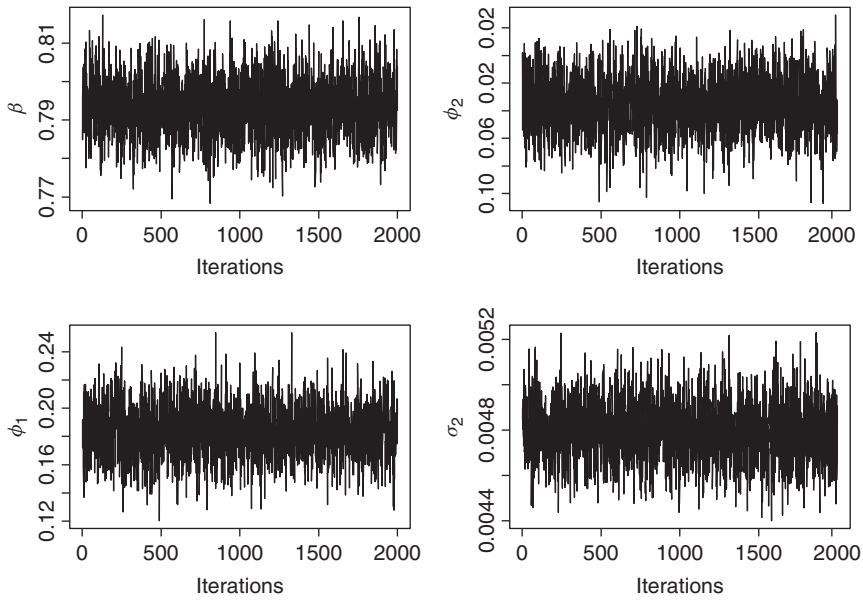

Figure 12.1 Time plots of Gibbs draws for the model in Eq. (12.13) with 2100 iterations. Results are based on last 2000 draws. Prior distributions and starting parameter values are given in text.

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.

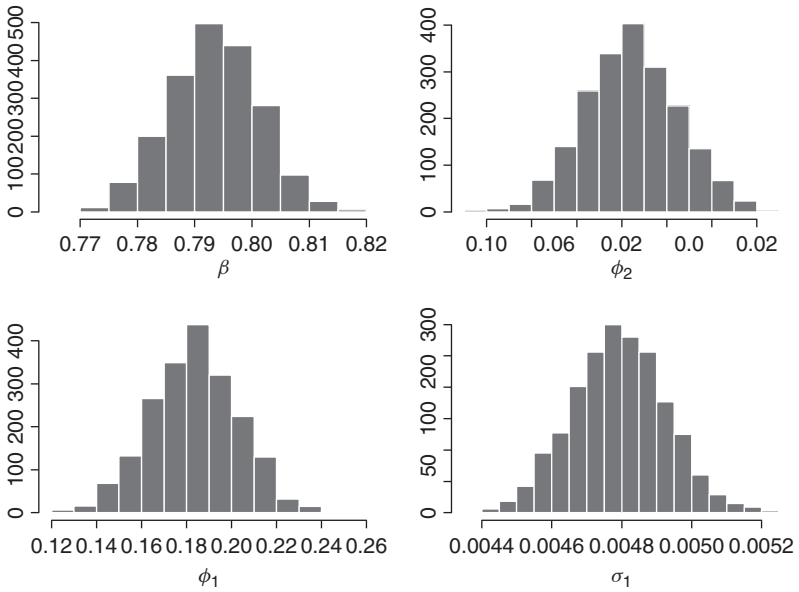

Figure 12.2 Histograms of Gibbs draws for model in Eq. (12.13) with 2100 iterations. Results are based on last 2000 draws. Prior distributions and starting parameter values are given in text.

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. □

12.6.2 Outlier Detection

Detection of additive outliers in Eq. (12.14) becomes straightforward under the MCMC framework. Except for the case of a patch of additive outliers with similar magnitudes, the simple Gibbs sampler of McCulloch and Tsay (1994a) seems to work well; see Justel, Peña, and Tsay (2001). Again we use an AR model to illustrate the problem. The method applies equally well to other time series models when the Metropolis–Hasting algorithm or the Griddy Gibbs is used to draw values of nonlinear parameters.

Assume that the observed time series is y_t, which may contain some additive outliers whose locations and magnitudes are unknown. We write the model for y_t as

$$y_t = \delta_t \beta_t + x_t, \qquad t = 1, \ldots, n, \tag{12.18}$$

where $\{\delta_t\}$ is a sequence of independent Bernoulli random variables such that $P(\delta_t = 1) = \epsilon$ and $P(\delta_t = 0) = 1 - \epsilon$, ϵ is a constant between 0 and 1, $\{\beta_t\}$ is a sequence of independent random variables from a given distribution, and x_t is an outlier-free AR(p) time series,

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t,$$

where $\{a_t\}$ is a Gaussian white noise with mean zero and variance σ^2. This model seems complicated, but it allows additive outliers to occur at every time point. The chance of being an outlier for each observation is ϵ.

Under the model in Eq. (12.18), we have n data points, but there are $2n + p + 3$ parameters—namely, $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_p)'$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$, σ^2, and ϵ. The binary parameters δ_t are governed by ϵ and the β_t are determined by the specified distribution. The parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are introduced by using the idea of data augmentation with δ_t denoting the presence or absence of an additive outlier at time t, and β_t is the magnitude of the outlier at time t when it is present.

Assume that the prior distributions are

$$\boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{\Sigma}_o), \qquad \frac{v\lambda}{\sigma^2} \sim \chi^2_v, \qquad \epsilon \sim \text{Beta}(\gamma_1, \gamma_2), \qquad \beta_t \sim N(0, \xi^2),$$

where the hyperparameters are known. These are conjugate prior distributions. To implement Gibbs sampling for model estimation with outlier detection, we need to consider the conditional posterior distributions of

$$f(\boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2), \qquad f(\delta_h|\boldsymbol{Y}, \boldsymbol{\delta}_{-h}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2), \qquad f(\beta_h|\boldsymbol{Y}, \boldsymbol{\delta}, \boldsymbol{\beta}_{-h}, \boldsymbol{\phi}, \sigma^2),$$
$$f(\epsilon|\boldsymbol{Y}, \boldsymbol{\delta}), \qquad f(\sigma^2|\boldsymbol{Y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}),$$

where $1 \le h \le n$, $\boldsymbol{Y}$ denotes the data, and $\boldsymbol{\theta}_{-i}$ denotes that the ith element of $\boldsymbol{\theta}$ is removed.

Conditioned on $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, the outlier-free time series x_t can be obtained by $x_t = y_t - \delta_t \beta_t$. Information of the data about $\boldsymbol{\phi}$ is then contained in the least-squares estimate

$$\widehat{\boldsymbol{\phi}} = \left(\sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}' \right)^{-1} \left(\sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} x_t \right),$$

where $\boldsymbol{x}_{t-1} = (1, x_{t-1}, \ldots, x_{t-p})'$, which is normally distributed with mean $\boldsymbol{\phi}$ and covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \sigma^2 \left(\sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}' \right)^{-1}.$$

The conditional posterior distribution of $\boldsymbol{\phi}$ is therefore multivariate normal with mean $\boldsymbol{\phi}_*$ and covariance matrix $\boldsymbol{\Sigma}_*$, which are given in Eq. (12.9) with $\boldsymbol{\beta}$ being replaced by $\boldsymbol{\phi}$ and $\boldsymbol{x}_{o,t}$ by $\boldsymbol{x}_{t-1}$. Similarly, the conditional posterior distribution of σ^2 is an inverted chi-squared distribution—that is,

$$\frac{v\lambda + \sum_{t=p+1}^{n} a_t^2}{\sigma^2} \sim \chi^2_{v+(n-p)},$$

where $a_t = x_t - \boldsymbol{\phi}' \boldsymbol{x}_{t-1}$ and $x_t = y_t - \delta_t \beta_t$.

The conditional posterior distribution of δ_h can be obtained as follows. First, δ_h is only related to $\{y_j, \beta_j\}_{j=h-p}^{h+p}$, $\{\delta_j\}_{j=h-p}^{h+p}$ with $j \neq h$, $\boldsymbol{\phi}$, and σ^2. More specifically, we have

$$x_j = y_j - \delta_j \beta_j, \qquad j \neq h.$$

Second, x_h can assume two possible values: $x_h = y_h - \beta_h$ if $\delta_h = 1$ and $x_h = y_h$, otherwise. Define

$$w_j = x_j^* - \phi_0 - \phi_1 x_{j-1}^* - \cdots - \phi_p x_{j-p}^*, \qquad j = h, \ldots, h+p,$$

where $x_j^* = x_j$ if $j \neq h$ and $x_h^* = y_h$. The two possible values of x_h give rise to two situations:

- Case I: $\delta_h = 0$. Here the hth observation is not an outlier and $x_h^* = y_h = x_h$. Hence, $w_j = a_j$ for $j = h, \ldots, h+p$. In other words, we have

$$w_j \sim N(0, \sigma^2), \qquad j = h, \ldots, h+p.$$

- Case II: $\delta_h = 1$. Now the hth observation is an outlier and $x_h^* = y_h = x_h + \beta_h$. The w_j defined before is contaminated by β_h. In fact, we have

$$w_h \sim N(\beta_h, \sigma^2) \quad \text{and} \quad w_j \sim N(-\phi_{j-h}\beta_h, \sigma^2), \quad j = h+1, \ldots, h+p.$$

  If we define $\psi_0 = -1$ and $\psi_i = \phi_i$ for $i = 1, \ldots, p$, then we have $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, \ldots, h+p$.

Based on the prior discussion, we can summarize the situation as follows:

1. Case I: $\delta_h = 0$ with probability $1 - \epsilon$. In this case, $w_j \sim N(0, \sigma^2)$ for $j = h, \ldots, h+p$.
2. Case II: $\delta_h = 1$ with probability ϵ. Here $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, \ldots, h+p$.

Since there are n data points, j cannot be greater than n. Let $m = \min(n, h+p)$. The posterior distribution of δ_h is therefore

$$P(\delta_h = 1 | \boldsymbol{Y}, \boldsymbol{\delta}_{-h}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2) = \frac{\epsilon \exp[-\sum_{j=h}^{m}(w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)]}{\epsilon \exp[-\sum_{j=h}^{m}(w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)] + (1-\epsilon)\exp[-\sum_{j=h}^{m} w_j^2/(2\sigma^2)]}. \tag{12.19}$$

This posterior distribution is simply to compare the weighted values of the likelihood function under the two situations with weight being the probability of each situation.

Finally, the posterior distribution of β_h is as follows.

- If $\delta_h = 0$, then y_h is not an outlier and $\beta_h \sim N(0, \xi^2)$.
- If $\delta_h = 1$, then y_h is contaminated by an outlier with magnitude β_h. The variable w_j defined before contains information of β_h for $j = h, h+1, \ldots, \min(h+p, n)$. Specifically, we have $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, h+1, \ldots, \min(h+p, n)$. The information can be put in a linear regression framework as

$$w_j = -\psi_{j-h}\beta_h + a_j, \qquad j = h, h+1, \ldots, \min(h+p, n).$$

  Consequently, the information is embedded in the least-squares estimate

$$\widehat{\beta}_h = \frac{\sum_{j=h}^{m} -\psi_{j-h} w_j}{\sum_{j=h}^{m} \psi_{j-h}^2}, \qquad m = \min(h+p, n),$$

which is normally distributed with mean β_h and variance $\sigma^2/(\sum_{j=h}^{m} \psi_{j-h}^2)$. By Result 12.1, the posterior distribution of β_h is normal with mean β_h^* and variance σ_{h*}^2, where

$$\beta_h^* = \frac{-(\sum_{j=h}^{m} \psi_{j-h} w_j)\xi^2}{\sigma^2 + (\sum_{j=h}^{m} \psi_{j-h}^2)\xi^2}, \qquad \sigma_{h*}^2 = \frac{\sigma^2 \xi^2}{\sigma^2 + (\sum_{j=h}^{m} \psi_{j-h}^2)\xi^2}.$$

Example 12.2. Consider the weekly change series of U.S. 3-year Treasury constant maturity interest rate from March 18, 1988, to September 10, 1999, for 600 observations. The interest rate is in percentage and is a subseries of the dependent variable c_{3t} of Example 12.1. The time series is shown in Figure 12.3(a). If AR models are entertained for the series, the partial autocorrelation function suggests an AR(3) model and we obtain

$$c_{3t} = 0.227 c_{3,t-1} + 0.006 c_{3,t-2} + 0.114 c_{3,t-2} + a_t, \qquad \widehat{\sigma}^2 = 0.0128,$$

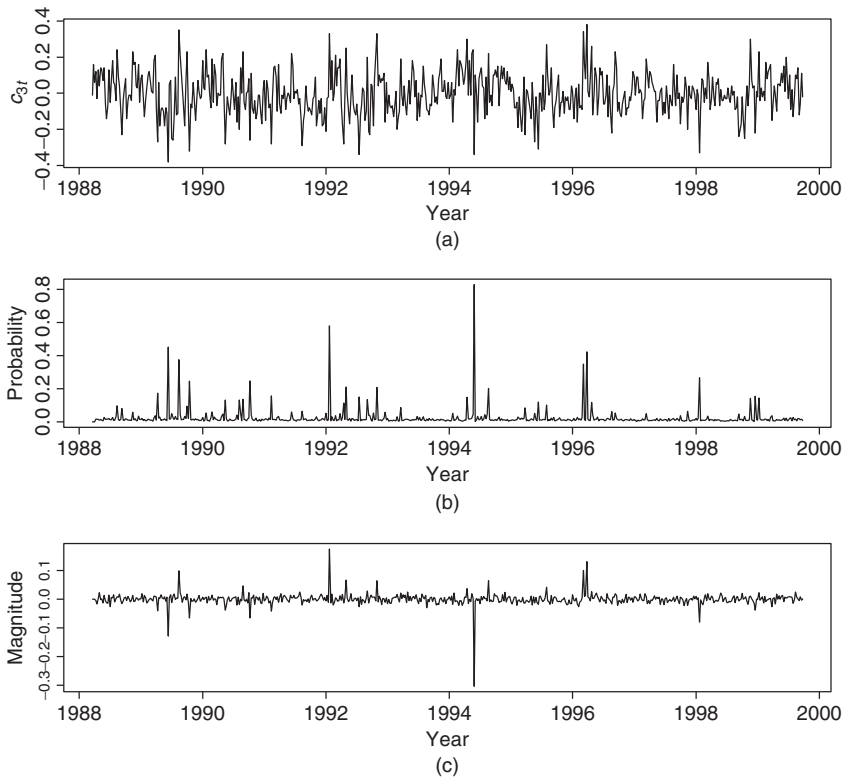

Figure 12.3 Time plots of weekly change series of U.S. 3-year Treasury constant maturity interest rate from March 18, 1988, to September 10, 1999: (a) data, (b) posterior probability of being an outlier, and (c) posterior mean of outlier size. Estimation is based on Gibbs sampling with 1050 iterations with first 50 iterations as burn-ins.

where standard errors of the coefficients are 0.041, 0.042, and 0.041, respectively. The Ljung–Box statistics of the residuals show $Q(12) = 11.4$, which is insignificant at the 5% level.

Next, we apply the Gibbs sampling to estimate the AR(3) model and to detect simultaneously possible additive outliers. The prior distributions used are

$$\boldsymbol{\phi} \sim N(\mathbf{0}, 0.25\boldsymbol{I}_3), \quad \frac{v\lambda}{\sigma^2} = \frac{5 \times 0.00256}{\sigma^2} \sim \chi^2_5, \quad \gamma_1 = 5, \quad \gamma_2 = 95, \quad \xi^2 = 0.1,$$

where $0.00256 \approx \widehat{\sigma}^2/5$ and $\xi^2 \approx 9\widehat{\sigma}^2$. The expected number of additive outliers is 5%. Using initial values $\epsilon = 0.05$, $\sigma^2 = 0.012$, $\phi_1 = 0.2$, $\phi_2 = 0.02$, and $\phi_3 = 0.1$, we run the Gibbs sampling for 1050 iterations but discard results of the first 50 iterations. Using posterior means of the coefficients as parameter estimates, we obtain the fitted model

$$c_{3t} = 0.252c_{3,t-1} + 0.003c_{3,t-2} + 0.110c_{3,t-2} + a_t, \qquad \widehat{\sigma}^2 = 0.0118,$$

where posterior standard deviations of the parameters are 0.046, 0.045, 0.046, and 0.0008, respectively. Thus, the Gibbs sampling produces results similar to that of the maximum-likelihood method. Figure 12.3(b) shows the time plot of posterior probability of each observation being an additive outlier, and Figure 12.3(c) plots the posterior mean of outlier magnitude. From the probability plot, some observations have high probabilities of being an outlier. In particular, $t = 323$ has a probability of 0.83 and the associated posterior mean of outlier magnitude is -0.304. This point corresponds to May 20, 1994, when the c_{3t} changed from 0.24 to -0.34 (i.e., about a 0.6% drop in the weekly interest rate within 2 weeks). The point with second highest posterior probability of being an outlier is $t = 201$, which is January 17, 1992. The outlying posterior probability is 0.58 and the estimated outlier size is 0.176. At this particular time point, c_{3t} changed from -0.02 to 0.33, corresponding to a jump of about 0.35% in the weekly interest rate.

Remark. Outlier detection via Gibbs sampling requires intensive computation but the approach performs a joint estimation of model parameters and outliers. Yet the traditional approach to outlier detection separates estimation from detection. It is much faster in computation, but may produce spurious detections when multiple outliers are present. For the data in Example 12.2, the SCA program also identifies $t = 323$ and $t = 201$ as the two most significant additive outliers. The estimated outlier sizes are -0.39 and 0.36, respectively. □

12.7 STOCHASTIC VOLATILITY MODELS

An important financial application of MCMC methods is the estimation of stochastic volatility models; see Jacquier, Polson, and Rossi (1994) and the references

therein. We start with a univariate stochastic volatility model. The mean and volatility equations of an asset return r_t are

$$r_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + a_t, \qquad a_t = \sqrt{h_t}\epsilon_t, \tag{12.20}$$

$$\ln h_t = \alpha_0 + \alpha_1 \ln h_{t-1} + v_t, \tag{12.21}$$

where $\{x_{it} | i = 1, \ldots, p\}$ are explanatory variables available at time $t - 1$, the β_j are parameters, $\{\epsilon_t\}$ is a Gaussian white noise sequence with mean 0 and variance 1, $\{v_t\}$ is also a Gaussian white noise sequence with mean 0 and variance σ_v^2, and $\{\epsilon_t\}$ and $\{v_t\}$ are independent. The log transformation is used to ensure that h_t is positive for all t. The explanatory variables x_{it} may include lagged values of the return (e.g., $x_{it} = r_{t-i}$). In Eq. (12.21), we assume that $|\alpha_1| < 1$ so that the log volatility process $\ln h_t$ is stationary. If necessary, a higher order AR(p) model can be used for $\ln h_t$.

Denote the coefficient vector of the mean equation by $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ and the parameter vector of the volatility equation by $\boldsymbol{\omega} = (\alpha_0, \alpha_1, \sigma_v^2)'$. Suppose that $\boldsymbol{R} = (r_1, \ldots, r_n)'$ is the collection of observed returns and $\boldsymbol{X}$ is the collection of explanatory variables. Let $\boldsymbol{H} = (h_1, \ldots, h_n)'$ be the vector of unobservable volatilities. Here $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are the "traditional" parameters of the model and $\boldsymbol{H}$ is an auxiliary variable. Estimation of the model would be complicated via the maximum-likelihood method because the likelihood function is a mixture over the n-dimensional $\boldsymbol{H}$ distribution as

$$f(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\omega}) = \int f(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{H}) f(\boldsymbol{H}|\boldsymbol{\omega})\, d\boldsymbol{H}.$$

However, under the Bayesian framework, the volatility vector $\boldsymbol{H}$ consists of augmented parameters. Conditioning on $\boldsymbol{H}$, we can focus on the probability distribution functions $f(\boldsymbol{R}|\boldsymbol{H}, \boldsymbol{\beta})$ and $f(\boldsymbol{H}|\boldsymbol{\omega})$ and the prior distribution $p(\boldsymbol{\beta}, \boldsymbol{\omega})$. We assume that the prior distribution can be partitioned as $p(\boldsymbol{\beta}, \boldsymbol{\omega}) = p(\boldsymbol{\beta})p(\boldsymbol{\omega})$; that is, prior distributions for the mean and volatility equations are independent. A Gibbs sampling approach to estimating the stochastic volatility in Eqs. (12.20) and (12.21) then involves drawing random samples from the following conditional posterior distributions:

$$f(\boldsymbol{\beta}|\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\omega}), \qquad f(\boldsymbol{H}|\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\omega}), \qquad f(\boldsymbol{\omega}|\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{H}).$$

In what follows, we give details of practical implementation of the Gibbs sampling used.

12.7.1 Estimation of Univariate Models

Given $\boldsymbol{H}$, the mean equation in (12.20) is a nonhomogeneous linear regression. Dividing the equation by $\sqrt{h_t}$, we can write the model as

$$r_{o,t} = \boldsymbol{x}'_{o,t}\boldsymbol{\beta} + \epsilon_t, \qquad t = 1, \ldots, n, \tag{12.22}$$

where $r_{o,t} = r_t/\sqrt{h_t}$ and $\boldsymbol{x}_{o,t} = \boldsymbol{x}_t/\sqrt{h_t}$, with $\boldsymbol{x}_t = (1, x_{1t}, \ldots, x_{pt})'$ being the vector of explanatory variables. Suppose that the prior distribution of $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{\beta}_o$ and covariance matrix $\boldsymbol{A}_o$. Then the posterior distribution of $\boldsymbol{\beta}$ is also multivariate normal with mean $\boldsymbol{\beta}_*$ and covariance matrix $\boldsymbol{A}_*$. These two quantities can be obtained as before via Result 12.1a, and they are

$$\boldsymbol{A}_*^{-1} = \sum_{t=1}^{n} \boldsymbol{x}_{o,t}\boldsymbol{x}_{o,t}' + \boldsymbol{A}_o^{-1}, \qquad \boldsymbol{\beta}_* = \boldsymbol{A}_* \left(\sum_{t=1}^{n} \boldsymbol{x}_{o,t} r_{o,t} + \boldsymbol{A}_o^{-1}\boldsymbol{\beta}_o \right),$$

where it is understood that the summation starts with $p + 1$ if r_{t-p} is the highest lagged return used in the explanatory variables.

The volatility vector $\boldsymbol{H}$ is drawn element by element. The necessary conditional posterior distribution is $f(h_t|\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{H}_{-t}, \boldsymbol{\beta}, \boldsymbol{\omega})$, which is produced by the normal distribution of a_t and the lognormal distribution of the volatility,

$$\begin{aligned} & f(h_t|\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{H}_{-t}, \boldsymbol{\omega}) \\ & \quad \propto f(a_t|h_t, r_t, \boldsymbol{x}_t, \boldsymbol{\beta}) f(h_t|h_{t-1}, \boldsymbol{\omega}) f(h_{t+1}|h_t, \boldsymbol{\omega}) \\ & \quad \propto h_t^{-0.5} \exp[-(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2/(2h_t)] h_t^{-1} \exp[-(\ln h_t - \mu_t)^2/(2\sigma^2)] \\ & \quad \propto h_t^{-1.5} \exp[-(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2/(2h_t) - (\ln h_t - \mu_t)^2/(2\sigma^2)], \end{aligned} \tag{12.23}$$

where $\mu_t = [\alpha_0(1 - \alpha_1) + \alpha_1(\ln h_{t+1} + \ln h_{t-1})]/(1 + \alpha_1^2)$ and $\sigma^2 = \sigma_v^2/(1 + \alpha_1^2)$. Here we have used the following properties: (a) $a_t|h_t \sim N(0, h_t)$; (b) $\ln h_t|\ln h_{t-1} \sim N(\alpha_0 + \alpha_1 \ln h_{t-1}, \sigma_v^2)$; (c) $\ln h_{t+1}|\ln h_t \sim N(\alpha_0 + \alpha_1 \ln h_t, \sigma_v^2)$; (d) $d \ln h_t = h_t^{-1}\, dh_t$, where d denotes differentiation; and (e) the equality

$$(x - a)^2 A + (x - b)^2 C = (x - c)^2 (A + C) + (a - b)^2 AC/(A + C),$$

where $c = (Aa + Cb)/(A + C)$ provided that $A + C \neq 0$. This equality is a scalar version of Lemma 1 of Box and Tiao (1973, p. 418). In our application, $A = 1$, $a = \alpha_0 + \ln h_{t-1}$, $C = \alpha_1^2$, and $b = (\ln h_{t+1} - \alpha_0)/\alpha_1$. The term $(a - b)^2 AC/(A + C)$ does not contain the random variable h_t and, hence, is integrated out in the derivation of the conditional posterior distribution. Jacquier, Polson, and Rossi (1994) use the Metropolis algorithm to draw h_t. We use Griddy Gibbs in this section, and the range of h_t is chosen to be a multiple of the unconditional sample variance of r_t.

To draw random samples of $\boldsymbol{\omega}$, we partition the parameters as $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$ and σ_v^2. The prior distribution of $\boldsymbol{\omega}$ is also partitioned accordingly [i.e., $p(\boldsymbol{\omega}) = p(\boldsymbol{\alpha})p(\sigma_v^2)$]. The conditional posterior distributions needed are

- $f(\boldsymbol{\alpha}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\beta}, \sigma_v^2) = f(\boldsymbol{\alpha}|\boldsymbol{H}, \sigma_v^2)$: Given $\boldsymbol{H}$, $\ln h_t$ follows an AR(1) model. Therefore, the result of AR models discussed in the previous two sections applies. Specifically, if the prior distribution of $\boldsymbol{\alpha}$ is multivariate

normal with mean $\boldsymbol{\alpha}_o$ and covariance matrix C_o, then $f(\boldsymbol{\alpha}|\boldsymbol{H}, \sigma_v^2)$ is multivariate normal with mean $\boldsymbol{\alpha}_*$ and covariance matrix $\boldsymbol{C}_*$, where

$$\boldsymbol{C}_*^{-1} = \frac{\sum_{t=2}^{n} \boldsymbol{z}_t \boldsymbol{z}_t'}{\sigma_v^2} + \boldsymbol{C}_o^{-1}, \qquad \boldsymbol{\alpha}_* = \boldsymbol{C}_* \left(\frac{\sum_{t=2}^{n} \boldsymbol{z}_t \ln h_t}{\sigma_v^2} + \boldsymbol{C}_o^{-1} \boldsymbol{\alpha}_o \right),$$

where $\boldsymbol{z}_t = (1, \ln h_{t-1})'$.

- $f(\sigma_v^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\sigma_v^2|\boldsymbol{H}, \boldsymbol{\alpha})$: Given $\boldsymbol{H}$ and $\boldsymbol{\alpha}$, we can calculate $v_t = \ln h_t - \alpha_0 - \alpha_1 \ln h_{t-1}$ for $t = 2, \ldots, n$. Therefore, if the prior distribution of σ_v^2 is $(m\lambda)/\sigma_v^2 \sim \chi_m^2$, then the conditional posterior distribution of σ_v^2 is an inverted chi-squared distribution with $m + n - 1$ degrees of freedom; that is,

$$\frac{m\lambda + \sum_{t=2}^{n} v_t^2}{\sigma_v^2} \sim \chi_{m+n-1}^2.$$

Remark. Formula (12.23) is for $1 < t < n$, where n is the sample size. For the two end data points h_1 and h_n, some modifications are needed. A simple approach is to assume that h_1 is fixed so that the drawing of h_t starts with $t = 2$. For $t = n$, one uses the result $\ln h_n \sim (\alpha_0 + \alpha_1 \ln h_{n-1}, \sigma_v^2)$. Alternatively, one can employ a forecast of h_{n+1} and a backward prediction of h_0 and continue to apply the formula. Since h_n is the variable of interest, we forecast h_{n+1} by using a 2-step-ahead forecast at the forecast origin $n - 1$. For the model in Eq. (12.21), the forecast of h_{n+1} is

$$\widehat{h}_{n-1}(2) = \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 \ln h_{n-1}).$$

The backward prediction of h_0 is based on the time reversibility of the model

$$(\ln h_t - \eta) = \alpha_1(\ln h_{t-1} - \eta) + v_t,$$

where $\eta = \alpha_0/(1 - \alpha_1)$ and $|\alpha_1| < 1$. The model of the reversed series is

$$(\ln h_t - \eta) = \alpha_1(\ln h_{t+1} - \eta) + v_t^*,$$

where $\{v_t^*\}$ is also a Gaussian white noise series with mean zero and variance σ_v^2. Consequently, the 2-step-backward prediction of h_0 at time $t = 2$ is

$$\widehat{h}_2(-2) = \alpha_1^2(\ln h_2 - \eta). \qquad \square$$

Remark. Formula (12.23) can also be obtained by using results of a missing value in an AR(1) model; see Section 12.6.1. Specifically, assume that $\ln h_t$ is missing. For the AR(1) model in Eq. (12.21), this missing value is related to $\ln h_{t-1}$ and $\ln h_{t+1}$ for $1 < t < n$. From the model, we have

$$\ln h_t = \alpha_0 + \alpha_1 \ln h_{t-1} + a_t.$$

Define $y_t = \alpha_0 + \alpha_1 y_{t-1}$, $x_t = 1$, and $b_t = -a_t$. Then we obtain

$$y_t = x_t \ln h_t + b_t. \tag{12.24}$$

Next, from

$$\ln h_{t+1} = \alpha_0 + \alpha_1 \ln h_t + a_{t+1},$$

we define $y_{t+1} = \ln h_{t+1} - \alpha_0$, $x_{t+1} = \alpha_1$, and $b_{t+1} = a_{t+1}$ and obtain

$$y_{t+1} = x_{t+1} \ln h_{t+1} + b_{t+1}. \tag{12.25}$$

Now Eqs. (12.24) and (12.25) form a special simple linear regression with two observations and an unknown parameter $\ln h_t$. Note that b_t and b_{t+1} have the same distribution because $-a_t$ is also $N(0, \sigma_v^2)$. The least-squares estimate of $\ln h_t$ is then

$$\widehat{\ln h_t} = \frac{x_t y_t + x_{t+1} y_{t+1}}{x_t^2 + x_{t+1}^2} = \frac{\alpha_0(1-\alpha_1) + \alpha_1(\ln h_{t+1} + \ln h_{t-1})}{1+\alpha_1^2},$$

which is precisely the conditional mean of $\ln h_t$ given in Eq. (12.23). In addition, this estimate is normally distributed with mean $\ln h_t$ and variance $\sigma_v^2/(1+\alpha_1^2)$. Formula (12.23) is simply the product of $a_t \sim N(0, h_t)$ and $\widehat{\ln h_t} \sim N[\ln h_t, \sigma_v^2/(1+\alpha_1^2)]$ with the transformation $d \ln h_t = h_t^{-1} dh_t$. This regression approach generalizes easily to other AR(p) models for $\ln h_t$. We use this approach and assume that $\{h_t\}_{t=1}^p$ are fixed for a stochastic volatility AR(p) model. □

Remark. Starting value of h_t can be obtained by fitting a volatility model of Chapter 3 to the return series. □

Example 12.3. Consider the monthly log returns of the S&P 500 index from January 1962 to December 2009 for 575 observations. The returns are computed using the first adjusted closing index of each month, that is, the closing index of the first trading day of each month. Figure 12.4(a) shows the time plot of the log level of the index, whereas Figure 12.4(b) shows the log returns measured in percentage. If GARCH models are entertained for the series, we obtain a Gaussian GARCH(1,1) model

$$\begin{aligned} r_t &= 0.552 + a_t, \qquad a_t = \sqrt{h_t}\epsilon_t, \\ h_t &= 0.878 + 0.125 a_{t-1}^2 + 0.837 h_{t-1}, \end{aligned} \tag{12.26}$$

where t ratios of the coefficients are all greater than 2.56. The Ljung–Box statistics of the standardized residuals and their squared series fail to indicate any model inadequacy. Specifically, we have $Q(12) = 10.04(0.61)$ and $6.14(0.91)$, respectively, for the standardized residuals and their squared series.

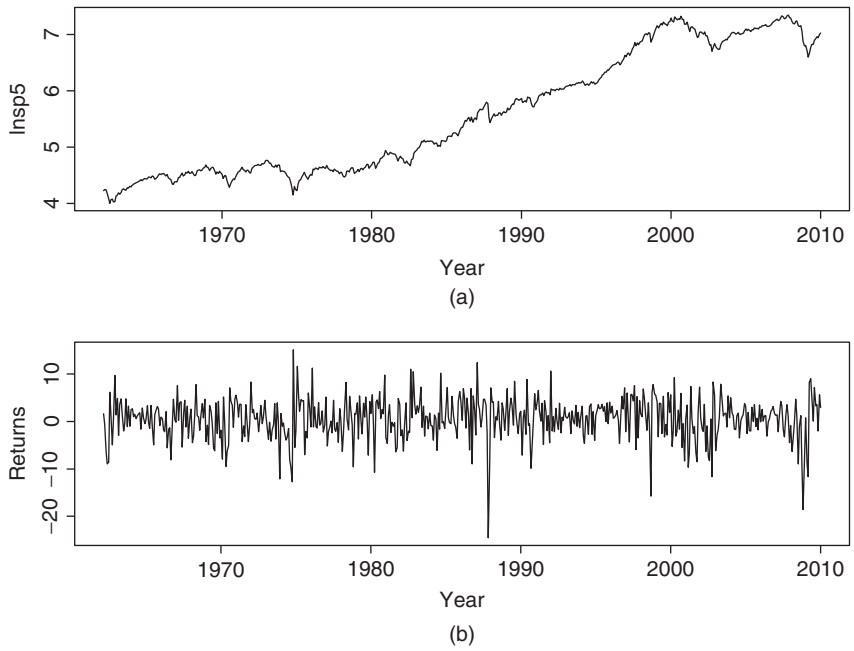

Figure 12.4 Time plot of monthly S&P 500 index from 1962 to 2009: (a) log level and (b) log return in percentage.

Next, consider the stochastic volatility model

$$r_t = \mu + a_t, \qquad a_t = \sqrt{h_t}\epsilon_t,$$
$$\ln h_t = \alpha_0 + \alpha_1 \ln h_{t-1} + v_t, \tag{12.27}$$

where the v_t are iid $N(0, \sigma_v^2)$. To implement the Gibbs sampling, we use the prior distributions

$$\mu \sim N(0, 4), \qquad \boldsymbol{\alpha} \sim N[\boldsymbol{\alpha}_o, \text{diag}(0.25, 0.04)], \qquad \frac{10 \times 0.1}{\sigma_v^2} \sim \chi_{10}^2,$$

where $\boldsymbol{\alpha}_o = (0, 0.6)'$. For initial parameter values, we use the fitted values of the GARCH(1,1) model in Eq. (12.26) for $\{h_t\}$, that is, $h_{0t} = h_t$, and set $\boldsymbol{\alpha}$ and σ_v^2 to the least-squares estimate of $\ln(h_{0t})$. The initial value of μ is the sample mean of the log returns. The volatility h_t is drawn by the Griddy Gibbs with 400 grid points. The possible range of h_t for the jth Gibbs iteration is $[\eta_{1t}, \eta_{2t}]$, where $\eta_{1t} = 0.6 \times \max(h_{j-1,t}, h_{0t})$ and $\eta_{2t} = 1.4 \times \min(h_{j-1,t}, h_{0t})$, where $h_{j-1,t}$ and h_{0t} denote, respectively, the estimate of h_t for the $(j-1)$th iteration and initial value.

We ran the Gibbs sampling for 2500 iterations but discarded results of the first 500 iterations. Figure 12.5 shows the density functions of the prior and posterior

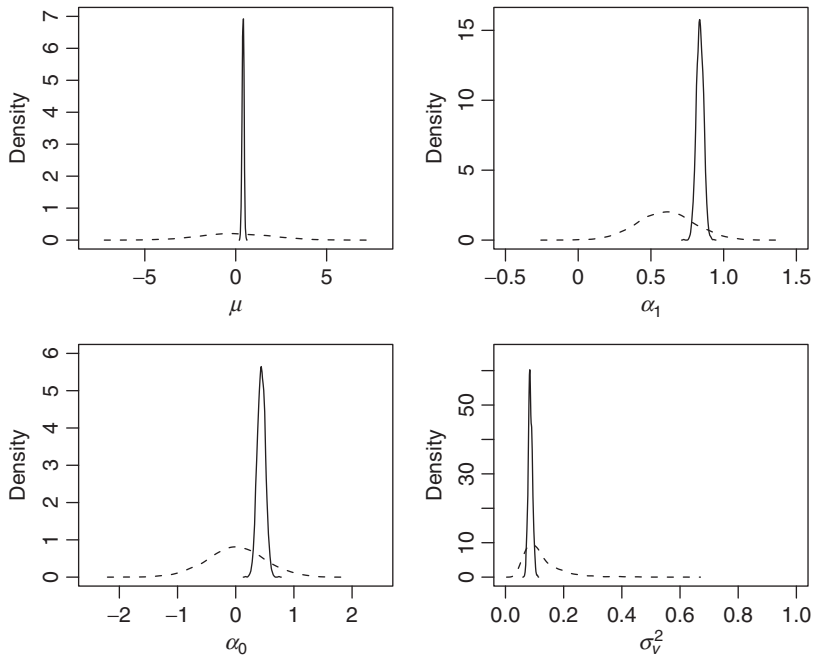

Figure 12.5 Density functions of prior and posterior distributions of parameters in stochastic volatility model for monthly log returns of S&P 500 index. Dashed line denotes prior density and solid line the posterior density, which is based on results of Gibbs sampling with 2000 iterations. See text for more details.

distributions of the four coefficient parameters. The prior distributions used are relatively noninformative. The posterior distributions are concentrated especially for μ and σ_v^2. Figure 12.6 shows the time plots of fitted volatilities. The upper panel shows the posterior mean of h_t over the 5000 iterations for each time point, whereas the lower panel shows the fitted values of the GARCH(1,1) model in Eq. (12.26). The two plots exhibit a similar pattern.

The posterior mean and standard error of the four coefficients are as follows:

Parameter	μ	α_0	α_1	σ_v^2
Mean	0.409	0.454	0.837	0.086
Standard error	0.157	0.068	0.025	0.007

The posterior mean of α_1 is 0.837, confirming strong serial dependence in the volatility series. This value is smaller than that obtained by Jacquier, Polson, and Rossi (1994) who used daily returns of the S&P 500 index. Finally, we have used different initial values, priors, and numbers of iterations for the Gibbs sampler. The

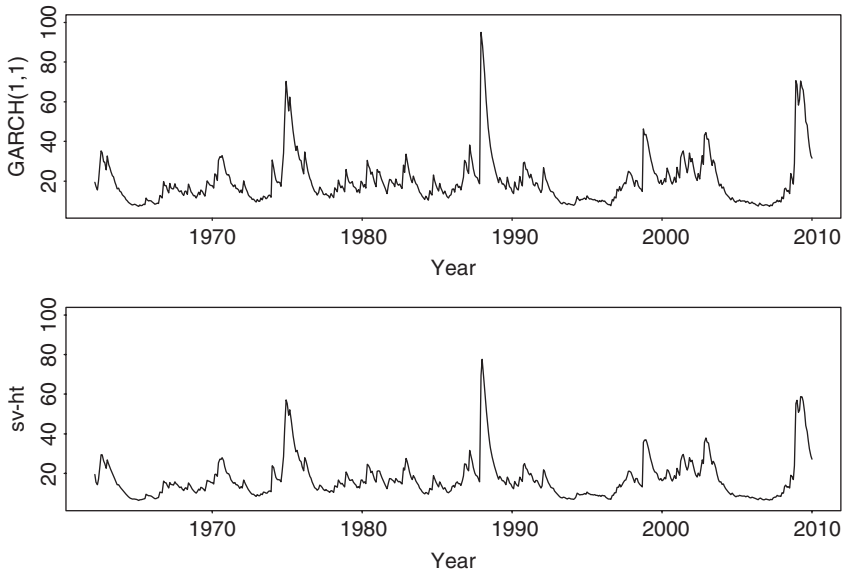

Figure 12.6 Time plots of fitted volatilities for monthly log returns of S&P 500 index from 1962 to 2009. Lower panel shows posterior means of a Gibbs sampler with 2000 iterations. Upper panel shows results of a Gaussian GARCH(1,1) model.

results are stable. Of course, as expected, the results and efficiency of the Griddy Gibbs algorithm depend on the specification of the range for h_t.

12.7.2 Multivariate Stochastic Volatility Models

In this section, we study multivariate stochastic volatility models using the Cholesky decomposition of Chapter 10. We focus on the bivariate case, but the methods discussed also apply to the higher dimensional case. Based on the Cholesky decomposition, the innovation $\boldsymbol{a}_t$ of a return series $\boldsymbol{r}_t$ is transformed into $\boldsymbol{b}_t$ such that

$$b_{1t} = a_{1t}, \qquad b_{2t} = a_{2t} - q_{21,t} b_{1t},$$

where b_{2t} and $q_{21,t}$ can be interpreted as the residual and least-squares estimate of the linear regression

$$a_{2t} = q_{21,t} a_{1t} + b_{2t}.$$

The conditional covariance matrix of $\boldsymbol{a}_t$ is parameterized by $\{g_{11,t}, g_{22,t}\}$ and $\{q_{21,t}\}$ as

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{21,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ q_{21,t} & 1 \end{bmatrix} \begin{bmatrix} g_{11,t} & 0 \\ 0 & g_{22,t} \end{bmatrix} \begin{bmatrix} 1 & q_{21,t} \\ 0 & 1 \end{bmatrix}, \tag{12.28}$$

where $g_{ii,t} = \text{Var}(b_{it}|F_{t-1})$ and $b_{1t} \perp b_{2t}$. Thus, the quantities of interest are $g_{11,t}$, $g_{22,t}$ and $q_{21,t}$.

A simple bivariate stochastic volatility model for the return $\boldsymbol{r}_t = (r_{1t}, r_{2t})'$ is as follows:

$$\boldsymbol{r}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_t + \boldsymbol{a}_t, \tag{12.29}$$

$$\ln g_{ii,t} = \alpha_{i0} + \alpha_{i1} \ln g_{ii,t-1} + v_{it}, \qquad i = 1, 2, \tag{12.30}$$

$$q_{21,t} = \gamma_0 + \gamma_1 q_{21,t-1} + u_t, \tag{12.31}$$

where $\{\boldsymbol{a}_t\}$ is a sequence of serially uncorrelated Gaussian random vectors with mean zero and conditional covariance matrix $\boldsymbol{\Sigma}_t$ given by Eq. (12.28), $\boldsymbol{\beta}_0$ is a two-dimensional constant vector, $\boldsymbol{x}_t$ denotes the explanatory variables, and $\{v_{1t}\}$, $\{v_{2t}\}$, and $\{u_t\}$ are three independent Gaussian white noise series such that $\text{Var}(v_{it}) = \sigma_{iv}^2$ and $\text{Var}(u_t) = \sigma_u^2$. Again log transformation is used in Eq. (12.30) to ensure the positiveness of $g_{ii,t}$.

Let $\boldsymbol{G}_i = (g_{ii,1}, \ldots, g_{ii,n})'$, $\boldsymbol{G} = [\boldsymbol{G}_1, \boldsymbol{G}_2]$, and $\boldsymbol{Q} = (q_{21,1}, \ldots, q_{21,n})'$. The "traditional" parameters of the model in Eqs. (12.29)–(12.31) are $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, $\boldsymbol{\alpha}_i = (\alpha_{i0}, \alpha_{i1})'$, and σ_{iv}^2 for $i = 1, 2$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)'$ and σ_u^2. The augmented parameters are $\boldsymbol{Q}$, $\boldsymbol{G}_1$, and $\boldsymbol{G}_2$. To estimate such a bivariate stochastic volatility model via Gibbs sampling, we use results of the univariate model in the previous section and two additional conditional posterior distributions. Specifically, we can draw random samples of

1. $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ row by row using the result (12.22)
2. $g_{11,t}$ using Eq. (12.23) with a_t being replaced by a_{1t}
3. $\boldsymbol{\alpha}_1$ and σ_{1v}^2 using exactly the same methods as those of the univariate case with a_t replaced by a_{1t}

To draw random samples of $\boldsymbol{\alpha}_2$, σ_{2v}^2, and $g_{22,t}$, we need to compute b_{2t}. But this is easy because $b_{2t} = a_{2t} - q_{21,t} a_{1t}$ given the augmented parameter vector $\boldsymbol{Q}$. Furthermore, b_{2t} is normally distributed with mean 0 and conditional variance $g_{22,t}$.

It remains to consider the conditional posterior distributions

$$f(\boldsymbol{\gamma}|\boldsymbol{Q}, \sigma_u^2), \qquad f(\sigma_u^2|\boldsymbol{Q}, \boldsymbol{\gamma}), \qquad f(q_{21,t}|\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{Q}_{-t}, \boldsymbol{\gamma}, \sigma_u^2),$$

where $\boldsymbol{A}$ denotes the collection of $\boldsymbol{a}_t$, which is known if $\boldsymbol{R}$, $\boldsymbol{X}$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$ are given. Given $\boldsymbol{Q}$ and σ_u^2, model (12.31) is a simple Gaussian AR(1) model. Therefore, if the prior distribution of $\boldsymbol{\gamma}$ is bivariate normal with mean $\boldsymbol{\gamma}_o$ and covariance matrix $\boldsymbol{D}_o$, then the conditional posterior distribution of $\boldsymbol{\gamma}$ is also bivariate normal with mean $\boldsymbol{\gamma}_*$ and covariance matrix $\boldsymbol{D}_*$, where

$$\boldsymbol{D}_*^{-1} = \frac{\sum_{t=2}^n \boldsymbol{z}_t \boldsymbol{z}_t'}{\sigma_u^2} + \boldsymbol{D}_o^{-1}, \qquad \boldsymbol{\gamma}_* = \boldsymbol{D}_* \left(\frac{\sum_{t=2}^n \boldsymbol{z}_t q_{21,t}}{\sigma_u^2} + \boldsymbol{D}_o^{-1} \boldsymbol{\gamma}_o \right),$$

where $z_t = (1, q_{21,t-1})'$. Similarly, if the prior distribution of σ_u^2 is $(m\lambda)/\sigma_u^2 \sim \chi_m^2$, then the conditional posterior distribution of σ_u^2 is

$$\frac{m\lambda + \sum_{t=2}^{n} u_t^2}{\sigma_u^2} \sim \chi_{m+n-1}^2,$$

where $u_t = q_{21,t} - \gamma_0 - \gamma_1 q_{21,t-1}$. Finally,

$$\begin{aligned} f(q_{21,t}|\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{Q}_{-t}, \sigma_u^2, \boldsymbol{\gamma}) &\propto f(\boldsymbol{b_{2t}}|g_{22,t}) f(q_{21,t}|q_{21,t-1}, \boldsymbol{\gamma}, \sigma_u^2) f(q_{21,t+1}|q_{21,t}, \boldsymbol{\gamma}, \sigma_u^2) \\ &\propto g_{22,t}^{-0.5} \exp[-(a_{2t} - q_{21,t}a_{1t})^2/(2g_{22,t})] \exp[-(q_{21,t} - \mu_t)^2(2\sigma^2)], \end{aligned} \tag{12.32}$$

where $\mu_t = [\gamma_0(1-\gamma_1) + \gamma_1(q_{21,t-1} + q_{21,t+1})]/(1+\gamma_1^2)$ and $\sigma^2 = \sigma_u^2/(1+\gamma_1^2)$. In general, μ_t and σ^2 can be obtained by using the results of a missing value in an AR(p) process. It turns out that Eq. (12.32) has a closed-form distribution for $q_{21,t}$. Specifically, the first term of Eq. (12.32), which is the conditional distribution of $q_{21,t}$ given $g_{22,t}$ and $\boldsymbol{a}_t$, is normal with mean a_{2t}/a_{1t} and variance $g_{22,t}/(a_{1t})^2$. The second term of the equation is also normal with mean μ_t and variance σ^2. Consequently, by Result 12.1, the conditional posterior distribution of $q_{21,t}$ is normal with mean μ_* and variance σ_*^2, where

$$\frac{1}{\sigma_*^2} = \frac{a_{1t}^2}{g_{22,t}} + \frac{1+\gamma_1^2}{\sigma_u^2}, \qquad \mu_* = \sigma_*^2 \left(\frac{1+\gamma_1^2}{\sigma_u^2} \times \mu_t + \frac{a_{1t}^2}{g_{22,t}} \times \frac{a_{2t}}{a_{1t}} \right)$$

where μ_t is defined in Eq. (12.32).

Example 12.4. In this example, we study bivariate volatility models for the monthly log returns of IBM stock and the S&P 500 index from January 1962 to December 2009. This is an expanded version of Example 12.3 by adding the IBM returns. Figure 12.7 shows the time plots of the two return series. Let $\boldsymbol{r}_t = (\text{IBM}_t, \text{SP}_t)'$. If time-varying correlation GARCH models with Cholesky decomposition of Chapter 10 are entertained, we obtain the model

$$\boldsymbol{r}_t = \boldsymbol{\beta}_0 + \boldsymbol{a}_t, \tag{12.33}$$

$$g_{11,t} = \alpha_{10} + \alpha_{11} g_{11,t-1} + \alpha_{12} a_{1,t-1}^2, \tag{12.34}$$

$$g_{22,t} = \alpha_{20} + \alpha_{22} b_{2,t-1}^2, \tag{12.35}$$

$$q_{21,t} = \gamma_0, \tag{12.36}$$

where $b_{2t} = a_{2t} - q_{21,t}a_{1t}$ and the estimates and their standard errors are given in Table 12.2(a). For comparison purpose, we also fit a BEKK(1,1) model and obtain $\widehat{\boldsymbol{\beta}}_0 = (0.70, 0.54)'$ and the coefficient matrices

$$\boldsymbol{A} = \begin{bmatrix} 0.80 & \\ 0.83 & 0.01 \end{bmatrix}, \quad \boldsymbol{A}_1 = \begin{bmatrix} 0.07 & 0.33 \\ -0.06 & 0.43 \end{bmatrix}, \quad \boldsymbol{B}_1 = \begin{bmatrix} 1.00 & -0.12 \\ 0.01 & 0.90 \end{bmatrix},$$

where the matrices are defined in Eq. (10.6) of Chapter 10.

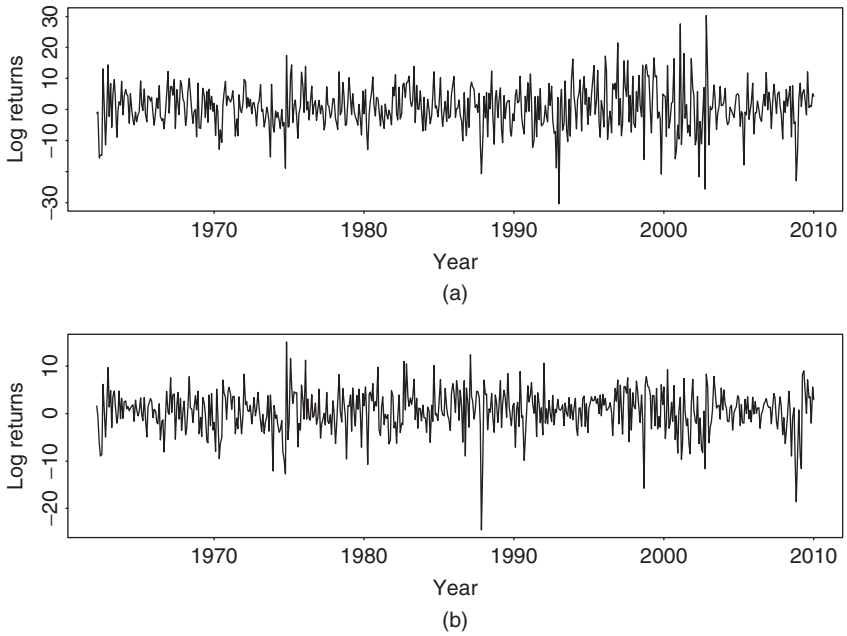

Figure 12.7 Time plots of monthly log returns of (a) IBM stock and (b) S&P 500 index from 1962 to 2009.

For stochastic volatility model, we employ the same mean equation in Eq. (12.33) and a stochastic volatility model similar to that in Eqs. (12.34)–(12.36). The volatility equations are

$$\ln g_{11,t} = \alpha_{10} + \alpha_{11} \ln g_{11,t-1} + v_{1t}, \qquad \text{Var}(v_{1t}) = \sigma_{1v}^2, \tag{12.37}$$

$$\ln g_{22,t} = \alpha_{20} + \alpha_{21} \ln g_{22,t-1} + v_{2t}, \qquad \text{Var}(v_{2t}) = \sigma_{2v}^2, \tag{12.38}$$

$$q_{21,t} = \gamma_0 + u_t, \qquad \text{Var}(u_t) = \sigma_u^2. \tag{12.39}$$

The prior distributions used are

$$\beta_{i0} \sim N(0, 4), \qquad \boldsymbol{\alpha}_i \sim N[(0, 0.7)', \text{diag}(0.25, 0.04)],$$

$$\gamma_0 \sim N(0, 1), \qquad \frac{10 \times 0.1}{\sigma_{iv}^2} \sim \chi_{10}^2, \qquad \frac{5 \times 0.2}{\sigma_u^2} \sim \chi_5^2,$$

where $i = 1$ and 2. These prior distributions are relatively noninformative. We obtained the initial values of $\{g_{11,t}, g_{22,t}, q_{21,t}\}$ from the results of the BEKK(1,1) model. In addition, we set the values of quantities at $t = 1$ as given. We then ran the Gibbs sampling for 2500 iterations but discarded results of the first 500 iterations. The random samples of $g_{ii,t}$ were drawn by Griddy Gibbs with 500 grid points in the intervals $[\eta_{i,1t}, \eta_{i,2t}]$ where the lower and upper bounds are set by

TABLE 12.2 Estimation of Bivariate Volatility Models for Monthly Log Returns of IBM Stock and S&P 500 Index from January 1962 to December 2009[a]

(a) Bivariate GARCH(1,1) Model With Time-Varying Correlations

Parameter	β_{01}	β_{02}	α_{10}	α_{11}	α_{12}	α_{20}	α_{22}	γ_0
Estimate	0.69	0.49	3.98	0.80	0.12	10.67	0.12	0.37
Standard error	0.30	0.18	1.22	0.04	0.03	0.53	0.04	0.01

(b) Stochastic Volatility Model

Parameter	β_{01}	β_{02}	α_{10}	α_{11}	σ_{1v}^2	α_{20}	α_{21}	σ_{2v}^2	γ_0	σ_u^2
Posterior mean	0.53	0.51	0.75	0.80	0.07	0.43	0.81	0.07	0.38	0.07
Standard error	0.26	0.17	0.11	0.03	0.01	0.06	0.03	0.01	0.03	0.01

[a]The stochastic volatility models are based on the last 2000 iterations of a Gibbs sampling with 2500 total iterations.

the same method as those of Example 12.3. Posterior means and standard errors of the "traditional" parameters of the bivariate stochastic volatility model are given in Table 12.2(b).

To check for convergence of the Gibbs sampling, we ran the procedure several times with different starting values and numbers of iterations. The results are stable. For illustration, Figure 12.8 shows the scatterplots of various quantities for two different Gibbs samples. The first Gibbs sample is based on $500 + 2000$ iterations, and the second Gibbs sample is based on $500 + 1000$ iterations, where $M + N$ denotes that the total number of Gibbs iterations is $M + N$, but results of the first M iterations are discarded. The scatterplots shown are posterior means of $g_{11,t}$, $g_{22,t}$, $g_{21,t}$, $\sigma_{22,t}$, $\sigma_{21,t}$, and the correlation $\rho_{21,t}$. The line $y = x$ is added to each plot to show the closeness of the posterior means. The stability of the Gibbs sampling results is clearly seen.

It is informative to compare the BEKK model and the GARCH model with time-varying correlations in Eqs. (12.33)–(12.36) with the stochastic volatility model. First, as expected, the mean equations of the three models are essentially identical. Second, Figure 12.9 shows the time plots of the conditional variance for IBM stock return. Figure 12.9(a) is for the GARCH model, Figure 12.9(b) is from the BEKK model, and Figure 12.9(c) shows the posterior mean of the stochastic volatility model. The three models show similar volatility characteristics; they exhibit volatility clustering and indicate an increasing trend in volatility. However, the GARCH model produces higher peak volatility values and an additional peak in 1993. Third, Figure 12.10 shows the time plots of conditional variance for the S&P 500 index return. The GARCH model produces an extra volatility peak around 1993. This additional peak does not appear in the univariate analysis shown in Figure 12.6. It seems that for this particular instance the bivariate GARCH model produces a spurious volatility peak. This spurious peak is induced by its dependence on IBM returns and does not appear in the stochastic volatility model or the BEKK model. Indeed, the fitted volatilities of the S&P 500 index

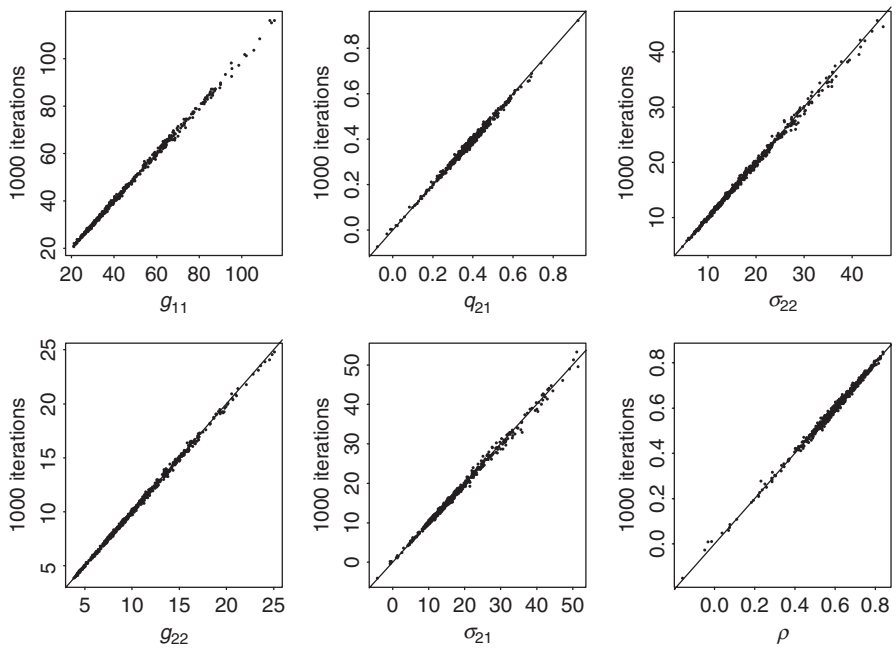

Figure 12.8 Scatterplots of posterior means of various statistics of two different Gibbs samples for bivariate stochastic volatility model for monthly log returns of IBM stock and S&P 500 index. The x axis denotes results based on $500 + 2000$ iterations and the y axis denotes results based on $500 + 1000$ iterations. Notation is defined in text.

return by the bivariate stochastic volatility model are similar to that of the univariate analysis. Fourth, Figure 12.11 shows the time plots of fitted conditional correlations. Here the three models differ substantially. The correlations of the GARCH model with Cholesky decomposition are relatively smooth and always positive with mean value 0.59 and standard deviation 0.07. The range of the correlations is (0.411,0.849). The correlations of the BEKK(1,1) model assume small negative values around 1993 and are more variable with mean 0.59, standard deviation 0.13 and range $(-0.020, 0.877)$. However, the correlations produced by the stochastic volatility model vary markedly from one month to another with mean value 0.60, standard deviation 0.14, and range $(-0.161, 0.839)$. Furthermore, the negative correlations occur in several isolated periods. The difference is understandable because $q_{21,t}$ contains the random shock u_t in the stochastic volatility model.

Remark. The Gibbs sampling estimation applies to other bivariate stochastic volatility models. The conditional posterior distributions needed require some extensions of those discussed in this section, but they are based on the same ideas. The BEKK model is estimated by using Matlab. □

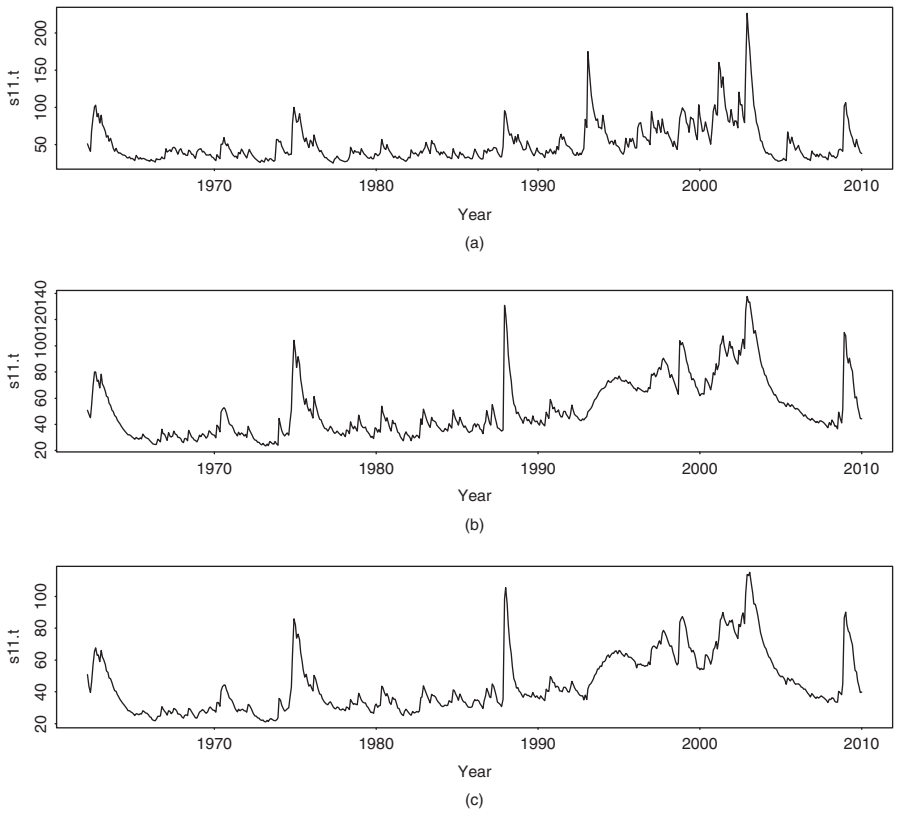

Figure 12.9 Time plots of fitted conditional variance for monthly log returns of IBM stock from 1962 to 2009: (a) GARCH model with time-varying correlations, (b) BEKK(1,1) model, and (c) bivariate stochastic volatility model estimated by Gibbs sampling with $500 + 2000$ iterations.

12.8 NEW APPROACH TO SV ESTIMATION

In this section, we discuss an alternative procedure to estimate stochastic volatility (SV) models. This approach makes use of the technique of *forward filtering and backward sampling* (FFBS) within the Kalman filter framework to improve the efficiency of Gibbs sampling. It can dramatically reduce the computing time by drawing the volatility process jointly with the help of a mixture of normal distributions. In fact, the approach can be used to estimate many stochastic diffusion models with leverage effects and jumps.

For ease in presentation, we reparameterize the univariate stochastic volatility model in Eqs. (12.20) and (12.21) as

$$r_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \sigma_0 \exp\left(\frac{z_t}{2}\right)\epsilon_t, \tag{12.40}$$

$$z_{t+1} = \alpha z_t + \eta_t, \tag{12.41}$$

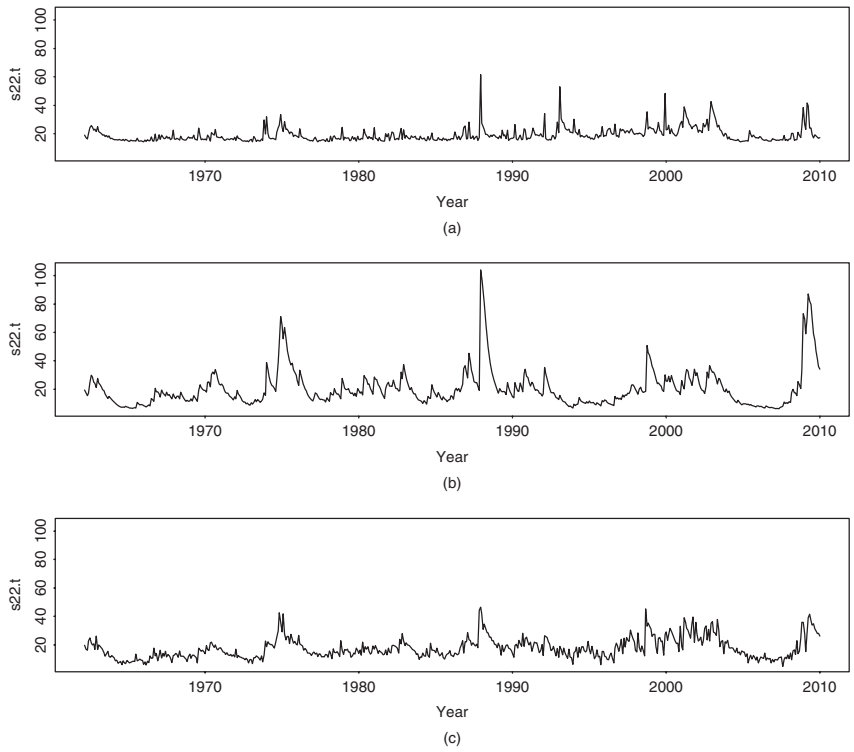

Figure 12.10 Time plots of conditional variance for monthly log returns of S&P 500 index from 1962 to 2009: (a) GARCH model with time-varying correlations, (b) BEKK(1,1) model, and (c) bivariate stochastic volatility model estimated by Gibbs sampling with $500 + 2000$ iterations.

where $\boldsymbol{x}_t = (1, x_{1t}, \ldots, x_{pt})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, $\sigma_0 > 0$, $\{z_t\}$ is a zero-mean log volatility series, and $\{\epsilon_t\}$ and $\{\eta_t\}$ are bivariate normal distributions with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho\sigma_\eta \\ \rho\sigma_\eta & \sigma_\eta^2 \end{bmatrix}.$$

The parameter ρ is the correlation between ϵ_t and η_t and represents the *leverage effect* of the asset return r_t. Typically, ρ is negative signifying that a negative return tends to increase the volatility of an asset price.

Compared with the model in Eqs. (12.22) and (12.20), we have $z_t = \ln(h_t) - \ln(\sigma_0^2)$ and $\sigma_0^2 = \exp\{E[\ln(h_t)]\}$. That is, z_t is a mean-adjusted log volatility series. This new parameterization has some nice characteristics. For example, the volatility series is $\sigma_0 \exp(z_t/2)$, which is always positive. More importantly, η_t is the innovation of z_{t+1} and is independent of z_t. This simple time shift enables us to handle the leverage effect. If one postulates $z_t = \alpha z_{t-1} + \eta_t$ for Eq. (12.41), then η_t and ϵ_t cannot be correlated because a nonzero correlation implies that z_t and ϵ_t are correlated in Eq. (12.40), which would lead to some identifiability issues.

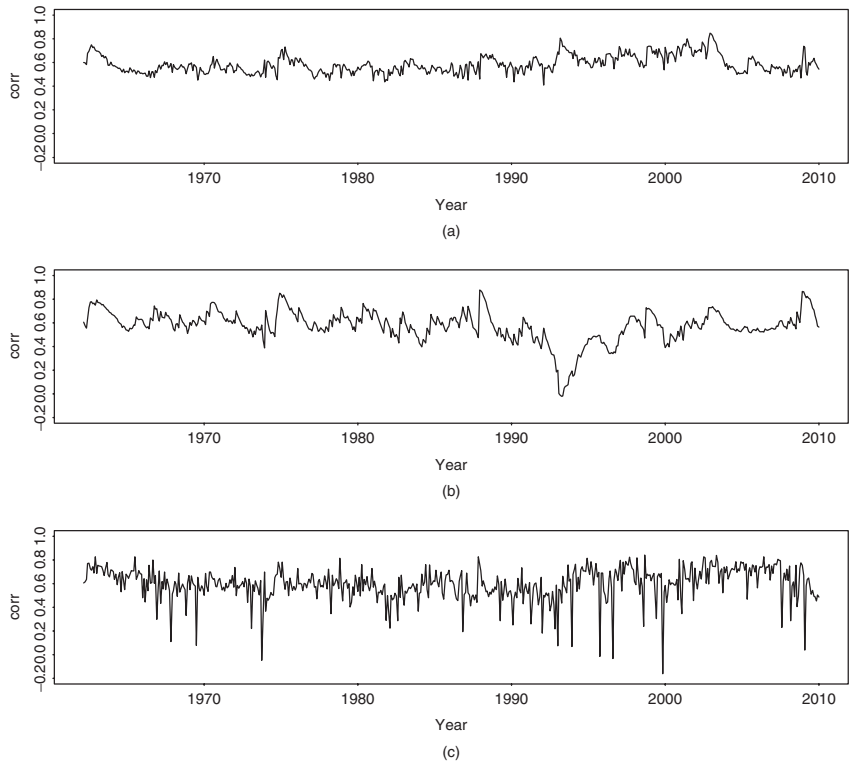

Figure 12.11 Time plots of fitted correlation coefficients between monthly log returns of IBM stock and S&P 500 index from 1962 to 2009: (a) GARCH model with time-varying correlations, (b) BEKK(1,1) model, and (c) bivariate stochastic volatility model estimated by Gibbs sampling with $500 + 2000$ iterations.

Remark. Alternatively, one can write the stochastic volatility model as

$$r_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \sigma_0 \exp\left(\frac{z_{t-1}}{2}\right)\epsilon_t,$$
$$z_t = \alpha z_{t-1} + \eta_t,$$

where $(\epsilon_t, \eta_t)'$ is a bivariate normal distribution as before. Yet another equivalent parameterization is

$$r_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \exp\left(\frac{z_{t-1}^*}{2}\right)\epsilon_t,$$
$$z_t^* = \alpha_0 + \alpha z_{t-1}^* + \eta_t,$$

where $E(z_t^*) = \alpha_0/(1-\alpha)$ is not zero. □

Parameters of the stochastic volatility model in Eqs. (12.40) and (12.41) are $\boldsymbol{\beta}$, σ_0, α, ρ, σ_η, and $\boldsymbol{z} = (z_1, \ldots, z_n)'$, where n is the sample size. For simplicity, we

assume z_1 is known. To estimate these parameters via MCMC methods, we need their conditional posterior distributions. In what follows, we discuss the needed conditional posterior distributions.

1. Given $\boldsymbol{z}$ and σ_0 and a normal prior distribution, $\boldsymbol{\beta}$ has the same conditional posterior distribution as that in Section 12.7.1 with $\sqrt{h_t}$ replaced by $\sigma_0 \exp(z_t/2)$; see Eq. (12.22).

2. Given $\boldsymbol{z}$ and σ_η^2, α is a simple AR(1) coefficient. Thus, with an approximate normal prior, the conditional posterior distribution of α is readily available; see Section 12.7.1.

3. Given $\boldsymbol{\beta}$ and $\boldsymbol{z}$, we define $v_t = (r_t - \boldsymbol{x}_t'\boldsymbol{\beta}) \exp(-z_t/2) = \sigma_0 \epsilon_t$. Thus, $\{v_t\}$ is a sequence of iid normal random variables with mean zero and variance σ_0^2. If the prior distribution of σ_0^2 is $(m\lambda)/\sigma_0^2 \sim \chi_m^2$, then the conditional posterior distribution of σ_0^2 is an inverted chi-squared distribution with $m + n$ degrees of freedom; that is,

$$\frac{m\lambda + \sum_{t=1}^{n} v_t^2}{\sigma_0^2} \sim \chi_{m+n}^2.$$

4. Given $\boldsymbol{\beta}$, σ_0, $\boldsymbol{z}$, and α, we can easily obtain the bivariate innovation $\boldsymbol{b}_t = (\epsilon_t, \eta_t)'$ for $t = 2, \ldots, n$. The likelihood function of (ρ, σ_η^2) is readily available as

$$\begin{aligned}\ell(\rho, \sigma_\eta^2) = \prod_{t=2}^{n} f(\boldsymbol{b}_t|\boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-(n-1)/2} \exp\left(-\frac{1}{2}\sum_{t=2}^{n} \boldsymbol{b}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_t\right) \\ &\propto |\boldsymbol{\Sigma}|^{-(n-1)/2} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{t=2}^{n}\boldsymbol{b}_t\boldsymbol{b}_t'\right)\right],\end{aligned}$$

where $\text{tr}(\boldsymbol{A})$ denotes trace of the matrix $\boldsymbol{A}$. However, this joint distribution is complicated because one cannot separate ρ and σ_η^2. We adopt the technique of Jacquier, Polson, and Rossi (2004) and reparameterize the covariance matrix as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho\sigma_\eta \\ \rho\sigma_\eta & \sigma_\eta^2 \end{bmatrix} = \begin{bmatrix} 1 & \varphi \\ \varphi & \omega + \varphi^2 \end{bmatrix},$$

where $\omega = \sigma_\eta^2(1 - \rho^2)$. It is easy to see that $|\boldsymbol{\Sigma}| = \omega$ and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\omega}\begin{bmatrix} \varphi^2 & -\varphi \\ -\varphi & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \equiv \frac{1}{\omega}\boldsymbol{S} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

where $\boldsymbol{S}$ contains φ only. Let $\boldsymbol{e} = (\epsilon_2, \ldots, \epsilon_n)'$ and $\boldsymbol{\eta} = (\eta_2, \ldots, \eta_n)'$ be the innovations of the model in Eqs. (12.40) and (12.41). The likelihood function then becomes (keeping terms related to parameters only)

$$\ell(\varphi, \omega) \propto \omega^{-(n-1)/2} \exp\left[-\frac{1}{2\omega}\text{tr}(\boldsymbol{S}\boldsymbol{R})\right],$$

where $\boldsymbol{R} = \sum_{t=2}^{n} \boldsymbol{b}_t \boldsymbol{b}_t' = (\boldsymbol{e}, \boldsymbol{\eta})'(\boldsymbol{e}, \boldsymbol{\eta})$, which is the 2×2 cross-product matrix of the innovations. For simplicity, we use conjugate priors such that ω is inverse gamma (IG) with hyperparameters $(\gamma_0/2, \gamma_1/2)$; that is, $\omega \sim \text{IG}(\gamma_0/2, \gamma_1/2)$, and $\varphi|\omega \sim N(0, \omega/2)$. Then, after some algebraic manipulation, the joint posterior distribution of (φ, ω) can be decomposed into a normal and an inverse gamma distribution. Specifically,

$$\varphi \sim N\left(\tilde{\varphi},\ \frac{\omega}{(2 + \boldsymbol{e}'\boldsymbol{e})}\right),$$

where $\tilde{\varphi} = \boldsymbol{e}'\boldsymbol{\eta}/(2 + \boldsymbol{e}'\boldsymbol{e})$, and

$$\omega \sim \text{IG}\left[\frac{1}{2}\,(n + 1 + \gamma_0)\,,\ \frac{1}{2}\left(\gamma_1 + \boldsymbol{\eta}'\boldsymbol{\eta} - \frac{(\boldsymbol{e}'\boldsymbol{\eta})^2}{2 + \boldsymbol{e}'\boldsymbol{e}}\right)\right].$$

In Gibbs sampling, once φ and ω are available, we can obtain ρ and σ_η^2 easily because $\sigma_\eta^2 = \omega + \varphi^2$ and $\rho = \varphi/\sigma_\eta$. Note that the probability density function of an IG(α, β) random variable ω is

$$f(\omega|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{-(\alpha+1)} \exp\left(-\frac{\beta}{\omega}\right), \qquad \text{for } \omega > 0,$$

where $\alpha > 2$ and $\beta > 0$.

5. Finally, we consider the joint distribution of the log volatility $\boldsymbol{z}$ given the data and other parameters. From Eq. (12.40), we have

$$\frac{(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{\sigma_0^2} = \exp(z_t)\epsilon_t^2.$$

Therefore, letting $y_t = \ln[(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2/\sigma_0^2]$, we obtain

$$y_t = z_t + \epsilon_t^*, \tag{12.42}$$

where $\epsilon_t^* = \ln(\epsilon_t^2)$. Since $\epsilon_t^2 \sim \chi_1^2$, ϵ_t^* is not normally distributed. Treating Eq. (12.42) as an observation equation and Eq. (12.40) as the state equation, we have the form of a state-space model except that ϵ_t^* is not Gaussian; see Eqs. (11.26) and (11.27). To overcome the difficulty associated with nonnormality, Kim, Shephard, and Chib (1998) use a mixture of seven normal distributions to approximate the distribution of ϵ_t^*. Specifically, we have

$$f(\epsilon_t^*) \approx \sum_{i=1}^{7} p_i N(\mu_i, \overline{\omega}_i^2),$$

where p_i, μ_i, and ϖ_i^2 are given in Table 12.3. See also Chib, Nardari, and Shephard (2002).

TABLE 12.3 Seven Components of Normal Distributions

Component i	Probability p_i	Mean μ_i	var. $\overline{\omega}_i^2$
1	0.00730	−11.4004	5.7960
2	0.10556	−5.2432	2.6137
3	0.00002	−9.8373	5.1795
4	0.04395	1.5075	0.1674
5	0.34001	−0.6510	0.6401
6	0.24566	0.5248	0.3402
7	0.25750	−2.3586	1.2626

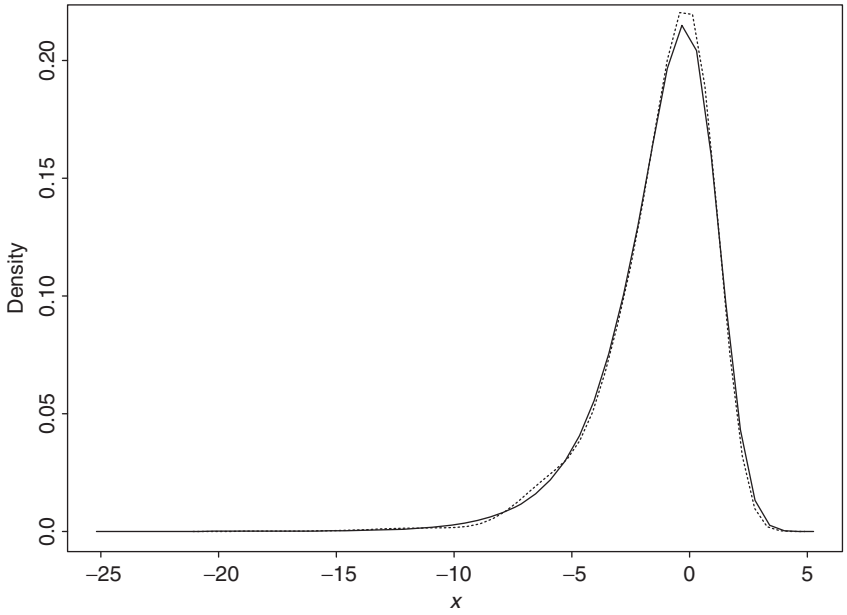

Figure 12.12 Density functions of $\log(\chi_1^2)$, solid line, and that of a mixture of seven normal distributions, dashed line. Results are based on 100,000 observations.

To demonstrate the adequacy of the approximation, Figure 12.12 shows the density function of ϵ_t^* (solid line) and that of the mixture of seven normals (dashed line) in Table 12.3. These densities are obtained using simulations with 100,000 observations. From the plot, the approximation by the mixture of seven normals is very good.

Why is it important to have a Gaussian state-space model? The answer is that such a Gaussian model enables us to draw the log volatility series $\boldsymbol{z}$ jointly and efficiently. To see this, consider the following special Gaussian state-space model,

where η_t and e_t are uncorrelated (i.e., no leverage effects):

$$z_{t+1} = \alpha z_t + \eta_t, \qquad \eta_t \sim_{\text{iid}} N(0, \sigma_\eta^2), \tag{12.43}$$

$$y_t = c_t + z_t + e_t, \qquad e_t \sim_{\text{ind.}} N(0, H_t), \tag{12.44}$$

where, as will be seen later, (c_t, H_t) assumes the value (μ_i, ϖ_i^2) of Table 12.3 for some i. For this special state-space model, we have the Kalman filter algorithm

$$\begin{aligned} v_t &= y_t - y_{t|t-1} = y_t - c_t - z_{t|t-1}, \\ V_t &= \Sigma_{t|t-1} + H_t, \\ z_{t|t} &= z_{t|t-1} + \Sigma_{t|t-1} V_t^{-1} v_t, \\ \Sigma_{t|t} &= \Sigma_{t|t-1} - \Sigma_{t|t-1} V_t^{-1} \Sigma_{t|t-1}, \\ z_{t+1|t} &= \alpha z_{t|t}, \\ \Sigma_{t+1|t} &= \alpha^2 \Sigma_{t|t} + \sigma_\eta^2, \end{aligned} \tag{12.45}$$

where $V_t = \text{Var}(v_t)$ is the variance of the 1-step-ahead prediction error v_t of y_t given $F_{t-1} = (y_1, \ldots, y_{t-1})$, and $z_{j|i}$ and $\Sigma_{j|i}$ are, respectively, the conditional expectation and variance of the state variable z_j given F_i. See the Kalman filter discussion of Chapter 11.

Forward Filtering and Backward Sampling

Let $p(\mathbf{z}|F_n)$ be the joint conditional posterior distribution of $\mathbf{z}$ given the return data and other parameters, where for simplicity the parameters are omitted from the condition set. We can partition the distribution as

$$\begin{aligned} p(\mathbf{z}|F_n) &= P(z_2, z_3, \ldots, z_n|F_n) \\ &= p(z_n|F_n)p(z_{n-1}|z_n, F_n)p(z_{n-2}|z_{n-1}, z_n, F_n)\cdots p(z_2|z_3, \ldots, z_n, F_n) \\ &= p(z_n|F_n)p(z_{n-1}|z_n, F_n)p(z_{n-2}|z_{n-1}, F_n)\cdots p(z_2|z_3, F_n), \end{aligned} \tag{12.46}$$

where the last equality holds because z_t in Eq. (12.43) is a Markov process so that conditioned on z_{t+1}, z_t is independent of z_{t+j} for $j > 1$.

From the Kalman filter in Eq. (12.45), we obtain that $p(z_n|F_n)$ is normal with mean $z_{n|n}$ and variance $\Sigma_{n|n}$. Next, consider the second term $p(z_{n-1}|z_n, F_n)$ of Eq. (12.46). We have

$$p(z_{n-1}|z_n, F_n) = p(z_{n-1}|z_n, F_{n-1}, y_n) = p(z_{n-1}|z_n, F_{n-1}, v_n), \tag{12.47}$$

where $v_n = y_n - y_{n|n-1}$ is the 1-step-ahead prediction error of y_n. From the state-space model in Eqs. (12.43) and (12.44), z_{n-1} is independent of v_n. Therefore,

$$p(z_{n-1}|z_n, F_n) = p(z_{n-1}|z_n, F_{n-1}). \tag{12.48}$$

This is an important property because it implies that we can derive the posterior distribution $p(z_{n-1}|z_n, F_n)$ from the joint distribution of (z_{n-1}, z_n) given F_{n-1} via Theorem 11.1. First, the joint distribution is bivariate normal under the Gaussian assumption. Second, the conditional mean and covariance matrix of (z_{n-1}, z_n) given F_{n-1} are readily available from the Kalman filter algorithm in Eq. (12.45). Specifically, we have

$$\begin{bmatrix} z_{n-1} \\ z_n \end{bmatrix}_{F_{n-1}} \sim N\left(\begin{bmatrix} z_{n-1|n-1} \\ z_{n|n-1} \end{bmatrix}, \begin{bmatrix} \Sigma_{n-1|n-1} & \alpha\Sigma_{n-1|n-1} \\ \alpha\Sigma_{n-1|n-1} & \Sigma_{n|n-1} \end{bmatrix}\right), \tag{12.49}$$

where the covariance is obtained by (i) multiplying z_{n-1} by Eq. (12.43) and (ii) taking conditional expectation. Note that all quantities involved in Eq. (12.49) are available from the Kalman filter. Consequently, by Theorem 11.1, we have

$$p(z_{n-1}|z_n, F_n) \sim N(\mu^*_{n-1}, \Sigma^*_{n-1}), \tag{12.50}$$

where

$$\begin{aligned} \mu^*_{n-1} &= z_{n-1|n-1} + \alpha\Sigma_{n-1|n-1}\Sigma^{-1}_{n|n-1}(z_n - z_{n|n-1}), \\ \Sigma^*_{n-1} &= \Sigma_{n-1|n-1} - \alpha^2\Sigma^2_{n-1|n-1}\Sigma^{-1}_{n|n-1}. \end{aligned}$$

Next, for the conditional posterior distribution $p(z_{n-2}|z_{n-1}, F_n)$, we have

$$\begin{aligned} p(z_{n-2}|z_{n-1}, F_n) &= p(z_{n-2}|z_{n-1}, F_{n-2}, y_{n-1}, y_n) \\ &= p(z_{n-2}|z_{n-1}, F_{n-2}, v_{n-1}, v_n) \\ &= p(z_{n-2}|z_{n-1}, F_{n-2}). \end{aligned}$$

Consequently, we can obtain $p(z_{n-2}|z_{n-1}, F_n)$ from the bivariate normal distribution of $p(z_{n-2}, z_{n-1}|F_{n-2})$ as before. In general, we have

$$p(z_t|z_{t+1}, F_n) = p(z_t|z_{t+1}, F_t), \qquad \text{for } 1 < t < n.$$

Furthermore, from the Kalman filter, $p(z_t, z_{t+1}|F_t)$ is bivariate normal as

$$\begin{bmatrix} z_t \\ z_{t+1} \end{bmatrix}_{F_t} \sim N\left(\begin{bmatrix} z_{t|t} \\ z_{t+1|t} \end{bmatrix}, \begin{bmatrix} \Sigma_{t|t} & \alpha\Sigma_{t|t} \\ \alpha\Sigma_{t|t} & \Sigma_{t+1|t} \end{bmatrix}\right). \tag{12.51}$$

Consequently,

$$p(z_t|z_{t+1}, F_t) \sim N(\mu^*_t, \Sigma^*_t),$$

where

$$\begin{aligned} \mu^*_t &= z_{t|t} + \alpha\Sigma_{t|t}\Sigma^{-1}_{t+1|t}(z_{t+1} - z_{t+1|t}), \\ \Sigma^*_t &= \Sigma_{t|t} - \alpha^2\Sigma^2_{t|t}\Sigma^{-1}_{t+1|t}. \end{aligned}$$

The prior derivation implies that we can draw the volatility series $\boldsymbol{z}$ jointly by a recursive method using quantities readily available from the Kalman filter algorithm. That is, given the initial values $z_{1|0}$ and $\Sigma_{1|0}$, one uses the Kalman filter in Eq. (12.45) to process the return data forward, then applies the recursive backward method to draw a realization of the volatility series $\boldsymbol{z}$. This scheme is referred to as *forward filtering and backward sampling* (FFBS); see Carter and Kohn (1994) and Frühwirth-Schnatter (1994). Because the volatility $\{z_t\}$ is serially correlated, drawing the series jointly is more efficient.

Remark. The FFBS procedure applies to general linear Gaussian state-space models. The main idea is to make use of the Markov property of the model and the structure of the state transition equation so that

$$p(\boldsymbol{S}_t|\boldsymbol{S}_{t+1}, F_n) = p(\boldsymbol{S}_t|\boldsymbol{S}_{t+1}, F_t, v_{t+1}, \ldots, v_n) = p(\boldsymbol{S}_t|\boldsymbol{S}_{t+1}, F_t),$$

where $\boldsymbol{S}_t$ denotes the state variable at time t and v_j is the 1-step-ahead prediction error. This identity enables us to apply Theorem 11.1 to derive a recursive method to draw the state vectors jointly. □

Return to the estimation of the SV model. As in Eq. (12.42), let $y_t = \ln[(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})^2/\sigma_0^2]$. To implement FFBS, one must determine c_t and H_t of Eq. (12.44) so that the mixture of normals provides a good approximation to the distribution of ϵ_t^*. To this end, we augment the model with a series of independent indicator variables $\{I_t\}$, where I_t assumes a value in $\{1, \ldots, 7\}$ such that $P(I_t = i) = p_{it}$ with $\sum_{i=1}^{7} p_{it} = 1$ for each t. In practice, conditioned on $\{z_t\}$, we can determine c_t and H_t as follows. Let

$$q_{it} = \Phi[(y_t - z_t - \mu_i)/\varpi_i], \qquad \text{for } i = 1, \ldots, 7,$$

where μ_i and ϖ_i are the mean and standard error of the normal distributions given in Table 12.3 and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal random variable. These probabilities q_{it} are the likelihood function of I_t given y_t and z_t. The probabilities p_i of Table 12.3 form a prior distribution of I_t. Therefore, the posterior distribution of I_t is

$$p_{it} = \frac{p_i q_{it}}{\sum_{j=1}^{7} p_j q_{jt}}, \qquad i = 1, \ldots, 7.$$

We can draw a realization of I_t using this posterior distribution. If the random draw is $I_t = j$, then we define $c_t = \mu_j$ and $H_t = \varpi_j^2$. In summary, conditioned on the return data and other parameters of the model, we employ the approximate linear Gaussian state-space model in Eqs. (12.43) and (12.44) to draw jointly the log volatility series $\boldsymbol{z}$. It turns out that the resulting Gibbs sampling is efficient in estimating univariate stochastic volatility models.

On the other hand, the square transformation involved in Eq. (12.42) fails to retain the correlation between η_t and ϵ_t if it exists, making the approximate state-space model in Eqs. (12.43) and (12.44) incapable of estimating the leverage effect. To overcome this inadequacy, Artigas and Tsay (2004) propose using a time-varying state-space model that maintains the leverage effect. Specifically, when $\rho \neq 0$, we have

$$\eta_t = \rho\sigma_\eta \epsilon_t + \eta_t^*,$$

where η_t^* is a normal random variable independent of ϵ_t and $\text{Var}(\eta_t^*) = \sigma_\eta^2(1-\rho^2)$. The state transition equation of Eq. (12.43) then becomes

$$z_{t+1} = \alpha z_t + \rho\sigma_\eta \epsilon_t + \eta_t^*.$$

Substituting $\epsilon_t = (1/\sigma_0)(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})\exp(-z_t/2)$, we obtain

$$\begin{aligned} z_{t+1} &= \alpha z_t + \frac{\rho\sigma_\eta(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})}{\sigma_0}\exp\left(\frac{-z_t}{2}\right) + \eta_t^* \\ &= G(z_t) + \eta_t^* \end{aligned} \tag{12.52}$$

where $G(z_t) = \alpha z_t + \rho\sigma_\eta(r_t - \boldsymbol{x}_t'\boldsymbol{\beta})\exp(-z_t/2)/\sigma_0$. This is a nonlinear transition equation for the state variable z_t. The Kalman filter in Eq. (12.45) is no longer applicable. To overcome this difficulty, Artigas and Tsay (2004) use a time-varying linear Kalman filter to approximate the system. Specifically, the last two equations of Eq. (12.45) are modified as

$$\begin{aligned} z_{t+1|t} &= G(z_{t|t}), \\ \Sigma_{t+1|t} &= g(z_{t|t})^2\Sigma_{t|t} + \sigma_\eta^2(1-\rho^2), \end{aligned} \tag{12.53}$$

where $g(z_{t|t}) = \partial G(x)/\partial x|_{x=z_{t|t}}$ is the first-order derivative of $G(z_t)$ evaluated at the smoothed state $z_{t|t}$.

Example 12.5. To demonstrate the FFBS procedure, we consider the monthly log returns of the S&P 500 index from January 1962 to November 2004 for 515 observations. This is a subseries of the data used in Example 12.3. See Figure 12.4 for time plots of the index and its log return. We consider two stochastic volatility models in the form:

$$\begin{aligned} r_t &= \mu + \sigma_o \exp(z_t/2)\epsilon_t, \qquad \epsilon_t \sim_{\text{iid}} N(0,1), \\ z_{t+1} &= \alpha z_t + \eta_t, \quad \eta_t \sim_{\text{iid}} N(0, \sigma_\eta^2). \end{aligned} \tag{12.54}$$

In model 1, $\{\epsilon_t\}$ and $\{\eta_t\}$ are two independent Gaussian white noise series. That is, there is no leverage effect in the model. In model 2, we assume that $\text{corr}(\epsilon_t, e_t) = \rho$, which denotes the leverage effect.

TABLE 12.4 Estimation of Stochastic Volatility Model in Eq. (12.54) for Monthly Log Returns of S&P 500 Index from January 1962 to November 2004 Using Gibbs Sampling with FFBS Algorithm[a]

Parameter	μ	σ_o	α	σ_η	ρ
With Leverage Effect					
Estimate	0.0081	0.0764	−0.0616	2.5639	−0.3892
Standard error	0.0274	0.0255	0.1186	0.3924	0.0292
Without Leverage Effect					
Estimate	0.0080	0.0775	−0.0613	2.5827	
Standard error	0.0279	0.0266	0.1164	0.3783	

[a]The results are based on 2000+8000 iterations with the first 2000 iterations as burn-ins.

We estimate the models via the FFBS procedure using a program written in Matlab. The Gibbs sampling was run for 2000+8000 iterations with the first 2000 iterations as burn-ins. Table 12.4 gives the posterior means and standard errors of the parameter estimates. In particular, we have $\hat{\rho} = -0.39$, which is close to the value commonly seen in the literature. Figure 12.13 shows the time plots of the posterior means of the estimated volatility. As expected, the two volatility series are very close. Compared with the results of Example 12.3, which uses a shorter

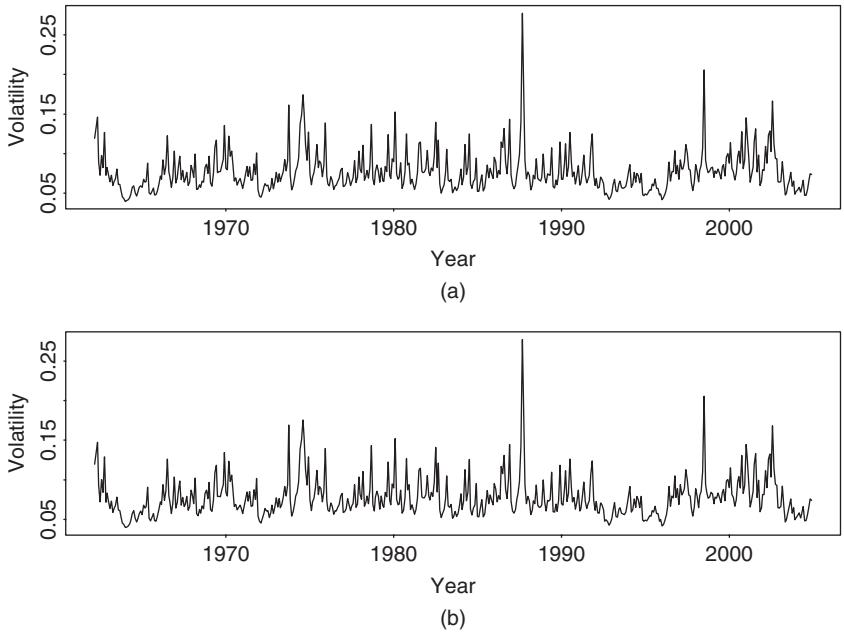

Figure 12.13 Estimated volatility of monthly log returns of S&P 500 index from January 1962 to November 2004 using stochastic volatility models: (a) with leverage effect and (b) without leverage effect.

series, the estimated volatility series exhibit similar patterns and are in the same magnitude. Note that the volatility shown in Figure 12.6 is conditional variance of percentage log returns whereas the volatility in Figure 12.13 is the conditional standard error of log returns.

12.9 MARKOV SWITCHING MODELS

The Markov switching model is another econometric model for which MCMC methods enjoy many advantages over the traditional likelihood method. McCulloch and Tsay (1994b) discuss a Gibbs sampling procedure to estimate such a model when the volatility in each state is constant over time. These authors applied the procedure to estimate a Markov switching model with different dynamics and mean levels for different states to the quarterly growth rate of U.S. real gross national product, seasonally adjusted, and obtained some interesting results. For instance, the dynamics of the growth rate are significantly different between periods of economic "contraction" and "expansion." Since this chapter is concerned with asset returns, we focus on models with volatility switching.

Suppose that an asset return r_t follows a simple two-state Markov switching model with different risk premiums and different GARCH dynamics:

$$r_t = \begin{cases} \beta_1\sqrt{h_t} + \sqrt{h_t}\epsilon_t, & h_t = \alpha_{10} + \alpha_{11}h_{t-1} + \alpha_{12}a_{t-1}^2 \quad \text{if } s_t = 1, \\ \beta_2\sqrt{h_t} + \sqrt{h_t}\epsilon_t, & h_t = \alpha_{20} + \alpha_{21}h_{t-1} + \alpha_{22}a_{t-1}^2 \quad \text{if } s_t = 2, \end{cases} \tag{12.55}$$

where $a_t = \sqrt{h_t}\epsilon_t$, $\{\epsilon_t\}$ is a sequence of Gaussian white noises with mean zero and variance 1, and the parameters α_{ij} satisfy some regularity conditions so that the unconditional variance of a_t exists. The probability transition from one state to another is governed by

$$P(s_t = 2|s_{t-1} = 1) = e_1, \qquad P(s_t = 1|s_{t-1} = 2) = e_2, \tag{12.56}$$

where $0 < e_i < 1$. A small e_i means that the return series has a tendency to stay in the ith state with expected duration $1/e_i$. For the model in Eq. (12.55) to be identifiable, we assume that $\beta_2 > \beta_1$ so that state 2 is associated with higher risk premium. This is not a critical restriction because it is used to achieve uniqueness in labeling the states. A special case of the model results if $\alpha_{1j} = \alpha_{2j}$ for all j so that the model assumes a GARCH model for all states. However, if $\beta_i\sqrt{h_t}$ is replaced by β_i, then model (12.55) reduces to a simple Markov switching GARCH model.

Model (12.55) is a Markov switching GARCH-M model. For simplicity, we assume that the initial volatility h_1 is given with value equal to the sample variance of r_t. A more sophisticated analysis is to treat h_1 as a parameter and estimate it jointly with other parameters. We expect the effect of fixing h_1 will be negligible in most applications, especially when the sample size is large. The "traditional"

parameters of the Markov switching GARCH-M model are $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $\boldsymbol{\alpha}_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})'$ for $i = 1$ and 2, and the transition probabilities $\boldsymbol{e} = (e_1, e_2)'$. The state vector $\boldsymbol{S} = (s_1, s_2, \ldots, s_n)'$ contains the augmented parameters. The volatility vector $\boldsymbol{H} = (h_2, \ldots, h_n)'$ can be computed recursively if h_1, $\boldsymbol{\alpha}_i$, and the state vector $\boldsymbol{S}$ are given.

Dependence of the return on volatility in model (12.55) implies that the return is also serially correlated. The model thus has some predictability in the return. However, states of the future returns are unknown and a prediction produced by the model is necessarily a mixture of those over possible state configurations. This often results in high uncertainty in point prediction of future returns.

Turn to estimation. The likelihood function of model (12.55) is complicated as it is a mixture over all possible state configurations. Yet the Gibbs sampling approach only requires the following conditional posterior distributions:

$$f(\boldsymbol{\beta}|\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{H}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), \qquad f(\boldsymbol{\alpha}_i|\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{H}, \boldsymbol{\alpha}_{j\neq i}),$$
$$P(\boldsymbol{S}|\boldsymbol{R}, h_1, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), \qquad f(e_i|\boldsymbol{S}), \quad i = 1, 2,$$

where $\boldsymbol{R}$ is the collection of observed returns. For simplicity, we use conjugate prior distributions discussed in Section 12.3—that is,

$$\beta_i \sim N(\beta_{io}, \sigma_{io}^2), \qquad e_i \sim \text{Beta}(\gamma_{i1}, \gamma_{i2}).$$

The prior distribution of parameter α_{ij} is uniform over a properly specified interval. Since α_{ij} is a nonlinear parameter of the likelihood function, we use the Griddy Gibbs to draw its random realizations. A uniform prior distribution simplifies the computation involved. Details of the prior conditional posterior distributions follow:

1. The posterior distribution of β_i only depends on the data in state i. Define

$$r_{it} = \begin{cases} r_t/\sqrt{h_t} & \text{if } s_t = i, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$r_{it} = \beta_i + \epsilon_t, \quad \text{for} \quad s_t = i.$$

Therefore, information of the data on β_i is contained in the sample mean of r_{it}. Let $\bar{r}_i = (\sum_{s_t=i} r_{it})/n_i$, where the summation is over all data points in state i and n_i is the number of data points in state i. Then the conditional posterior distribution of β_i is normal with mean β_i^* and variance σ_{i*}^2, where

$$\frac{1}{\sigma_{i*}^2} = n_i + \frac{1}{\sigma_{io}^2}, \qquad \beta_i^* = \sigma_{i*}^2 \left(n_i \bar{r}_i + \beta_{io}/\sigma_{io}^2\right), \qquad i = 1, 2.$$

2. Next, the parameters α_{ij} can be drawn one by one using the Griddy Gibbs method. Given h_1, $\boldsymbol{S}$, $\boldsymbol{\alpha}_{v\neq i}$, and α_{iv} with $v \neq j$, the conditional posterior distribution function of α_{ij} does not correspond to a well-known distribution, but it can be evaluated easily as

$$f(\alpha_{ij}|.) \propto -\frac{1}{2}\left[\ln h_t + \frac{(r_t - \beta_i\sqrt{h_t})^2}{h_t}\right], \qquad \text{if} \quad s_t = i,$$

where h_t contains α_{ij}. We evaluate this function at a grid of points for α_{ij} over a properly specified interval. For example, $0 \leq \alpha_{11} < 1 - \alpha_{12}$.

3. The conditional posterior distribution of e_i only involves $\boldsymbol{S}$. Let ℓ_1 be the number of switches from state 1 to state 2 and ℓ_2 be the number of switches from state 2 to state 1 in $\boldsymbol{S}$. Also, let n_i be the number of data points in state i. Then by Result 12.3 of conjugate prior distributions, the posterior distribution of e_i is Beta$(\gamma_{i1} + \ell_i, \gamma_{i2} + n_i - \ell_i)$.

4. Finally, elements of $\boldsymbol{S}$ can be drawn one by one. Let $\boldsymbol{S}_{-j}$ be the vector obtained by removing s_j from $\boldsymbol{S}$. Given $\boldsymbol{S}_{-j}$ and other information, s_j can assume two possibilities (i.e., $s_j = 1$ or $s_j = 2$), and its conditional posterior distribution is

$$P(s_j|.) \propto \prod_{t=j}^{n} f(a_t|\boldsymbol{H})P(s_j|\boldsymbol{S}_{-j}).$$

The probability

$$P(s_j = i|\boldsymbol{S}_{-j}) = P(s_j = i|s_{j-1}, s_{j+1}), \qquad i = 1, 2$$

can be computed by the Markov transition probabilities in Eq. (12.56). In addition, assuming $s_j = i$, one can compute h_t for $t \geq j$ recursively. The relevant likelihood function, denoted by $L(s_j)$, is given by

$$L(s_j = i) \equiv \prod_{t=j}^{n} f(a_t|\boldsymbol{H}) \propto \exp(f_{ji}), \qquad f_{ji} = \sum_{t=j}^{n} -\frac{1}{2}\left[\ln(h_t) + \frac{a_t^2}{h_t}\right],$$

for $i = 1$ and 2, where $a_t = r_t - \beta_1\sqrt{h_t}$ if $s_t = 1$ and $a_t = r_t - \beta_2\sqrt{h_t}$ otherwise. Consequently, the conditional posterior probability of $s_j = 1$ is

$$P(s_j = 1|\cdot) = \frac{P(s_j = 1|s_{j-1}, s_{j+1})L(s_j = 1)}{P(s_j = 1|s_{j-1}, s_{j+1})L(s_j = 1) + P(s_j = 2|s_{j-1}, s_{j+1})L(s_j = 2)}.$$

The state s_j can then be drawn easily using a uniform distribution on the unit interval [0, 1].

Remark. Since s_j and s_{j+1} are highly correlated when e_1 and e_2 are small, it is more efficient to draw several s_j jointly. However, the computation involved in enumerating the possible state configurations increases quickly with the number of states drawn jointly. □

Example 12.6. In this example, we consider the monthly log stock returns of General Electric Company from January 1926 to December 1999 for 888 observations. The returns are in percentages and shown in Figure 12.14(a). For comparison purposes, we start with a GARCH-M model for the series and obtain

$$\begin{aligned} r_t &= 0.182\sqrt{h_t} + a_t, \qquad a_t = \sqrt{h_t}\epsilon_t, \\ h_t &= 0.546 + 1.740h_{t-1} - 0.775h_{t-2} + 0.025a_{t-1}^2, \end{aligned} \tag{12.57}$$

where r_t is the monthly log return and $\{\epsilon_t\}$ is a sequence of independent Gaussian white noises with mean zero and variance 1. All parameter estimates are highly significant with p values less than 0.0006. The Ljung–Box statistics of the standardized residuals and their squared series fail to suggest any model inadequacy. It

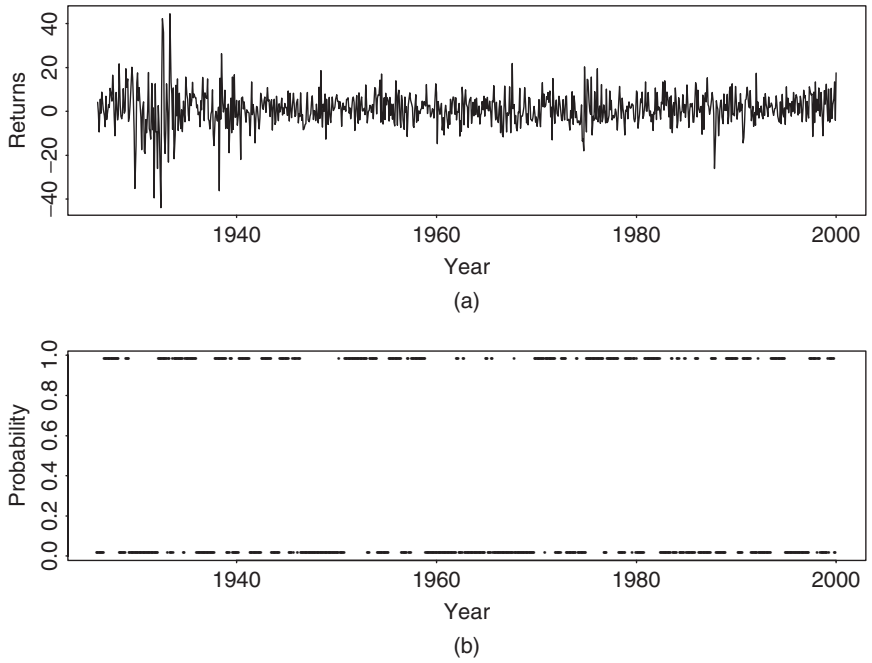

Figure 12.14 (a) Time plot of monthly log returns, in percentages, of GE stock from 1926 to 1999. (b) Time plot of the posterior probability of being in state 2 based on results of last 2000 iterations of Gibbs sampling with 5000 + 2000 total iterations. Model used is two-state Markov switching GARCH-M model.

is reassuring to see that the risk premium is positive and significant. The GARCH model in Eq. (12.57) can be written as

$$(1 - 1.765B + 0.775B^2)a_t^2 = 0.546 + (1 - 0.025B)\eta_t,$$

where $\eta_t = a_t^2 - h_t$ and B is the back-shift operator such that $Ba_t^2 = a_{t-1}^2$. As discussed in Chapter 3, the prior equation can be regarded as an ARMA(2,1) model with nonhomogeneous innovations for the squared series a_t^2. The AR polynomial can be factorized as $(1 - 0.945B)(1 - 0.820B)$, indicating two real characteristic roots with magnitudes less than 1. Consequently, the unconditional variance of r_t is finite and equal to $0.546/(1 - 1.765 + 0.775) \approx 49.64$.

Turn to Markov switching models. We use the following prior distributions:

$$\beta_1 \sim N(0.3, 0.09), \qquad \beta_2 \sim N(1.3, 0.09), \qquad \epsilon_i \sim \text{Beta}(5, 95).$$

The initial parameter values used are (a) $e_i = 0.1$, (b) s_1 is a Bernoulli trial with equal probabilities and s_t is generated sequentially using the initial transition probabilities, and (c) $\boldsymbol{\alpha}_1 = (1.0, 0.6, 0.2)'$ and $\boldsymbol{\alpha}_2 = (2, 0.7, 0.1)'$. Gibbs samples of α_{ij} are drawn using the Griddy Gibbs with 400 grid points, equally spaced over the following ranges: $\alpha_{i0} \in [0, 6.0]$, $\alpha_{i1} \in [0, 1]$, and $\alpha_{i2} \in [0, 0.5]$. In addition, we implement the constraints $\alpha_{i1} + \alpha_{i2} < 1$ for $i = 1, 2$. The Gibbs sampler is run for $5000 + 2000$ iterations, but only results of the last 2000 iterations are used to make inference.

Table 12.5 shows the posterior means and standard deviations of parameters of the Markov switching GARCH-M model in Eq. (12.55). In particular, it also contains some statistics showing the difference between the two states such as $\theta = \beta_2 - \beta_1$. The difference between the risk premiums is statistically significant at the 5% level. The differences in posterior means of the volatility parameters between the two states appear to be insignificant. Yet the posterior distributions of volatility parameters show some different characteristics. Figures 12.15 and 12.16 show the histograms of all parameters in the Markov switching GARCH-M model. They exhibit some differences between the two states. Figure 12.17 shows the time plot of the persistent parameter $\alpha_{i1} + \alpha_{i2}$ for the two states. It shows that the persistent parameter of state 1 reaches the boundary 1.0 frequently, but that of state 2 does not. The expected durations of the two states are about 11 and 9 months, respectively. Figure 12.14(b) shows the posterior probability of being in state 2 for each observation.

Finally, we compare the fitted volatility series of the simple GARCH-M model in Eq. (12.57) and the Markov switching GARCH-M model in Eq. (12.55). The two fitted volatility series (Figure 12.18) show similar patterns and are consistent with the behavior of the squared log returns. The simple GARCH-M model produces a smoother volatility series with lower estimated volatilities.

TABLE 12.5 Fitted Markov Switching GARCH-M Model for Monthly Log Returns of GE Stock from January 1926 to December 1999[a]

State 1					
Parameter	β_1	e_1	α_{10}	α_{11}	α_{12}
Posterior mean	0.111	0.089	2.070	0.844	0.033
Posterior standard error	0.043	0.012	1.001	0.038	0.033
State 2					
Parameter	β_2	e_2	α_{20}	α_{21}	α_{22}
Posterior mean	0.247	0.112	2.740	0.869	0.068
Posterior standard Error	0.050	0.014	1.073	0.031	0.024
Difference Between States					
Parameter	$\beta_2 - \beta_1$	$e_2 - e_1$	$\alpha_{20} - \alpha_{10}$	$\alpha_{21} - \alpha_{11}$	$\alpha_{22} - \alpha_{12}$
Posterior mean	0.135	0.023	0.670	0.026	−0.064
Posterior standard error	0.063	0.019	1.608	0.050	0.043

[a]The numbers shown are the posterior means and standard deviations based on a Gibbs sampling with 5000 + 2000 iterations. Results of the first 5000 iterations are discarded. The prior distributions and initial parameter estimates are given in the text.

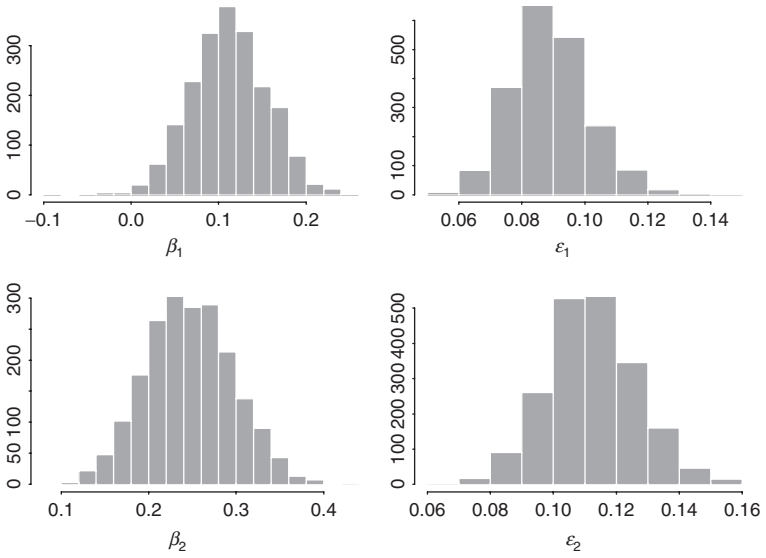

Figure 12.15 Histograms of risk premium and transition probabilities of a two-state Markov switching GARCH-M model for monthly log returns of GE stock from 1926 to 1999. Results based on last 2000 iterations of Gibbs sampling with 5000 + 2000 total iterations.

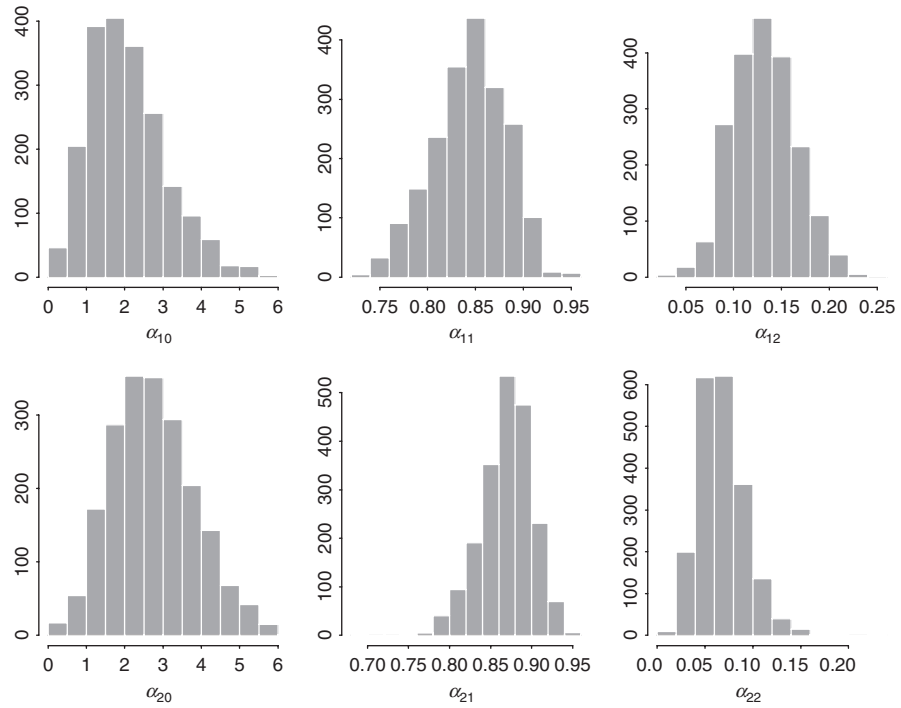

Figure 12.16 Histograms of volatility parameters of two-state Markov switching GARCH-M model for monthly log returns of GE stock from 1926 to 1999. Results based on last 2000 iterations of Gibbs sampling with 5000 + 2000 total iterations.

12.10 FORECASTING

Forecasting under the MCMC framework can be done easily. The procedure is simply to use the fitted model in each Gibbs iteration to generate samples for the forecasting period. In a sense, forecasting here is done by using the fitted model to simulate realizations for the forecasting period. We use the univariate stochastic volatility model to illustrate the procedure; forecasts of other models can be obtained by the same method.

Consider the stochastic volatility model in Eqs. (12.20) and (12.21). Suppose that there are n returns available and we are interested in predicting the return r_{n+i} and volatility h_{n+i} for $i = 1, \ldots, \ell$, where $\ell > 0$. Assume that the explanatory variables x_{jt} in Eq. (12.20) are either available or can be predicted sequentially during the forecasting period. Recall that estimation of the model under the MCMC framework is done by Gibbs sampling, which draws parameter values from their conditional posterior distributions iteratively. Denote the parameters by $\boldsymbol{\beta}_j = (\beta_{0,j}, \ldots, \beta_{p,j})'$, $\boldsymbol{\alpha}_j = (\alpha_{0,j}, \alpha_{1,j})'$, and $\sigma^2_{v,j}$ for the jth Gibbs iteration. In other words, at the jth

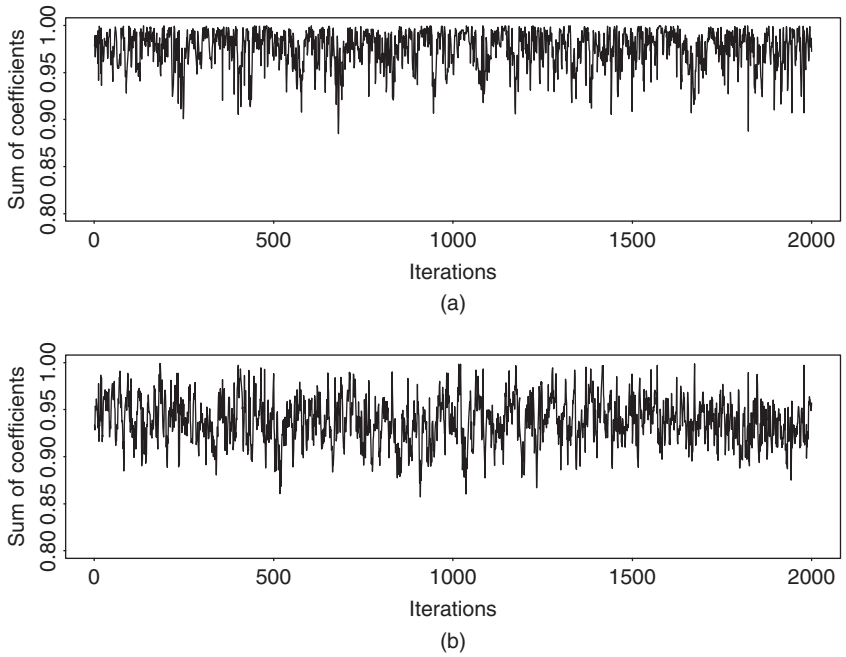

Figure 12.17 Time plots of persistent parameter $\alpha_{i1} + \alpha_{i2}$ of two-state Markov switching GARCH-M model for monthly log returns of GE stock from 1926 to 1999. Results based on last 2000 iterations of Gibbs sampling with 5000 + 2000 total iterations.

Gibbs iteration, the model is

$$r_t = \beta_{0,j} + \beta_{1,j} x_{1t} + \cdots + \beta_{p,j} x_{pt} + a_t, \tag{12.58}$$

$$\ln h_t = \alpha_{0,j} + \alpha_{1,j} \ln h_{t-1} + v_t, \qquad \text{Var}(v_t) = \sigma^2_{v,j}. \tag{12.59}$$

We can use this model to generate a realization of r_{n+i} and h_{n+i} for $i = 1, \ldots, \ell$. Denote the simulated realizations by $r_{n+i,j}$ and $h_{n+i,j}$, respectively. These realizations are generated as follows:

- Draw a random sample v_{n+1} from $N(0, \sigma^2_{v,j})$ and use Eq. (12.59) to compute $h_{n+1,j}$.
- Draw a random sample ϵ_{n+1} from $N(0, 1)$ to obtain $a_{n+1,j} = \sqrt{h_{n+1,j}}\epsilon_{n+1}$ and use Eq. (12.58) to compute $r_{n+1,j}$.
- Repeat the prior two steps sequentially for $n + i$ with $i = 2, \ldots, \ell$.

If we run a Gibbs sampling for $M + N$ iterations in model estimation, we only need to compute the forecasts for the last N iterations. This results in a random

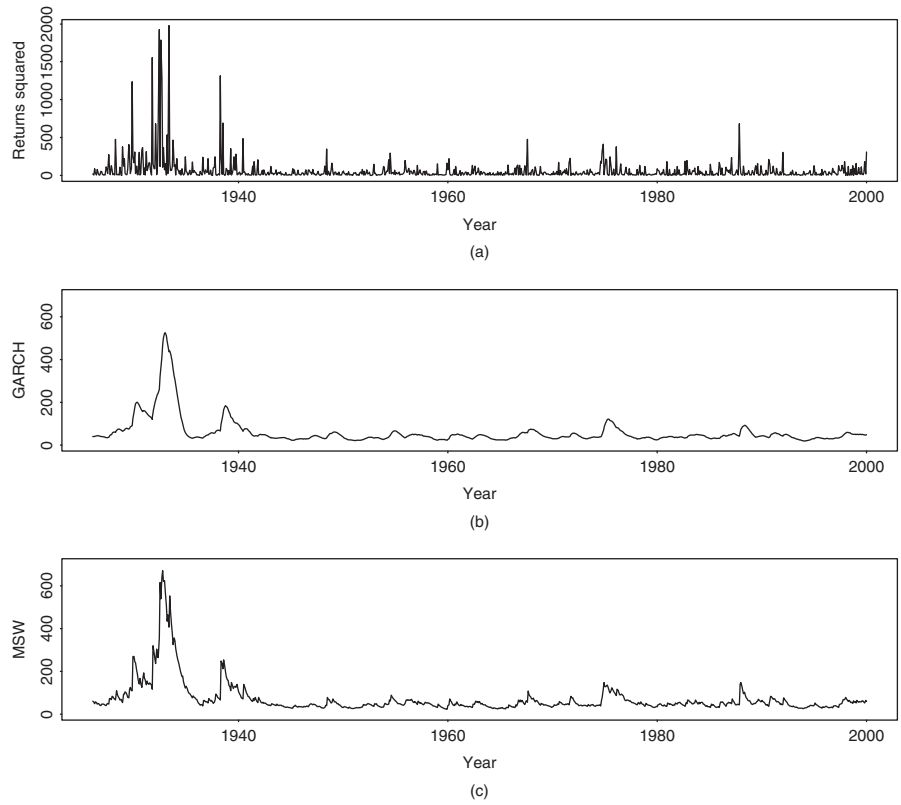

Figure 12.18 Fitted volatility series for monthly log returns of GE stock from 1926 to 1999: (a) squared log returns, (b) GARCH-M model in Eq. (12.59), and (c) two-state Markov switching GARCH-M model in Eq. (12.57).

sample for r_{n+i} and h_{n+i}. More specifically, we obtain

$$\{r_{n+1,j}, \ldots, r_{n+\ell,j}\}_{j=1}^{N}, \qquad \{h_{n+1,j}, \ldots, h_{n+\ell,j}\}_{j=1}^{N}.$$

These two random samples can be used to make inference. For example, point forecasts of the return r_{n+i} and volatility h_{n+i} are simply the sample means of the two random samples. Similarly, the sample standard deviations can be used as the variances of forecast errors. To improve the computational efficiency in volatility forecast, importance sampling can be used; see Gelman, Carlin, Stern, and Rubin (2003).

Example 12.7. (Example 12.3 continued) As a demonstration, we consider the monthly log return series of the S&P 500 index from 1962 to 1999. Table 12.6 gives the point forecasts of the return and its volatility for five forecast horizons starting with December 1999. Both the GARCH model in Eq. (12.26) and the

TABLE 12.6 Volatility Forecasts for Monthly Log Return of S&P 500 Index[a]

Horizon	1	2	3	4	5
			Log Return		
GARCH	0.66	0.66	0.66	0.66	0.66
SVM	0.53	0.78	0.92	0.88	0.84
			Volatility		
GARCH	17.98	18.12	18.24	18.34	18.42
SVM	19.31	19.36	19.35	19.65	20.13

[a]The data span is from January 1962 to December 1999 and the forecast origin is December 1999. Forecasts of the stochastic volatility model are obtained by a Gibbs sampling with 2000 + 2000 iterations.

stochastic volatility model in Eq. (12.27) are used in the forecasting. The volatility forecasts of the GARCH(1,1) model increase gradually with the forecast horizon to the unconditional variance $3.349/(1 - 0.086 - 0.735) = 18.78$. The volatility forecasts of the stochastic volatility model are higher than those of the GARCH model. This is understandable because the stochastic volatility model takes into consideration the parameter uncertainty in producing forecasts. In contrast, the GARCH model assumes that the parameters are fixed and given in Eq. (12.26). This is an important difference and is one of the reasons that GARCH models tend to underestimate the volatility in comparison with the implied volatility obtained from derivative pricing.

Remark. Besides the advantage of taking into consideration parameter uncertainty in forecast, the MCMC method produces in effect a predictive distribution of the volatility of interest. The predictive distribution is more informative than a simple point forecast. It can be used, for instance, to obtain the quantiles needed in value at risk calculation. □

12.11 OTHER APPLICATIONS

The MCMC method is applicable to many other financial problems. For example, Zhang, Russell, and Tsay (2008) use it to analyze information determinants of bid and ask quotes, McCulloch and Tsay (2001) use the method to estimate a hierarchical model for IBM transaction data, and Eraker (2001) and Elerian, Chib, and Shephard (2001) use it to estimate diffusion equations. The method is also useful in value at risk calculation because it provides a natural way to evaluate predictive distributions. The main question is not whether the methods can be used in most financial applications, but how efficient the methods can become. Only time and experience can provide an adequate answer to the question.

EXERCISES

12.1. Suppose that x is normally distributed with mean μ and variance 4. Assume that the prior distribution of μ is also normal with mean 0 and variance 25. What is the posterior distribution of μ given the data point x?

12.2. Consider the linear regression model with time series errors in Section 12.5. Assume that z_t is an AR(p) process (i.e., $z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t$). Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ be the vector of AR parameters. Derive the conditional posterior distributions of $f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\phi}, \sigma^2)$, $f(\boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$, and $f(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi})$ using the conjugate prior distributions, that is, the priors are

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o), \qquad \boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{A}_o), \qquad (v\lambda)/\sigma^2 \sim \chi^2_v.$$

12.3. Consider the linear AR(p) model in Section 12.6.1. Suppose that x_h and x_{h+1} are two missing values with a joint prior distribution being multivariate normal with mean $\boldsymbol{\mu}_o$ and covariance matrix $\boldsymbol{\Sigma}_o$. Other prior distributions are the same as that in the text. What is the conditional posterior distribution of the two missing values?

12.4. Consider the monthly log returns of Ford Motors stock from January 1965 to December 2008: (a) Build a GARCH model for the series, (b) build a stochastic volatility model for the series, and (c) compare and discuss the two volatility models. The simple returns of the stock are in the file `m-fsp6508.txt`.

12.5. Build a stochastic volatility model for the daily log return of Cisco Systems stock from January 2001 to December 2008. You may download the simple return of the stock from the CRSP database or the file `d-csco0108.txt`. Transform the data into log returns in percentage. Use the model to obtain a predictive distribution for 1-step-ahead volatility forecast at the forecast origin December 31, 2008. Finally, use the predictive distribution to compute the value at risk of a long position worth \$1 million with probability 0.01 for the next trading day.

12.6. Build a bivariate stochastic volatility model for the monthly log returns of Ford Motors stock and the S&P composite index for the sample period from January 1965 to December 2008. Discuss the relationship between the two volatility processes and compute the time-varying beta for the Ford stock.

12.7. Consider the monthly log returns of Procter & Gamble stock and the value-weighted index from January 1965 to December 2008. The simple returns are given in the file `m-pgvw6508.txt`. Transform the data into log returns in percentages. (a) Build a bivariate stochastic volatility model for the two return series. (b) Build a BEKK(1,1) model for the two series. (c) Compare and discuss the two models.

12.8. Consider the monthly data of 30-year mortgage rate and the 3-month Treasury Bill rate of the secondary market from April 1971 to September 2009.

The data are in `m-mort3mtb7109.txt`. (a) Build a regression model with time series error to study the effect of 3-month Treasury Bill rate on the mortgage rate. (b) Reestimate the model using MCMC method. (c) Compare and discuss the two fitted models.

REFERENCES

Artigas, J. C. and Tsay, R. S. (2004). Effective estimation of stochastic diffusion models with leverage effects and jumps. Working paper, Graduate School of Business, University of Chicago.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

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

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**: 541–553.

Chang, I., Tiao, G. C., and Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics* **30**: 193–204.

Chib, S., Nardari, F., and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* **108**: 281–316.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**: 1–38.

Elerian, O., Chib, S. and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69**: 959–993.

Eraker, B. (2001). Markov Chain Monte Carlo analysis of diffusion with application to finance. *Journal of Business & Economic Statistics* **19**: 177–191.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**: 183–202.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**: 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. Chapman and Hall/CRC, London.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.

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

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business & Economic Statistics* **12**: 371–417.

Jacquier, E., Polson, N. G., and Rossi, P. E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics* **122**: 185–212.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**: 389–395.

Justel, A., Peña, D., and Tsay, R. S. (2001). Detection of outlier patches in autoregressive time series. *Statistica Sinica* **11**: 651–673.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**: 361–393.

Liu, J., Wong, W. H., and Kong, A. (1994). Correlation structure and convergence rate of the Gibbs samplers. I. Applications to the comparison of estimators and augmentation schemes. *Biometrika* **81**: 27–40.

McCulloch, R. E. and Tsay, R. S. (1994a). Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis* **15**: 235–250.

McCulloch, R. E. and Tsay, R. S. (1994b). Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis* **15**: 523–539.

McCulloch, R. E. and Tsay, R. S. (2001). Nonlinearity in high-frequency financial data and hierarchical models. *Studies in Nonlinear Dynamics and Econometrics* **5**: 1–17.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association* **44**: 335–341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21** 1087–1092.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. Springer, New York.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**: 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**: 1701–1762.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting* **7**: 1–20.

Tsay, R. S., Peña, D., and Pankratz, A. (2000). Outliers in multivariate time series. *Biometrika* **87**: 789–804.

Zhang, M. Y., Russell, J. R., and Tsay, R. S. (2008). Determinants of bid and ask quotes and implications for the cost of trading. *Journal of Empirical Finance* **15**: 656–678.