Lecture 17: Some MCMC Applications in Time Series Analysis
Bus 41910, Time Series Analysis, Mr. R. Tsay

The key reference of this lecture is Chapter 10 of Tsay (2001) or Chapter 12 of Tsay (2005).

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 Markov Chain Monte Carlo (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 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 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.

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:

$$f_1(\theta_1|\theta_2, \theta_3, \boldsymbol{X}, M); \quad f_2(\theta_2|\theta_3, \theta_1, \boldsymbol{X}, M); \quad f_3(\theta_3|\theta_1, \theta_2, \boldsymbol{X}, M), \tag{1}$$

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:

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

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

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

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

$$\widehat{\theta}_i = \frac{1}{n-m} \sum_{j=m+1}^{n} \theta_{i,j}, \quad \widehat{\sigma}_i^2 = \frac{1}{n-m-1} \sum_{j=m+1}^{n} (\theta_{i,j} - \widehat{\theta}_i)^2. \tag{3}$$

The Gibbs sample in Eq. (2) can be used in many ways. For example, if one is interested in testing the null hypothesis $H_o : \theta_1 = \theta_2$ versus the alternative hypothesis $H_a : \theta_1 \neq \theta_2$, then she can simply obtain point estimate of $\theta = \theta_1 - \theta_2$ and its variance as

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

The null hypothesis can then be tested by using the conventional t ratio statistic $t = \widehat{\theta}/\widehat{\sigma}$. From the prior introduction, Gibbs sampling has the advantage to decompose 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.

3

3   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.

3.1   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 $t = 1, 2, \cdots$

   (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. (1995). 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. (1995).

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 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 $\boldsymbol{\theta}_o$.

3.2 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.

3.3 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 $j = 1, \cdots, m$.

2. Use $w_1, \cdots, w_m$ 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.

4 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. A simple version of the model is

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

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

$$
y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + z_t, \quad z_t = \phi z_{t-1} + a_t, \quad t = 1, \cdots, n, \tag{4}
$$

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, then we can estimate the regression model easily by using the least squares method. However, if the regression model is known, then 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:

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

where $\boldsymbol{Y} = (y_1, \cdots, y_n)'$ 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

$$
\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o), \quad \phi \sim N(\phi_o, \sigma_o^2), \quad \frac{v\lambda}{\sigma^2} \sim \chi_v^2, \tag{5}
$$

where again $\sim$ denotes distribution, $\boldsymbol{\beta}_o$, $\boldsymbol{\Sigma}_o$, λ, v, 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 $\boldsymbol{\beta}_o = \boldsymbol{0}$, $\phi_o = 0$, and $\boldsymbol{\Sigma}_o$ is a diagonal matrix with large diagonal elements. The prior distributions in Eq. (5) 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 conjugate priors in Bayesian inference. Specifically, given ϕ, we define

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

Using Eq. (4), we have

$$
y_{o,t} = \boldsymbol{\beta}'\boldsymbol{x}_{o,t} + a_t, \quad t = 2, \cdots, n. \tag{6}
$$

Under the assumption of $\{a_t\}$, Eq. (6) 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 Results 1a, the posterior distribution of $\boldsymbol{\beta}$, given the data, ϕ, and σ^2, is multivariate normal. We write the result as

$$(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma) \sim N(\boldsymbol{\beta}_*, \boldsymbol{\Sigma}_*), \tag{7}$$

where the parameters are given by

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

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

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

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 1, the posterior distribution of ϕ is also normal with mean ϕ_* and variance σ_*^2, where

$$\sigma_*^{-2} = \frac{\sum_{t=2}^{n} z_{t-1}^2}{\sigma^2} + \sigma_o^{-2}, \quad \phi_* = \sigma_*^2 \left(\frac{\sum_{t=2}^{n} z_{t-1}^2}{\sigma^2}\widehat{\phi} + \sigma_o^{-2}\phi_o\right). \tag{8}$$

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

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

Based on conjugate priors, the posterior distribution of σ^2 is an inverted chi-squared distribution − that is,

$$\frac{v\lambda + \sum_{t=2}^{n} a_t^2}{\sigma^2} \sim \chi_{v+(n-1)}^2, \tag{9}$$

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

Using the three conditional posterior distributions in Eqs. (7)−(9), we can estimate Eq. (4) via Gibbs sampling as follows:

1. Specify the hyperparameter values of the priors in Eq. (5).

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. (7) to draw a random realization for $\boldsymbol{\beta}$.

4. Use the univariate normal distribution in Eq. (8) to draw a random realization for ϕ.

5. Use the chi-squared distribution in Eq. (9) 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 (4).

5   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

$$
y_t = \begin{cases} x_h + \omega & \text{if } t = h \\ x_t & \text{otherwise,} \end{cases}
\tag{10}
$$

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.

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. 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.

5.1 Missing values

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

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t, \tag{11}$$

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 $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)'$. Thus, we treat the missing value x_h as an unknown parameter. If we assume that the prior distributions are

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

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 $\{x_{h-p}, \cdots, x_{h-1}, x_{h+1}, \cdots, x_{h+p}\}$. Keeping in mind that x_h is an unknown parameter, we can write the relationship as follows:

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

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

    Let $y_h = \phi_1 x_{h-1} + \cdots + \phi_p x_{h-p}$ 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

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

    Let $y_{h+1} = x_{h+1} - \phi_2 x_{h-1} - \cdots - \phi_p x_{h+1-p}$ 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 $j = 1, \cdots, p$, 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

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

9

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

$$y_{h+j} = \phi_j x_h + b_{h+j}, \quad j = 0, \cdots, p, \tag{12}$$

where $\phi_0 = 1$. Since a normal distribution is symmetric with respective to its mean, a_h and $-a_h$ have the same distribution. Consequently, Eq. (12) is a special simple linear regression model with $p+1$ data points. The least squares estimate of x_h and its variance are

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

For instance, when $p = 1$, we have $\widehat{x}_h = \frac{\phi_1}{1+\phi_1^2}(x_{h-1} + x_{h+1})$, 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 conjugate prior, we obtain that the posterior distribution of x_h is normal with mean μ_* and variance σ_*^2, where

$$\mu_* = \frac{\sigma^2 \mu_o + \sigma_o^2 (\sum_{j=0}^p \phi_j^2) \widehat{x}_o}{\sigma^2 + \sigma_o^2 (\sum_{j=0}^p \phi_j^2)}, \quad \sigma_*^2 = \frac{\sigma^2 \sigma_o^2}{\sigma^2 + \sigma_o^2 \sum_{j=0}^p \phi_j^2}. \tag{13}$$

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 $\{x_{h-p}, \cdots, x_{h-1}; x_{h+2}, \cdots, x_{h+p+1}\}$. 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. (13) 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 assume $h - p \geq 1$ 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.

5.2 Outlier detection

Detection of additive outliers in Eq. (10) 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 (1994) 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

$$y_t = \delta_t \beta_t + x_t, \quad t = 1, \cdots, n, \tag{14}$$

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. (14), we have n data points, but there are $2n + p + 3$ parameters − namely, $\boldsymbol{\phi} = (\phi_0, \cdots, \phi_p)'$, $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)'$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)'$, σ^2, and ϵ. The binary parameters δ_t are governed by ϵ and β_ts 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

$$\boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{\Sigma}_o), \quad \frac{v\lambda}{\sigma^2} \sim \chi_v^2, \quad \epsilon \sim \text{beta}(\gamma_1, \gamma_2), \quad \beta_t \sim N(0, \xi^2),$$

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

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

$$f(\epsilon|\boldsymbol{Y}, \boldsymbol{\delta}), \quad f(\sigma^2|\boldsymbol{Y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}),$$

where $1 \leq h \leq n$, $\boldsymbol{Y}$ 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 $\boldsymbol{x}_{t-1} = (1, x_{t-1}, \cdots, x_{t-p})'$, 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. (7) 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

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

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

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

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 $j = h, \cdots, h+p$. In other words, we have

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

- 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

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

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

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 $j = h, \cdots, h+p$.

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

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

$$P(\delta_h = 1|\boldsymbol{Y}, \boldsymbol{\delta}_{-h}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2) =$$

$$\frac{\epsilon \exp[-\sum_{j=h}^m (w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)]}{\epsilon \exp[-\sum_{j=h}^m (w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)] + (1 - \epsilon) \exp[-\sum_{j=h}^m w_j^2/(2\sigma^2)]}. \tag{15}$$

This posterior distribution is simply to compare the weighted values of 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 $j = h, h + 1, \cdots, \min(h + p, n)$. Specifically, we have $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, h + 1, \cdots, \min(h + p, n)$. The information can be put in a linear regression framework as

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

Consequently, the information is embedded in the least squares estimate

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

which is normally distributed with mean β_h and variance $\sigma^2/(\sum_{j=h}^m \psi_{j-h}^2)$. By Result 1, the posterior distribution of β_h is normal with mean β_h^* and variance σ_{h*}^2, where

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