Financial Econometrics
Lecture 12: Some MCMC Applications in Time Series Analysis

Prof Hamed Ghoddusi
2019

A Short Overview of Bayesian Econometrics

Bayesian versus Frequentist View

- Frequentist view:
  - Probability: the result of repeated experiment
  - Parameters: unknown constants
  - Confidence interval (CI)
- Bayesian view
  - Probability: subjective belief about parameters
  - Updating beliefs based on new evidence (data)
  - Parameters: stochastic

Reminder Bayes Formula

- Conditional Probability and Bayes theorem

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

- In the context of an *inference* problem:
- A = Parameters , B = Data

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{2}$$

- $P(Parameters|Data) = \frac{P(Data|Parameters)P(Parameters)}{P(Data)}$

Bayesian View to Estimation and Inference

- Assume a continuous distribution for θ

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} = \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta} \tag{3}$$

- This can be viewed as

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \propto P(y|\theta)P(\theta) \tag{4}$$

- Posterior Density for $\theta \propto$ Prior Density for θ * Likelihood Function

Bayesian View to Estimation

- Prior belief about the distribution of parameters
  - Parameters used in a very broad sense
  - Hyperparameter: a parameter of a prior distribution
- Likelihood function
- Posterior distribution
- Normalizing factor

Example: Estimating Bernoulli Distribution

- Estimating the probability of success in a Bernoulli model
- Unifirm prior for θ: $P(\theta) = 1$, $0 <= \theta <= 1$
- Likelihood function based on a random sample of a n observations

$$L(\theta|y) = \prod_{i=1}^{n}[\theta^{y_i}(1-\theta)^{1-y_i}] = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \tag{5}$$

- This is a Beta p.d.f (after adding some normalizing constants to make it a pdf)
- Maximum likelihood estimation (MLE): maximize $L(\theta|y) \Rightarrow$ provide one point

Example: Financial Econometrics

- We observe a vector T of returns $R = [r_1, r_2, ..., r_T]$
- Each return is normally distributed $r_i \sim N(\mu, \sigma^2)$
  - μ is a stochastic random variable denoting the mean return
  - Apply Bayes rule: $\underbrace{P(\mu|R,\sigma^2)}_{\text{Posterior}} \propto \underbrace{P(\mu)}_{\text{Prior}} \underbrace{P(R|\mu,\sigma^2)}_{\text{Likelihood}}$
- Likelihood function of individual normal is known: $P(r|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(r_t-\mu)^2}$

Posterior

- Since returns are assumed to be IID, the joint likelihood of all realized returns is

$$P(R|\mu,\sigma) = [\frac{1}{\sqrt{2\pi\sigma^2}}]^T e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{T}(r_i-\mu)^2} \tag{6}$$

- With diffuse prior and normal likelihood, the posterior is proportional to the likelihood function

$$P(\mu|R,\sigma) \propto e^{\frac{-1}{2\sigma^2}[vs^2+T(\mu-\hat{\mu})]^2} \tag{7}$$

Imposing an Informative Prior

- A normally distributed prior
- Posterior

$$P(\mu|R,\sigma) \propto e^{\frac{(\mu-\mu_a)^2}{2\sigma_p^2}+\frac{T(\mu-\hat{\mu})^2}{2\sigma^2}} \tag{8}$$

- where μ_a is the mean of prior, $\hat{\mu} = \frac{\sum_{i=1}^{T} r_i}{T}$

Posterior Distribution of Parameters

- Posterior mean and variance of the return are a combination of prior and data driven evidence:
- $\frac{1}{\tilde{\sigma}} = \frac{1}{\sigma_p^2} + \frac{T}{\sigma^2}$
- $\tilde{\mu} = \tilde{\sigma^2}[\frac{\mu_0}{\sigma_P^2} + \frac{T\hat{\mu}}{\sigma^2}]$
- It is common to think in terms of the precision parameter $\lambda = \frac{1}{\sigma^2}$
- Note what happens if $T \to \infty$

Some Jargons about Priors

- Objective versus subjective priors
- Conjugate priors: induces a posterior distributions in the same probability distribution family as the prior probability
  - Example: Normal, Gamma, Beta distributions
- Informative prior: expresses specific, definite information about a variable.
- Diffuse (uninformative) prior: providing vague or general information about a variable
- Improper priors: infinitesimal over an infinite range, in order to add to one
  - Example: the uniform prior over all real numbers

Conjugate Priors

- Conjugate priors guarantee that a closed-form solution for the conditional posterior distribution exists
- Great news for MCMC: we can use standard computer commands to generate random draws (samples)
- Some famous cases:
  - Normal distribution with know variance but unknown μ: assume a normal distribution for μ
  - Multivariate normal distribution with know VCV but unknown μ: multivariate normal
  - Normal distribution with know mean but unknown σ: assume a gamma distribution for σ
  - Normal distribution with unknown mean and unknown σ: assume a normal distribution for μ and a gamma distribution for σ

Wrap Up: Why Bayesian econometrics?

- Philosophically appealing
- Produce a range of possible values for a parameter
- Specify a much richer model sets (BMA = Bayesian Model Averaging)
- Include subjective beliefs in the estimation (Black-Litterman model of asset allocation)
- Flexibility in using computational methods

Wrap Up: Why MCM?

- MCMC= Markov-chain that has as its equilibrium distribution the target posterior distribution
- Generating posteriors with non-standard distributions
- Evaluating large multi-variate integrals
- Calculating the *normalization factor of Bayesian models.*

Outline

Markov Chain Simulation

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

Markov Chain Simulation

- The key to Markov chain simulation is:
  - To create a Markov process whose stationary transition distribution is a specified $P(\theta|X)$.
  - To run the simulation sufficiently long so that the distribution of the current values of the process is close enough to the stationary transition distribution.
- In other words, the values of the process can be regarded as random draws from the transition distribution.
- It turns out that, for a given $P(\theta|X)$, many Markov chains with the desired property can be constructed.
- We refer to methods that use Markov chain simulation to obtain the distribution $P(\theta|X)$ as Markov Chain Monte Carlo (MCMC) methods.

Markov Chain Simulation

- 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 course were developed under the assumption of “complete data” (i.e., there is no missing value).
- For example, in forecasting U.S. quarterly unemployment rates, we assume that the unemployment rates are available for each quarter in the sample period.
- What should we do if there is a missing value?

Markov Chain Simulation

- 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 time series model for the unemployment rates.
  - 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.

Markov Chain Simulation

- 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.

Outline

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.
- 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.

Gibbs Sampling

- Denote the three parameters by θ_1, θ_2, and θ_3.
- Let 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:

$$f_1(\theta_1|\theta_2, \theta_3, X, M); f_2(\theta_2|\theta_3, \theta_1, X, M); f_3(\theta_3|\theta_1, \theta_2, X, M), \qquad (9)$$

where $f_i(\theta_i|\theta_{j\neq i}, X, M)$ denotes the conditional distribution of the parameter θ_i given the data, the model, and the other two parameters.

Gibbs Sampling

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

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

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

Gibbs Sampling

- 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 $f(\theta_1, \theta_2, \theta_3|X, M)$ 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:

$$(\theta_{1,m+1}, \theta_{2,m+1}, \theta_{3,m+1}), \ldots, (\theta_{1,n}, \theta_{2,n}, \theta_{3,n}). \quad (10)$$

Gibbs Sampling

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

$$\hat{\theta}_i = \frac{1}{n-m}\sum_{j=m+1}^{n}\theta_{i,j}, \quad \hat{\sigma}_i^2 = \frac{1}{n-m-1}\sum_{j=m+1}^{n}(\theta_{i,j} - \hat{\theta}_i)^2. \quad (11)$$

- The Gibbs sample in Eq. (10) can be used in many ways:
  - For example, if one is interested in testing the null hypothesis $H_0 : \theta_1 = \theta_2$ 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:

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

Gibbs Sampling

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

Gibbs Sampling

- Consider the three-parameter illustrative example:
  - If θ_1 and θ_2 are highly correlated, then one should employ the conditional distributions $f(\theta_1, \theta_2|\theta_3, X, M)$ and $f_3(\theta_3|\theta_1, \theta_2, X, M)$ whenever possible.
- A Gibbs iteration then consists of:
  a. drawing jointly (θ_1, θ_2) given θ_3
  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).

Gibbs Sampling

- 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.
- 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.

Outline

Metropolis Algorithm

- Applicable when the conditional posterior distribution is known except for a normalization constant.
- Suppose that we want to draw a random sample from the distribution $f(\theta|X)$, which contains a complicated normalization constant so that a direct draw is either too time-consuming or infeasible.

Metropolis Algorithm

- 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 $f(\theta|X)$.
- Performs a random walk in the parameter space, and will stay at a parameter value proportional to its posterior probability.

Metropolis Algorithm

1. Draw a random starting value θ_0 such that $f(\theta_0|X) > 0$.
2. For $t = 1, 2, \ldots$
   a. Draw a candidate sample $\theta_\star$ from a known distribution at iteration t given the previous draw θ_{t-1}. Denote the known distribution by $J_t(\theta_t|\theta_{t-1})$. The jumping distribution must be symmetric - that is, $J_t(\theta_i|\theta_j) = J_t(\theta_j|\theta_i)$ for all θ_i, θ_j , and t.
   b. Calculate the ratio
   $$r = \frac{f(\theta_*|X)}{f(\theta_{t-1}|X)}.$$
   c. Set
   $$\theta_t = \begin{cases} \theta_* & \text{with probability} \min(r, 1) \\ \theta_{t-1} & \text{otherwise.} \end{cases}$$
   Under some regularity conditions, the sequence $\{\theta_t\}$ converges in distribution to $f(\theta|X)$; see Gelman et al. (1995).

Metropolis Algorithm

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

Metropolis Algorithm

- The acceptance and rejection rule of the algorithm can be stated as follows:
  i. if the jump from θ_{t-1} to $\theta_\star$ increases the conditional posterior density, then accept $\theta_\star$ as θ_t
  ii. if the jump decreases the posterior density, then set $\theta_t = \theta_\star$ with probability equal to the density ratio r, and set $\theta_t = \theta_{t-1}$ 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 $f(\theta_i|\theta_j) = f(\theta_j|\theta_i)$, where $f(\theta|\theta_o)$ denotes a multivariate normal density function with mean vector θ_o.

Metropolis-Hasting algorithm

- Hasting (1970) generalizes the Metropolis algorithm in two ways:
  1. The jumping distribution does not have to be symmetric.
  2. The jumping rule is modified to:

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

This modified algorithm is referred to as the Metropolis-Hasting algorithm.

Griddy Gibbs

- In economic or 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.

Griddy Gibbs

- Let θ_i be a scalar parameter with conditional posterior distribution $f(\theta_i|X, \theta_{-i})$, where θ_{-i} is the parameter vector after removing θ_i.
- For instance, if $\theta = (\theta_1, \theta_2, \theta_3)'$, then $\theta_{-1} = (\theta_2, \theta_3)'$.
- 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 $w_j = f(\theta_{ij}|X, \theta_{-i})$ for $j = 1, \ldots, m$.
  2. Use $w_1, \ldots, w_m$ to obtain an approximation to the inverse cumulative distribution function (CDF) of $f(\theta_i|X, \theta_{-i})$.
  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.

Outline

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 economics and finance.
- The first example is to estimate a regression model with serially correlated errors.

Linear Regression With Time-Series Errors

- 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.

Linear Regression With Time-Series Errors

Denote the parameters of the model by $\theta = (\beta', \phi, \sigma_2)'$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$, and let $x_t = (1, x_{1t}, \ldots, x_{kt})'$ be the vector of all regressors at time t, including a constant of unity.
The model becomes:

$$y_t = x_t'\beta + z_t, \quad z_t = \phi z_{t-1} + a_t, \quad t = 1, \ldots, n, \tag{4}$$

where n is the sample size.

Linear Regression With Time-Series Errors

- 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 $z_t = y_t - x_t'\beta$ and use the series to estimate the AR(1) model.
- We need the following conditional posterior distributions:

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

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

Linear Regression With Time-Series Errors

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

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

  where again $\sim$ denotes distribution, $\beta_o, \Sigma_o, \lambda, v, \phi_o$, 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 $\beta_o = 0$, $\phi_o = 0$, and Σ_o is a diagonal matrix with large diagonal elements.

Linear Regression With Time-Series Errors

- The prior distributions in Eq. (5) are assumed to be independent of each other.
- Thus, we use independent priors based on thepartition of the parameter vector θ.
- The conditional posterior distribution $f(\beta|Y, X, \phi, \sigma^2)$ can be obtained by conjugate priors in Bayesian inference.
- Specifically, given ϕ, we define

$$y_{o,t} = y_t - \phi y_{t-1}, \quad x_{o,t} = x_t - \phi x_{t-1}.$$

- Using Eq. (4), we have

$$y_{o,t} = \beta' x_{o,t} + a_t, \quad t = 2, \dots, n. \tag{6}$$

- Under the assumption of at, Eq. (6) is a multiple linear regression.

Linear Regression With Time-Series Errors

- Therefore, information of the data about the parameter vector β is contained in its least squares estimate

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

  which has a multivariate normal distribution

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

- Using Results 1a of Tsay (2005, Ch. 12), the posterior distribution of β, given the data,ϕ, and σ^2, is multivariate normal. We write the result as

$$(\beta|Y, X, \phi, \sigma) \sim N(\beta_*, \Sigma_*), \tag{7}$$

Linear Regression With Time-Series Errors

where the parameters are given by

$$\Sigma_*^{-1} = \frac{\sum_{t=2}^{n} x_{o,t} x'_{o,t}}{\sigma^2} + \Sigma_o^{-1}, \quad \beta_* = \Sigma_* \left(\frac{\sum_{t=2}^{n} x_{o,t} x'_{o,t}}{\sigma^2} \hat{\beta} + \Sigma_o^{-1} \beta_o\right).$$

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

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

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

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

Linear Regression With Time-Series Errors

which is normally distributed with mean ϕ and variance $\sigma^2(\sum_{t=2}^n z_{t-1}^2)^{-1}$.

- Based on Result 1 of Tsay (2005, Ch. 12), the posterior distribution of ϕ is also normal with mean ϕ_* and variance σ_*^2 where

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

- Finally, turn to the posterior distribution of σ^2 given β, ϕ, and the data.
- Because β and ϕ are known, we can calculate

$$a_t = z_t - \phi z_{t-1}, \quad z_t = y_t - \beta' x_t, \quad t = 2, \ldots, n.$$

Linear Regression With Time-Series Errors

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

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

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 β, ϕ, and σ^2 (e.g., the ordinary least squares estimate of β without time-series errors).

Linear Regression With Time-Series Errors

3. Use the multivariate normal distribution in Eq. (7) to draw a random realization for β.
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).

Outline

Missing values and outliers

- In this section, we discuss MCMC $\{y_t\}_{t=1}^{n}$ 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.

Missing values and outliers

- 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 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.
- If the chance to observe y_h is very small under the derived distribution, then y_h can be classified as an additive outlier.
- Detection of additive outliers and treatment of missing values in time-series analysis are based on the same idea.

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 $\theta = (\phi', x_h, \sigma^2)'$, where $\phi = (\phi_1, \ldots, \phi_p)'$.
- Thus, we treat the missing value x_h as an unknown parameter.

Missing values

- If we assume that the prior distributions are

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

- where the hyperparameters are known, then the conditional posterior distributions $f(\phi|X, x_h, \sigma^2)$ and $f(\sigma^2|X, x_h, \phi)$ are exactly as those given in the previous section, where X denotes the observed data.
- The conditional posterior distribution $f(x_h|X, \phi, \sigma^2)$ 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 $\{x_{h-p}, \ldots, x_{h-1}, x_{h+1}, \ldots, x_{h+p}\}$.

Missing values

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

Missing values

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
$$y_{h+j} = \phi_j x_h + b_{h+j}, \quad j = 0, \ldots, p, \tag{12}$$
where $\phi_0 = 1$.

Missing values

- 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.

Missing values

- Finally, using conjugate prior, we obtain that the posterior distribution of x_h is normal with mean μ_* and variance σ_*^2, where
$$\mu_* = \frac{\sigma^2 \mu_o + \sigma_o^2 (\sum_{j=0}^{p} \phi_j^2) \widehat{x}_h}{\sigma^2 + \sigma_o^2 (\sum_{j=0}^{p} \phi_j^2)}, \quad \sigma_*^2 = \frac{\sigma^2 \sigma_o^2}{\sigma^2 + \sigma_o^2 \sum_{j=0}^{p} \phi_j^2}. \tag{13}$$
- Missing values may occur in patches, resulting in the situation of multiple consecutive missing values.
- These missing values can be handled in **two** ways.

Missing values

- **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.

Missing values

- **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$(x_h|X, x_{h+1}, \phi, \sigma^2)$ and $f(x_{h+1}|X, x_h, \phi, \sigma^2)$ 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.

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.

Outlier detection

- We write the model for y_t as
$$y_t = \delta_t \beta_t + x_t, \quad t = 1, \ldots, n, \tag{14}$$
where $\{\delta_t\}$ is a sequence of independent Bernoulli random variables such that $P(\delta_t = 1) = \epsilon$ and $P(\delta_t = 0) = 1 - \epsilon, \epsilon$ 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.

Outlier detection

- Under the model in Eq. (14), we have n data points, but there are $2n+p+3$ parameters - namely, $\phi=(\phi_0,\ldots,\phi_p)'$, $\delta=(\delta_1,\ldots,\delta_n)'$, $\beta=(\beta_1,\ldots,\beta_n)'$, σ^2, and ϵ.
- The binary parameters δ_t are governed by ϵ and $\beta_t s$ are determined by the specified distribution.
- The parameters δ and β 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.

Outlier detection

- Assume that the prior distributions are

$$\phi \sim N(\phi_o,\Sigma_o),\quad \frac{v\lambda}{\sigma^2}\sim\chi^2_v,\quad \epsilon\sim \text{beta}(\gamma_1,\gamma_2),\quad \beta_t\sim N(0,\xi^2),$$

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

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

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

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

Outlier detection

- Conditioned on δ and β, the outlier-free time series x_t can be obtained by $x_t=y_t-\delta_t\beta_t$.
- Information of the data about ϕ is then contained in the least squares estimate

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

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

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

Outlier detection

- The conditional posterior distribution of ϕ is therefore multivariate normal with mean ϕ_* and covariance matrix Σ_*, which are given in Eq. (7) with β being replaced by ϕ and $x_{o,t}$ by 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 $a_t=x_t-\phi' x_{t-1}$ and $x_t=y_t-\delta_t\beta_t$.

Outlier detection

- 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 $j \neq h$, ϕ, 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

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

    where $x_j^* = x_j$ if $j \neq h$ and $x_h^* = y_h$.

Outlier detection

- 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

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

  - 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

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

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

Outlier detection

- 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

$$P(\delta_h = 1 | Y, \delta_{-h}, \beta, \phi, \sigma^2) = \frac{\epsilon \exp[-\sum_{j=h}^{m}(w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)]}{\epsilon \exp[-\sum_{j=h}^{m}(w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)] + (1-\epsilon)\exp[-\sum_{j=h}^{m} w_j^2/(2\sigma^2)]}. \quad (15)$$

Outlier detection

- 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

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

Outlier detection

- 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 of Tsay (2005, Ch. 12), the posterior distribution of β_h is normal with mean β_h^* and variance σ_{h*}^2, where

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

- For demonstration, see Chapter 12 of Tsay (2005) and the references therein.