Chapter 12

INTRODUCTION TO BAYESIAN APPROACH

Over the last decade, Bayesian analysis has become an increasingly popular method in economics. As you will see in this chapter, the Bayesian approach differs from the classical frequentist approach in various aspects. The fundamental difference lies in its probabilistic interpretation of the object of interest such as unknown parameters and random events. In the Bayesian framework, unknown parameters are treated as random variables while the observed data are treated as fixed. This interpretation allows us to assign a probability distribution associated with the parameters upon which Bayesian inferences are made.

This chapter introduces basic concepts and implementation of Bayesian analysis. Next section explains probability density functions in Bayesian statistics, followed by their application to generating point estimates and constructing Bayesian credible intervals. We then discuss posterior odds ratio tests for hypothesis testing and model comparison. Details of each topic can be found in DeJong and Dave (2007), Judge *et al*(1985), and Zellner (1996). The appendix to this chapter explains simulation methods that are widely used in the implementation of Bayesian analysis.

12.1 Bayes Theorem

Bayesian analysis centers around the representation of our uncertainty about the object of interest such as true values of unknown parameters. A prior distribution represents our initial knowledge or subjective beliefs about the unknown parameters held prior to observing data. After the data has been observed, sample information is incorporated into the prior to form a posterior distribution which assigns a probability to alternative parameter values based on the information from the prior and the data. Bayes' theorem is a mathematical formula in probability theory that relates the posterior distribution to the prior and the sample information represented by a likelihood function.

Suppose we are interested in a vector of unknown parameters $\boldsymbol{\theta}$. Let $p(\boldsymbol{\theta})$ denote a prior density function for $\boldsymbol{\theta}$, and $\mathbf{y}$ a vector of sample observations from a density $f(\mathbf{y}|\boldsymbol{\theta})$. A joint probability density for $\boldsymbol{\theta}$ and $\mathbf{y}$ is given by

$$P(\boldsymbol{\theta}, \mathbf{y}) = p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}). \tag{12.1}$$

Rearranging the second equality in (12.1) yields a posterior density function for $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y})}.$$

This result is Bayes' theorem, showing how our prior knowledge $p(\boldsymbol{\theta})$ is combined with sample information $f(\mathbf{y}|\boldsymbol{\theta})$ to generate the posterior distribution. Since we are interested in the distribution of $\boldsymbol{\theta}$, $f(\mathbf{y})$ may be treated as a normalizing constant, and $p(\boldsymbol{\theta}|\mathbf{y})$ is in general analyzed up to constant proportionality:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}). \tag{12.2}$$

Here, $f(\mathbf{y}|\boldsymbol{\theta})$ is algebraically identical to a likelihood function $l(\boldsymbol{\theta}|\mathbf{y})$, and (12.2) may be expressed as $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})l(\boldsymbol{\theta}|\mathbf{y})$; that is, the posterior distribution is proportional to the product of the prior and the likelihood function. The posterior distribution serves as an essential element of Bayesian inferences such as generating point estimates, constructing confidence intervals, and conducting hypothesis testing which we discuss in next sections.

12.2 Parameter Estimates

In general, Bayesian point estimates are obtained by specifying a loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ which quantifies the consequences of choosing $\hat{\boldsymbol{\theta}}$ when the true value is $\boldsymbol{\theta}$. An optimal point estimate is the value $\hat{\boldsymbol{\theta}}$ which minimizes the expected loss where the expectations are with respect to the posterior distribution of $\boldsymbol{\theta}$:

$$\min_{\hat{\boldsymbol{\theta}}} E\left(L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})\right) = \min_{\hat{\boldsymbol{\theta}}} \int L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

In the case of a quadratic loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \boldsymbol{\Phi} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ where $\boldsymbol{\Phi}$ is a symmetric positive definite matrix, an optimal point estimate is given by the mean of the posterior distribution. Alternatively, if the loss is measured by an absolute error $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|$, then the median of the posterior distribution becomes an optimal point estimate.

12.3 Bayesian Intervals and Regions

A Bayesian counterpart of a classical confidence interval is called a posterior credible interval (or region if $\boldsymbol{\theta}$ is a vector of parameters). For a scalar θ in a parameter space

Ω, a $100 \cdot (1-\alpha)$ percent posterior credible interval is a subset $S \subset \Omega$ such that

$$Pr(\theta \in S|\mathbf{y}) = \int_S p(\theta|\mathbf{y})d\theta = 1-\alpha. \tag{12.3}$$

For any given α, the interval S satisfying (12.3) may not be unique. Of those satisfying (12.3), a highest posterior density interval is obtained by imposing an additional condition that the value of $p(\theta|\mathbf{y})$ at any θ inside S is at least as large as that evaluated outside S; that is,

$$p(\theta_i|\mathbf{y}) \geq p(\theta_j|\mathbf{y}) \quad \text{for all} \quad \theta_i \in S \quad \text{and} \quad \theta_j \notin S,$$

which implies that the end points of the interval, say $\underline{\theta}$ and $\overline{\theta}$, satisfy $p(\underline{\theta}|\mathbf{y}) = p(\overline{\theta}|\mathbf{y})$. If the posterior density is unimodal, a highest posterior density interval is an interval that satisfies (12.3) with a minimum distance between $\underline{\theta}$ and $\overline{\theta}$.

While a highest posterior density interval is identical to a $100 \cdot (1-\alpha)$ percent confidence interval in the classical framework, their interpretations are different. A classical confidence interval is a random interval which would contain a fix value θ with probability $(1-\alpha)$ if we repeatedly draw samples from population and construct an interval each time. On the other hand, a highest posterior density interval is a fixed interval within which a random variable θ lies with probability $(1-\alpha)$.

12.4 Posterior Odds Ratio and Hypothesis Testing

Posterior distributions are also employed to assess relative plausibility of competing hypotheses. We evaluate the relative plausibility with a ratio of posterior probabilities associated with the hypotheses, called a posterior odds ratio. Unlike the classical hypothesis testing, a posterior odds ratio test treats the competing hypotheses symmetrically, and its conclusion is not designed to necessarily accept or reject

the hypotheses. Instead, the test merely infers which hypothesis is more likely given the priors and sample information.

Suppose we are interested in comparing two hypotheses, H_0 and H_1, with prior probabilities $p(H_0)$ and $p(H_1)$. Let $\boldsymbol{\theta}_i$ denote a parameter vector associated with hypothesis H_i, $i = 0, 1$. For H_0, the joint density function for $\mathbf{y}$, $\boldsymbol{\theta}_0$, H_0 is,

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}_0, H_0) &= f(\mathbf{y}) p(\boldsymbol{\theta}_0, H_0|\mathbf{y}) \\ &= p(\boldsymbol{\theta}_0, H_0) f(\mathbf{y}|\boldsymbol{\theta}_0, H_0) \end{aligned}$$

or

$$\begin{aligned} p(\boldsymbol{\theta}_0, H_0|\mathbf{y}) &= \frac{p(\boldsymbol{\theta}_0, H_0) f(\mathbf{y}|\boldsymbol{\theta}_0, H_0)}{f(\mathbf{y})} \\ &= \frac{p(H_0) h(\boldsymbol{\theta}_0|H_0) f(\mathbf{y}|\boldsymbol{\theta}_0, H_0)}{f(\mathbf{y})}, \end{aligned} \tag{12.4}$$

where $h(\boldsymbol{\theta}_0|H_0)$ is the conditional prior distribution for $\boldsymbol{\theta}_0$ given H_0. The posterior distribution of H_0 can be obtained by integrating (12.4) with respect to $\boldsymbol{\theta}_0$:

$$p(H_0|\mathbf{y}) = \frac{p(H_0) \int h(\boldsymbol{\theta}_0|H_0) f(\mathbf{y}|\boldsymbol{\theta}_0, H_0) d\boldsymbol{\theta}_0}{f(\mathbf{y})}.$$

Given that $p(H_1|\mathbf{y})$ has been obtained in an analogous way, the posterior odds ratio is,

$$\begin{aligned} \frac{p(H_0|\mathbf{y})}{p(H_1|\mathbf{y})} &= \frac{p(H_0)}{p(H_1)} \frac{\int h(\boldsymbol{\theta}_0|H_0) f(\mathbf{y}|\boldsymbol{\theta}_0, H_0) d\boldsymbol{\theta}_0}{\int h(\boldsymbol{\theta}_1|H_1) f(\mathbf{y}|\boldsymbol{\theta}_1, H_1) d\boldsymbol{\theta}_1} \\ &= \frac{p(H_0)}{p(H_1)} \frac{f(\mathbf{y}|H_0)}{f(\mathbf{y}|H_1)}. \end{aligned} \tag{12.5}$$

The larger the value of this ratio, the more the test is in favor of H_0.

The first term in (12.5), $p(H_0)/p(H_1)$, is called a prior odds ratio, and the second term $f(\mathbf{y}|H_0)/f(\mathbf{y}|H_1)$ is the ratio of averaged likelihoods, called a Bayes factor. If

we assume, prior to observing the data, that the two hypotheses are equally likely, then the prior odds ratio is 1. In that case, the relative plausibility is determined by the Bayes factor, and we can conveniently interpret its value using the following scale developed by Jeffreys (1961):

Bayes factor	Evidence in favor of H_0
1:1 - 3:1	Very slight
3:1 - 10:1	Slight
10:1 - 100:1	Strong to very strong
100:1 -	Decisive

Although the posterior odds ratio itself does not make an explicit conclusion about accepting or rejecting one hypothesis with respect to the other, it is still possible to make an explicit choice between the two, if necessary. In such cases, a loss function is assumed to measure the consequences of choosing each hypothesis, and we accept one which yields the lowest expected loss, with the expectation with respect to the posterior probability of the hypothesis.

One useful application of a posterior odds ratio is the assessment of relative plausibility of competing models which may not be nested (for empirical applications, see Lubik and Schorfheide, 2007; Rabanal and Rubio-Ramirez, 2005). Its implementation follows the same procedure as simple hypothesis testing, but now the probabilities are conditional on the model specification, considering all possible parameter values rather than the parameters used by the model. Suppose we are interested in comparing two structural models $\mathcal{M}_1$ and $\mathcal{M}_2$ with an associated parameter vector $\boldsymbol{\theta}_i$ and prior probability $p(\mathcal{M}_i)$, $i = 1, 2$. Let $\mathbf{y}$ denote sample observations on variables in

the model. As in (12.5), the posterior odds ratio is given by

$$\begin{aligned}\frac{p(\mathcal{M}_1|\mathbf{y})}{p(\mathcal{M}_2|\mathbf{y})} &= \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}\frac{\int h(\boldsymbol{\theta}_1|\mathcal{M}_1)f(\mathbf{y}|\boldsymbol{\theta}_1,\mathcal{M}_1)d\boldsymbol{\theta}_1}{\int h(\boldsymbol{\theta}_2|\mathcal{M}_2)f(\mathbf{y}|\boldsymbol{\theta}_2,\mathcal{M}_2)d\boldsymbol{\theta}_2} \\ &= \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}\frac{f(\mathbf{y}|\mathcal{M}_1)}{f(\mathbf{y}|\mathcal{M}_2)}.\end{aligned}$$

Again, if the two models are equally likely a priori, the prior odds ratio is 1, and the Bayes factor can be interpreted according to Jeffreys' scale.

Appendix

12.A Numerical Approximation Methods

As we have seen, calculating an explicit form of posterior distributions often involves evaluation of high-dimensional integrals. In practice, the integrals of high-order functions are increasingly difficult to solve analytically, and, as a result, the posterior distribution may be intractable. To overcome this difficulty, numerical approximation methods are prominently used in the Bayesian analysis. This section explains three leading simulation techniques popularly used in the literature: the Importance Sampling, the Gibbs sampler and the Metropolis-Hastings algorithm. The latter two are in the class of the Markov chain Monte Carlo methods.

12.A.1 Importance Sampling

The idea behind the importance sampling is to obtain sample draws $\{\theta_i\}$ from some known distribution and assign weights to each draw so that the limiting distribution of the weighted sample converges to the target distribution.

Suppose we are interested in evaluating

$$E[h(\theta)] = \int h(\theta)f(\theta)d\theta \tag{12.A.1}$$

but $f(\theta)$ is not available as a sampling distribution. Let $I(\theta|\mu)$ denote a known distribution from which $\{\theta_i\}$ can be obtained. This distribution is called the importance sampler and μ represents its parameterization. Equation (12.A.1) can be rewritten as

$$\begin{aligned} E[h(\theta)] &= \int h(\theta)\frac{f(\theta)}{I(\theta)}I(\theta)d\theta \\ &= \int h(\theta)w(\theta)I(\theta)d\theta. \end{aligned} \tag{12.A.2}$$

where $w(\theta) \equiv f(\theta)/I(\theta)$. In (12.A.2), $w(\theta)$ serves to mitigate the direct influence of $I(\theta|\mu)$ on θ_i by assigning the weight or "importance" of different points in the sample space.

After a sample $\{\theta_i\}_{i=1}^N$ has been obtained from $I(\theta)$ rather than $f(\theta)$ for some large N, $E[h(\theta)]$ is approximated by the sample mean:

$$\hat{h} = \frac{1}{N}\sum_{i=1}^{N} h(\theta_i)w(\theta_i).$$

Geweke (1989) outlines criteria for choosing an importance sampler and formal diagnostics for the adequacy of a chosen sampler. Poor samplers tend to assign weights on only a small fraction of the sample rather than being approximately uniform, requiring a large number of draws to achieve convergence.

12.A.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are iterative sampling schemes to generate sample draws $\{x_i\}$ with the Markov property:

$$Pr(x_{i+1}|x_i, x_{i-1}, x_{i-2}, \cdots) = Pr(x_{i+1}|x_i) \text{ for all } i$$

where i indexes the Monte Carlo draws. These computer-intensive algorithms are particularly powerful in approximating multi-dimensional integrals with high accu-

racy. This section explains two widely used methods to simulate Markov chains: the Gibbs sampler and the Metropolis-Hastings algorithm. Further details are provided by Casella and George (1992) for the Gibbs sampler and Chib and Greenberg (1995) for the Metropolis-Hastings algorithm.

The Gibbs Sampler

Consider a q-dimensional vector of parameters $\boldsymbol{\theta}$ that is partitioned into k blocks, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_k)$, $k \leq q$. Suppose we wish to obtain the marginal distribution of the i^{th} block:

$$P(\boldsymbol{\theta}_i|\mathbf{x}) = \int \cdots \int P(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k|\mathbf{y}) d\boldsymbol{\theta}_1, \cdots, d\boldsymbol{\theta}_{i-1} d\boldsymbol{\theta}_{i+1} \cdots d\boldsymbol{\theta}_k$$

when the joint density $P(\boldsymbol{\theta}|\mathbf{y})$ is intractable. We assume that, for all i, the conditional posterior probability density for $\boldsymbol{\theta}_i$, $P(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i})$, is available as a sampling distribution where $\boldsymbol{\theta}_{-i}$ denotes all components of $\boldsymbol{\theta}$ excluding $\boldsymbol{\theta}_i$. The Gibbs sampler generates a Markov chain of random variables $\boldsymbol{\theta}_i^{(1)}, \cdots, \boldsymbol{\theta}_i^{(N)} \sim P(\boldsymbol{\theta}_i|\mathbf{y})$ by sampling from $P(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i})$.

The algorithm is initiated with some starting values $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \cdots, \boldsymbol{\theta}_k^{(0)}\right)$, and the subsequent sampling proceeds as follows.

(i) Draw a random observation $\boldsymbol{\theta}_1^{(1)}$ from $P(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(0)}, \boldsymbol{\theta}_3^{(0)}, \cdots, \boldsymbol{\theta}_k^{(0)})$.

(ii) Draw a random observation $\boldsymbol{\theta}_2^{(1)}$ from $P(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_3^{(0)}, \cdots, \boldsymbol{\theta}_k^{(0)})$.

$\vdots$

(iii) Draw a random observation $\boldsymbol{\theta}_k^{(1)}$ from $P(\boldsymbol{\theta}_k|\mathbf{y}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}, \cdots, \boldsymbol{\theta}_{k-1}^{(1)})$.

(iv) Return to step 1 and draw $\boldsymbol{\theta}_1^{(2)}$ from $P(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(1)}, \boldsymbol{\theta}_3^{(1)}, \cdots, \boldsymbol{\theta}_k^{(1)})$, and so on.

Repeating this process N times generates a Markov chain of length N, $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{N}$.

The effect of the fixed starting values $\boldsymbol{\theta}^{(0)}$ is eliminated by discarding some iterations at the beginning of the chain, a practice called a burn-in. With the remaining m observations, $P(\boldsymbol{\theta}_i|\mathbf{y})$ is approximated by

$$\hat{P}(\boldsymbol{\theta}_i|\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m} P(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}^{(j)}).$$

Alternatively, Gelfand and Smith (1990) suggest generating s independent Markov chains of length N and using the final value $\boldsymbol{\theta}^{(N)}$ from each sequence. Other approaches to exploiting convergence are discussed in Casella and George (1992).

Metropolis-Hastings Algorithm

The Gibbs sampler described above requires that the full conditional distribution is available in a tractable form as a sampling distribution for $\theta^{(i)}$. There are also MCMC methods for the case in which it is unavailable. The best known of these is the Metropolis-Hastings algorithm.

Suppose the target density $P(\boldsymbol{\theta}|\mathbf{x})$ is not available as a sampling distribution, but there is a known density $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$, where $\int g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})d\boldsymbol{\theta} = 1$, from which $\theta^{(i)}$ can be obtained. The Metropolis-Hastings algorithm is initialized with a starting value $\boldsymbol{\theta}^{(0)}$ and, given $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{i-1}$, $\boldsymbol{\theta}^{(i)}$ is obtained as follows:

(i) Draw a random sample $\tilde{\boldsymbol{\theta}}^{(i)}$ from $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$. This serves as a candidate for $\boldsymbol{\theta}^{(i)}$.

(ii) Define the probability of accepting $\tilde{\boldsymbol{\theta}}^{(i)}$ for $\boldsymbol{\theta}^{(i)}$:

$$\pi\left(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}\right) = \min\left(1, \frac{P(\tilde{\boldsymbol{\theta}}^{(i)}|\mathbf{x})}{P(\boldsymbol{\theta}^{(i-1)}|\mathbf{x})}\frac{g(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})}{g(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})}\right). \tag{12.A.3}$$

(iii) Draw a value δ from a uniform distribution on $[0, 1]$.

(iv) If $\pi(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}) > \delta$, set $\boldsymbol{\theta}^{(i)} = \tilde{\boldsymbol{\theta}}^{(i)}$; otherwise, discard $\tilde{\boldsymbol{\theta}}^{(i)}$ and draw a new candidate.

A sequence of accepted draws $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ is a Markov chain with transition probability $\lambda(\tilde{\boldsymbol{\theta}}^{(i)}|\tilde{\boldsymbol{\theta}}^{(i-1)}) = \pi(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)})g(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$ for $i = 1, \cdots, N$ and $\tilde{\theta}^{(i)} \neq \tilde{\theta}^{(i-1)}$. Under mild regularity conditions, this converges in distribution to $P(\boldsymbol{\theta}|\mathbf{x})$ as N increases.

Note that the calculation of $\pi(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{i-1})$ does not require knowledge about a normalizing constant in $P(\cdot)$ or $g(\cdot)$ since they appear in both the numerator and the denominator of (12.A.3) and simply cancel out. This is one of the attractive features of this algorithm for approximating posterior distributions since they are often known up to constant proportionality as in (12.2).

In application, the candidate-generating density $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$ can be specified in various ways. A random walk chain utilizes $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu}) = g_1(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}|\boldsymbol{\mu})$, and $\tilde{\boldsymbol{\theta}}^{(i)}$ follows the process $\tilde{\boldsymbol{\theta}}^{(i)} = \boldsymbol{\theta}^{(i-1)} + \boldsymbol{\varepsilon}_i$ where $\boldsymbol{\varepsilon}_i \sim g(\boldsymbol{\varepsilon})$ (Metropolis *et al*, 1953). Choices for g_1 include the multivariate normal and the multivariate-t densities. Alternatively, an independent chain draws a candidate independently of the last accepted draw. This is implemented by choosing a density that is independent across all Monte Carlo replications: $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu}) = g_2(\boldsymbol{\theta}|\boldsymbol{\mu})$ (Hastings, 1970). Another possibility is an autoregressive chain. A vector autoregressive process of order 1 follows $\tilde{\boldsymbol{\theta}}^{(i)} = \mathbf{a}+\mathbf{B}(\boldsymbol{\theta}^{(i-1)}-1)+\boldsymbol{\upsilon}_i$ drawn from the density $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu}) = g(\tilde{\boldsymbol{\theta}}^{(i)}-\mathbf{a}-\mathbf{B}(\boldsymbol{\theta}^{(i-1)}-1))$ where $\mathbf{a}$ is a vector, $\mathbf{B}$ is a matrix, and $\boldsymbol{\upsilon}_i \sim g(\boldsymbol{\upsilon})$ (Tierney, 1994).

12.B Application of the MCMC methods

In this section, we describe an application of the MCMC methods by nan Chen, Watanabe, and Yabu (1990). They propose a new method of data augmentation based on the Gibbs sampler to account for an endogeneity problem arising from the

use of time-aggregated data. Their application considers the estimation of the effects of foreign exchange interventions by a central bank.

Suppose the exchange rate movements and the central bank's intervention in the foreign exchange market can be represented by the following two-equation system:

$$s_{t,h} - s_{t,h-1} = \alpha I_{t,h} + \varepsilon_{t,h} \tag{12.B.4}$$

$$I_{t,h} = \beta(s_{t,h-1} - s_{t-1,h-1}) + \eta_{t,h} \tag{12.B.5}$$

for $t = 1, \cdots, T$ and $h = 1, \cdots, 24$, where $s_{t,h}$ is the log price of the domestic currency per unit of the foreign currency at hour h of day t , $I_{t,h}$ is the central bank's purchase of the domestic currency between $h-1$ and h of day t, $\varepsilon_{t,h} \sim i.i.d.N(0, \sigma_\varepsilon^2)$, and $\eta_{t,h} \sim i.i.d.N(0, \sigma_\eta^2)$. If $s_{t,h}$ and $I_{t,h}$ are both observable at the hourly frequency, we can obtain unbiased estimates of α and β by estimating (12.B.4) and (12.B.5) by OLS.

Suppose instead that $I_{t,h}$ is not observable and only the daily sum of hourly interventions $I_t \equiv \sum_{h=1}^{24} I_{t,h}$ is available. The above model can be transformed into a daily-frequency model by summing up both sides of (12.B.4) and (12.B.5) over h:

$$s_{t,24} - s_{t-1,24} = \alpha I_t + \varepsilon_t \tag{12.B.6}$$

$$I_t = \beta \sum_{h=1}^{24} (s_{t,h-1} - s_{t-1,h-1}) + \eta_t \tag{12.B.7}$$

where $s_{t,24} - s_{t-1,24} = \sum_{h=1}^{24}(s_{t,h} - s_{t,h-1})$ and $x_t = \sum_{h=1}^{24} x_{t,h}$ for $x = \{I, \varepsilon, \eta\}$. This model, however, suffers from an endogeneity problem, and the OLS estimates from (16.56) and (16.57) may be biased. To see this, consider a rise in $\varepsilon_{t,h}$. It increases $s_{t,h} - s_{t,h-1}$ in (12.B.4) and $I_{t,h+1}$ in (12.B.5) for $\beta > 0$, and hence I_t and ε_t are positively correlated. Alternatively, a rise in $\eta_{t,h}$ increases $I_{t,h}$ in (12.B.5) and

appreciates the currency in (12.B.4) for $\alpha < 0$, implying that $\sum(s_{t,h} - s_{t,h-1})$ and η_t are negatively correlated.

Recognizing this problem, nan Chen, Watanabe, and Yabu (1990) propose an algorithm to obtain a posterior distribution of the parameters using the Gibbs sampler. They first introduce an auxiliary variable $I_{t,h}$ to substitute the unobserved hourly interventions, and assume a flat distribution as the priors of α and β, and distributions $IG\left(\frac{\upsilon_\varepsilon}{2}, \frac{\delta_\varepsilon}{2}\right)$ and $IG\left(\frac{\upsilon_\eta}{2}, \frac{\delta_\eta}{2}\right)$ for σ_ε^2 and σ_η^2. The algorithm proceeds as follows.[1]

(i) Generate α conditional on $s_{t,h}$, $I_{t,h}$, and σ_ε^2. The posterior distribution is $\alpha \sim N(\phi_s, \omega_s)$ where $\phi_s = \sum I_{t,h}(s_{t,h} - s_{t,h-1}) / \sum I_{t,h}^2$ and $\omega_s = \sigma_\varepsilon^2 / \sum I_{t,h}^2$.

(ii) Generate β conditional on $s_{t,h}$, $I_{t,h}$, and σ_η^2. The posterior distribution is $\beta \sim N(\phi_I, \omega_I)$ where $\phi_I = \sum I_{t,h}(s_{t,h-1} - s_{t-1,h-1}) / \sum(s_{t,h-1} - s_{t-1,h-1})^2$, and $\omega_I = \sigma_\eta^2 / \sum(s_{t,h-1} - s_{t-1,h-1})^2$.

(iii) Generate σ_ε^2 conditional on $s_{t,h}$, $I_{t,h}$, and α. The posterior distribution is $\sigma_\varepsilon^2 \sim IG(\frac{\upsilon_\varepsilon + T}{2}, \frac{\delta_\varepsilon + RRS_s}{2})$ where $RRS_s = \sum(s_{t,h} - s_{t,h-1} - \alpha I_{t,h})^2$.

(iv) Generate σ_η^2 conditional on $s_{t,h}$, $I_{t,h}$, and β. The posterior distribution is $\sigma_\eta^2 \sim IG(\frac{\upsilon_\eta + T}{2}, \frac{\delta_\eta + RRS_I}{2})$ where $RRS_I = \sum(I_{t,h} - \beta(s_{t,h-1} - s_{t-1,h-1}))^2$.

(v) Generate $I_{t,h}$ conditional on $s_{t,h}$, I_t, α, β, σ_ε^2, and σ_η^2. The posterior distribution is derived as follows.

If I_t is unknown, the posterior distribution is given by $(I_{t,1}, \cdots, I_{t,24})' \sim N(\mathbf{\Xi}_t, \mathbf{\Phi})$ where $\mathbf{\Xi}_t = (\xi_{t,1}, \cdots, \xi_{t,24})'$ and $\mathbf{\Phi} = diag(\psi, \cdots, \psi)$ with $\psi = (\frac{1}{\sigma_\eta^2} + \frac{\alpha^2}{\sigma_\varepsilon^2})^{-1}$ and $\xi_{t,h} = \left(\psi \frac{1}{\sigma_\eta^2}\right)[\beta(s_{t,h-1} - s_{t-1,h-1})] + \left(\psi \frac{\alpha^2}{\sigma_\varepsilon^2}\right)[\alpha^{-1}(s_{t,h} - s_{t,h-1})]$. Since I_t is known, consider the posterior distribution $(I_{t,1}, \cdots, I_{t,23}, I_t)' \sim N(\mathbf{\Xi}_t^*, \mathbf{\Phi}^*)$ where $\mathbf{\Xi}_t^* = \mathbf{B}\mathbf{\Xi}_t$

[1]The summations indicate $\sum \equiv \sum_{t=1}^{T} \sum_{h=1}^{24}$.

and $\mathbf{\Phi}^* = \mathbf{B}\mathbf{\Phi}\mathbf{B}'$ with

$$\underset{(24\times24)}{\mathbf{B}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

Partition $\mathbf{\Xi}_t^*$ and $\mathbf{\Phi}^*$ as follows:

$$\mathbf{\Xi}_t^* = \begin{bmatrix} \underset{(23\times1)}{\mathbf{\Xi}_{t,1}^*} \\ \\ \underset{(1\times1)}{\mathbf{\Xi}_{t,2}^*} \end{bmatrix}, \mathbf{\Phi}^* = \begin{bmatrix} \underset{(23\times23)}{\mathbf{\Phi}_{11}^*} & \underset{(23\times1)}{\mathbf{\Phi}_{12}^*} \\ \\ \underset{(1\times23)}{\mathbf{\Phi}_{21}^*} & \underset{(1\times1)}{\mathbf{\Phi}_{22}^*} \end{bmatrix}.$$

The posterior distribution of $(I_{t,1} \cdots, I_{t,23})$ conditional on I_t is given by

$$(I_{t,1} \cdots, I_{t,23}|I_t)' \sim N(\mathbf{\Xi}_{t,1}^* + \mathbf{\Phi}_{12}^*(\mathbf{\Phi}_{22}^*)^{-1}(I_t - \mathbf{\Xi}_{t,2}^*), \mathbf{\Phi}_{11}^* - \mathbf{\Phi}_{12}^*(\mathbf{\Phi}_{22}^*)^{-1}\mathbf{\Phi}_{21}^*).$$

After $(I_{t,1} \cdots, I_{t,23})$ has been generated from this posterior distribution, $I_{t,24}$ is obtained from $I_{t,24} = I_t - \sum_{h=1}^{23} I_{t,h}$.

Applying this method to the Japanese data, nan Chen, Watanabe, and Yabu (1990) generate three Markov chains of the length 2,000 after discarding the first 2,000 draws in each chain as a burn-in phase. They obtain the point estimate of each parameter using the mean of the generated posterior distribution, and find that the effect of intervention is more than twice as large as the magnitude estimated by OLS using daily observations, suggesting the quantitative significance of the endogeneity problem.

References

Casella, G., and E. I. George (1992): "Explaining the Gibbs Sampler," *American Statistician*, 46(3), 167–174.

Chib, S., and E. Greenberg (1995): "Understanding the Metropolis-Hastings Algorithm," *American Statistician*, 49(4), 327–335.

DeJong, D. N., and C. Dave (2007): *Structural Macroeconometrics*. Princeton University Press.

Gelfand, A. E., and A. F. M. Smith (1990): "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85(410), 398–409.

Geweke, J. (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57(6), 1317–1339.

Hastings, W. K. (1970): "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Jeffreys, H. (1961): *Theory of Probability*. Oxford University Press, New York.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T. Lee (1985): *The Theory and Practice of Econometrics*. Wiley, New York, 2nd edn.

Lubik, T., and F. Schorfheide (2007): "Do Central Banks Respond to Exchange Rate Movements? A Structural Investigation," *Journal of Monetary Economics*, 54(4), 1069–1087.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953): "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21(6), 1087–1092.

nan Chen, C., T. Watanabe, and T. Yabu (1990): "A New Method for Identifying the Effects of Foreign Exchange Interventions," IMES Discussion Paper Series, No. 2009-E-6, Bank of Japan.

Rabanal, P., and J. F. Rubio-Ramirez (2005): "Comparing New Keynesian Models of the Business Cycle: A Bayesian Approach," *Journal of Monetary Economics*, 52(6), 1151–1166.

Tierney, L. (1994): "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22(4), 1701–1728.

Zellner, A. (1996): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York, Reprint of 1971 ed.