6 Bayesian Estimation — Examples

6.1 Heteroscedasticity Model

In Section 6.1, Tanizaki and Zhang (2001) is re-computed using the random number generators.

Here, we show how to use Bayesian approach in the multiplicative heteroscedasticity model discussed by Harvey (1976).

The Gibbs sampler and the Metropolis-Hastings (MH) algorithm are applied to the multiplicative heteroscedasticity model, where some sampling densities are considered in the MH algorithm.

We carry out Monte Carlo study to examine the properties of the estimates via Bayesian approach and the traditional counterparts such as the modified two-step estimator (M2SE) and the maximum likelihood estimator (MLE).

The results of Monte Carlo study show that the sampling density chosen here is suitable, and Bayesian approach shows better performance than the traditional counterparts in the criterion of the root mean square error (RMSE) and the interquartile range (IR).

6.1.1 Introduction

For the heteroscedasticity model, we have to estimate both the regression coefficients and the heteroscedasticity parameters.

In the literature of heteroscedasticity, traditional estimation techniques include the two-step estimator (2SE) and the maximum likelihood estimator (MLE).

Harvey (1976) showed that the 2SE has an inconsistent element in the heteroscedasticity parameters and furthermore derived the consistent estimator based on the 2SE, which is called the modified two-step estimator (M2SE).

These traditional estimators are also examined in Amemiya (1985), Judge, Hill, Griffiths and Lee (1980) and Greene (1997).

Ohtani (1982) derived the Bayesian estimator (BE) for a heteroscedasticity linear model.

Using a Monte Carlo experiment, Ohtani (1982) found that among the Bayesian estimator (BE) and some traditional estimators, the Bayesian estimator (BE) shows the best properties in the mean square error (MSE) criterion.

Because Ohtani (1982) obtained the Bayesian estimator by numerical integration, it is not easy to extend to the multi-dimensional cases of both the regression coefficient and the heteroscedasticity parameter.

Recently, Boscardin and Gelman (1996) developed a Bayesian approach in which a Gibbs sampler and the Metropolis-Hastings (MH) algorithm are used to estimate the parameters of heteroscedasticity in the linear model.

They argued that through this kind of Bayesian approach, we can average over our uncertainty in the model parameters instead of using a point estimate via the traditional estimation techniques.

Their modeling for the heteroscedasticity, however, is very simple and limited. Their choice of the heteroscedasticity is $V(y_i) = \sigma^2 w_i^{-\theta}$, where w_i are known "weights" for the problem and θ is an unknown parameter.

In addition, they took only one candidate for the sampling density used in the MH algorithm and compared it with 2SE.

In Section 6.1, we also consider Harvey's (1976) model of multiplicative heteroscedasticity.

This modeling is very flexible, general, and includes most of the useful formulations for heteroscedasticity as special cases.

The Bayesian approach discussed by Ohtani (1982) and Boscardin and Gelman (1996) can be extended to the multi-dimensional and more complicated cases, using the model introduced here.

The Bayesian approach discussed here includes the MH within Gibbs algorithm, where through Monte Carlo studies we examine two kinds of candidates for the sampling density in the MH algorithm and compare the Bayesian approach with the two traditional estimators, i.e., M2SE and MLE, in the criterion of the root mean square error (RMSE) and the interquartile range (IR).

We obtain the results that the Bayesian estimator significantly has smaller RMSE and IR than M2SE and MLE at least for the heteroscedasticity parameters.

Thus, the results of the Monte Carlo study show that the Bayesian approach performs better than the traditional estimators.

6.1.2 Multiplicative Heteroscedasticity Regression Model

The multiplicative heteroscedasticity model discussed by Harvey (1976) can be shown as follows:

$$y_t = X_t\beta + u_t, \qquad u_t \sim N(0, \sigma_t^2), \tag{7}$$

$$\sigma_t^2 = \sigma^2 \exp(q_t\alpha), \tag{8}$$

for $t = 1, 2, \cdots, n$, where y_t is the tth observation, X_t and q_t are the tth $1 \times k$ and $1 \times (J-1)$ vectors of explanatory variables, respectively.
β and α are vectors of unknown parameters.

The model given by equations (7) and (8) includes several special cases such as the model in Boscardin and Gelman (1996), in which $q_t = \log w_t$ and $\theta = -\alpha$.

As shown in Greene (1997), there is a useful simplification of the formulation.

Let $z_t = (1, q_t)$ and $\gamma = (\log \sigma^2, \alpha')'$, where z_t and γ denote $1 \times J$ and $J \times 1$ vectors.

Then, we can simply rewrite equation (8) as:

$$\sigma_t^2 = \exp(\mathrm{z}_t \gamma). \tag{9}$$

Note that $\exp(\gamma_1)$ provides σ^2, where γ_1 denotes the first element of γ.

As for the variance of u_t, hereafter we use (9), rather than (8).

The generalized least squares (GLS) estimator of β, denoted by $\hat{\beta}_{GLS}$, is given by:

$$\hat{\beta}_{GLS} = \Big(\sum_{t=1}^{n} \exp(-z_t\gamma)X_t'X_t\Big)^{-1} \sum_{t=1}^{n} \exp(-z_t\gamma)X_t'y_t, \tag{10}$$

where $\hat{\beta}_{GLS}$ depends on γ, which is the unknown parameter vector.

To obtain the feasible GLS estimator, we need to replace γ by its consistent estimate.

We have two traditional consistent estimators of γ, i.e., M2SE and MLE, which are briefly described as follows.

Modified Two-Step Estimator (M2SE): First, define the ordinary least squares (OLS) residual by $e_t = y_t - X_t\hat{\beta}_{OLS}$, where $\hat{\beta}_{OLS}$ represents the OLS estimator, i.e., $\hat{\beta}_{OLS} = (\sum_{t=1}^{n} X_t'X_t)^{-1} \sum_{t=1}^{n} X_t'y_t$.

For 2SE of γ, we may form the following regression:

$$\log e_t^2 = z_t\gamma + v_t.$$

The OLS estimator of γ applied to the above equation leads to the 2SE of γ, because e_t is obtained by OLS in the first step.

Thus, the OLS estimator of γ gives us 2SE, denoted by $\hat{\gamma}_{2SE}$, which is given by:

$$\hat{\gamma}_{2SE} = (\sum_{t=1}^{n} z_t' z_t)^{-1} \sum_{t=1}^{n} z_t' \log e_t^2.$$

A problem with this estimator is that v_t, $t = 1, 2, \cdots, n$, have non-zero means and are heteroscedastic.

If e_t converges in distribution to u_t, the v_t will be asymptotically independent with mean $\mathrm{E}(v_t) = -1.2704$ and variance $\mathrm{V}(v_t) = 4.9348$, which are shown in Harvey (1976).

Then, we have the following mean and variance of $\hat{\gamma}_{2SE}$:

$$\mathrm{E}(\hat{\gamma}_{2SE}) = \gamma - 1.2704(\sum_{t=1}^{n} z_t' z_t)^{-1} \sum_{t=1}^{n} z_t', \tag{11}$$

$$\mathrm{V}(\hat{\gamma}_{2SE}) = 4.9348(\sum_{t=1}^{n} z_t' z_t)^{-1}.$$

For the second term in equation (11), the first element is equal to -1.2704 and the remaining elements are zero, which can be obtained by simple calculation. Therefore, the first element of $\hat{\gamma}_{2SE}$ is biased but the remaining elements are still unbiased.

To obtain a consistent estimator of γ_1, we consider M2SE of γ, denoted by $\hat{\gamma}_{M2SE}$, which is given by:

$$\hat{\gamma}_{M2SE} = \hat{\gamma}_{2SE} + 1.2704(\sum_{t=1}^{n} z_t' z_t)^{-1} \sum_{t=1}^{n} z_t'.$$

Let Σ_{M2SE} be the variance of $\hat{\gamma}_{M2SE}$.

Then, Σ_{M2SE} is represented by:

$$\Sigma_{M2SE} \equiv \mathrm{V}(\hat{\gamma}_{M2SE}) = \mathrm{V}(\hat{\gamma}_{2SE}) = 4.9348(\sum_{t=1}^{n} z_t' z_t)^{-1}.$$

The first element of $\hat{\gamma}_{2SE}$ and $\hat{\gamma}_{M2SE}$ corresponds to the estimate of σ^2, which value does not influence $\hat{\beta}_{GLS}$.

Since the remaining elements of $\hat{\gamma}_{2SE}$ are equal to those of $\hat{\gamma}_{M2SE}$, $\hat{\beta}_{2SE}$ is equivalent to $\hat{\beta}_{M2SE}$, where $\hat{\beta}_{2SE}$ and $\hat{\beta}_{M2SE}$ denote 2SE and M2SE of β, respectively.
Note that $\hat{\beta}_{2SE}$ and $\hat{\beta}_{M2SE}$ can be obtained by substituting $\hat{\gamma}_{2SE}$ and $\hat{\gamma}_{M2SE}$ into γ in (10).

Maximum Likelihood Estimator (MLE): The density of $Y_n = (y_1, y_2, \cdots, y_n)$ based on (7) and (9) is:

$$f(Y_n|\beta,\gamma) \propto \exp\left(-\frac{1}{2}\sum_{t=1}^{n}\Big(\exp(-z_t\gamma)(y_t - X_t\beta)^2 + z_t\gamma\Big)\right), \tag{12}$$

which is maximized with respect to β and γ, using the method of scoring.

That is, given values for $\beta^{(j)}$ and $\gamma^{(j)}$, the method of scoring is implemented by the following iterative procedure:

$$\beta^{(j)} = \Big(\sum_{t=1}^{n} \exp(-z_t\gamma^{(j-1)})X_t'X_t\Big)^{-1} \sum_{t=1}^{n} \exp(-z_t\gamma^{(j-1)})X_t'y_t,$$

$$\gamma^{(j)} = \gamma^{(j-1)} + 2(\sum_{t=1}^{n} z_t'z_t)^{-1}\frac{1}{2} \sum_{t=1}^{n} z_t'\Big(\exp(-z_t\gamma^{(j-1)})e_t^2 - 1\Big),$$

for $j = 1, 2, \cdots$, where $e_t = y_t - X_t\beta^{(j-1)}$.

The starting value for the above iteration may be taken as $(\beta^{(0)}, \gamma^{(0)}) = (\hat{\beta}_{OLS}, \hat{\gamma}_{2SE})$, $(\hat{\beta}_{2SE}, \hat{\gamma}_{2SE})$ or $(\hat{\beta}_{M2SE}, \hat{\gamma}_{M2SE})$.

Let $\theta = (\beta, \gamma)$.

The limit of $\theta^{(j)} = (\beta^{(j)}, \gamma^{(j)})$ gives us the MLE of θ, which is denoted by $\hat{\theta}_{MLE} = (\hat{\beta}_{MLE}, \hat{\gamma}_{MLE})$.

Based on the information matrix, the asymptotic covariance matrix of $\hat{\theta}_{MLE}$ is represented by:

$$
\begin{aligned}
\mathrm{V}(\hat{\theta}_{MLE}) &= \left(-\mathrm{E}\left(\frac{\partial^2 \log f(Y_n|\theta)}{\partial\theta\partial\theta'} \right) \right)^{-1} \\
&= \begin{pmatrix} \left(\sum_{t=1}^{n} \exp(-z_t\gamma) X_t' X_t\right)^{-1} & 0 \\ 0 & 2(\sum_{t=1}^{n} z_t' z_t)^{-1} \end{pmatrix}.
\end{aligned} \tag{13}
$$

Thus, from (13), asymptotically there is no correlation between $\hat{\beta}_{MLE}$ and $\hat{\gamma}_{MLE}$, and furthermore the asymptotic variance of $\hat{\gamma}_{MLE}$ is represented by: $\Sigma_{MLE} \equiv \mathrm{V}(\hat{\gamma}_{MLE}) =$

$2(\sum_{t=1}^{n} z_t' z_t)^{-1}$, which implies that $\hat{\gamma}_{M2SE}$ is asymptotically inefficient because $\Sigma_{M2SE} - \Sigma_{MLE}$ is positive definite.

Remember that the variance of $\hat{\gamma}_{M2SE}$ is given by: $\mathrm{V}(\hat{\gamma}_{M2SE}) = 4.9348(\sum_{t=1}^{n} z_t' z_t)^{-1}$.

6.1.3 Bayesian Estimation

We assume that the prior distributions of the parameters β and γ are noninformative, which are represented by:

$$f_\beta(\beta) = \text{constant}, \qquad f_\gamma(\gamma) = \text{constant}. \tag{14}$$

Combining the prior distributions (14) and the likelihood function (12), the posterior distribution $f_{\beta\gamma}(\beta, \gamma|y)$ is obtained as follows:

$$f_{\beta\gamma}(\beta, \gamma|Y_n) \propto \exp\left(-\frac{1}{2}\sum_{t=1}^{n}\Big(\exp(-z_t\gamma)(y_t - X_t\beta)^2 + z_t\gamma\Big)\right).$$

The posterior means of β and γ are not operationally obtained.

Therefore, by generating random draws of β and γ from the posterior density $f_{\beta\gamma}(\beta, \gamma|Y_n)$, we consider evaluating the mathematical expectations as the arithmetic averages based on the random draws.

Now we utilize the Gibbs sampler, which has been introduced in Section 5.7.5, to sample random draws of β and γ from the posterior distribution.

Then, from the posterior density $f_{\beta\gamma}(\beta, \gamma|Y_n)$, we can derive the following two conditional densities:

$$f_{\gamma|\beta}(\gamma|\beta, Y_n) \propto \exp\left(-\frac{1}{2}\sum_{t=1}^{n}\Big(\exp(-z_t\gamma)(y_t - X_t\beta)^2 + z_t\gamma\Big)\right), \tag{15}$$

$$f_{\beta|\gamma}(\beta|\gamma, Y_n) = N(B_1, H_1), \tag{16}$$

where

$$H_1^{-1} = \sum_{t=1}^{n}\exp(-z_t\gamma)X_t'X_t, \qquad B_1 = H_1\sum_{t=1}^{n}\exp(-z_t\gamma)X_t'y_t.$$

Sampling from (16) is simple since it is a k-variate normal distribution with mean B_1 and variance H_1.

However, since the J-variate distribution (15) does not take the form of any standard density, it is not easy to sample from (15).

In this case, the MH algorithm discussed in Section 5.7.3 can be used within the Gibbs sampler.

See Tierney (1994) and Chib and Greeberg (1995) for a general discussion.

Let γ_{i-1} be the $(i-1)$th random draw of γ and γ^* be a candidate of the ith random draw of γ.

The MH algorithm utilizes another appropriate distribution function $f_*(\gamma|\gamma_i)$, which is called the sampling density or the proposal density.

Let us define the acceptance rate $\omega(\gamma_{i-1}, \gamma^*)$ as:

$$\omega(\gamma_{i-1}, \gamma^*) = \min\left(\frac{f_{\gamma|\beta}(\gamma^*|\beta_{i-1}, Y_n)/f_*(\gamma^*|\gamma_{i-1})}{f_{\gamma|\beta}(\gamma_{i-1}|\beta_{i-1}, Y_n)/f_*(\gamma_{i-1}|\gamma^*)},\ 1\right).$$

The sampling procedure based on the MH algorithm within Gibbs sampling is as follows:

(i) Set the initial value β_{-M}, which may be taken as $\hat{\beta}_{M2SE}$ or $\hat{\beta}_{MLE}$.

(ii) Given β_{i-1}, generate a random draw of γ, denoted by γ_i, from the conditional density $f_{\gamma|\beta}(\gamma|\beta_{i-1}, Y_n)$, where the MH algorithm is utilized for random number generation because it is not easy to generate random draws of γ from (15).

The Metropolis-Hastings algorithm is implemented as follows:

(a) Given γ_{i-1}, generate a random draw γ^* from $f_*(\cdot|\gamma_{i-1})$ and compute the acceptance rate $\omega(\gamma_{i-1}, \gamma^*)$.

We will discuss later about the sampling density $f_*(\gamma|\gamma_{i-1})$.

(b) Set $\gamma_i = \gamma^*$ with probability $\omega(\gamma_{i-1}, \gamma^*)$ and $\gamma_i = \gamma_{i-1}$ otherwise,

(iii) Given γ_i, generate a random draw of β, denoted by β_i, from the conditional density $f_{\beta|\gamma}(\beta|\gamma_i, Y_n)$, which is $\beta|\gamma_i, Y_n \sim N(B_1, H_1)$ as shown in (16).

(iv) Repeat (ii) and (iii) for $i = -M+1, -M+2, \cdots, N$.

Note that the iteration of Steps (ii) and (iii) corresponds to the Gibbs sampler, which iteration yields random draws of β and γ from the joint density $f_{\beta\gamma}(\beta, \gamma|Y_n)$ when i is large enough.

It is well known that convergence of the Gibbs sampler is slow when β is highly correlated with γ.

That is, a large number of random draws have to be generated in this case.

Therefore, depending on the underlying joint density, we have the case where the Gibbs sampler does not work at all.

For example, see Chib and Greenberg (1995) for convergence of the Gibbs sampler.

In the model represented by (7) and (8), however, there is asymptotically no correlation between $\hat{\beta}_{MLE}$ and $\hat{\gamma}_{MLE}$, as shown in (13).

It might be expected that correlation between $\hat{\beta}_{MLE}$ and $\hat{\gamma}_{MLE}$ is not too high even in the small sample.

Therefore, it might be appropriate to consider that the Gibbs sampler works well in this model.

In Step (ii), the sampling density $f_*(\gamma|\gamma_{i-1})$ is utilized.

We consider the multivariate normal density function for the sampling distribution, which is discussed as follows.

Choice of the Sampling Density in Step (ii): Several generic choices of the sampling density are discussed by Tierney (1994) and Chib and Greenberg (1995). Here, we take $f_*(\gamma|\gamma_{i-1}) = f_*(\gamma)$ as the sampling density, which is called the independence chain because the sampling density is not a function of γ_{i-1}.

We consider taking the multivariate normal sampling density in the independence MH algorithm, because of its simplicity.

Therefore, $f_*(\gamma)$ is taken as follows:

$$f_*(\gamma) = N(\gamma^+, c^2\Sigma^+), \tag{17}$$

which represents the J-variate normal distribution with mean γ^+ and variance $c^2\Sigma^+$.

The tuning parameter c is introduced into the sampling density (17).

We have mentioned that for the independence chain (Sampling Density I) the sampling density with the variance which gives us the maximum acceptance probability is not necessarily the best choice.

From some Monte Carlo experiments, we have obtained the result that the sampling density with the 1.5 – 2.5 times larger standard error is better than that with the standard error which maximizes the acceptance probability.

Therefore, $c = 2$ is taken in the next section, and it is the larger value than the c which gives us the maximum acceptance probability.

This detail discussion is given in Section 6.1.4.

Thus, the sampling density of γ is normally distributed with mean γ^+ and variance $c^2\Sigma^+$.

As for (γ^+, Σ^+), in the next section we choose one of $(\hat{\gamma}_{M2SE}, \Sigma_{M2SE})$ and $(\hat{\gamma}_{MLE}, \Sigma_{MLE})$ from the criterion of the acceptance rate.

As shown in Section 2, both of the two estimators $\hat{\gamma}_{M2SE}$ and $\hat{\gamma}_{MLE}$ are consistent estimates of γ.

Therefore, it might be very plausible to consider that the sampling density is distributed around the consistent estimates.

Bayesian Estimator: From the convergence theory of the Gibbs sampler and the MH algorithm, as i goes to infinity we can regard γ_i and β_i as random draws from the target density $f_{\beta\gamma}(\beta, \gamma|Y_n)$.

Let M be a sufficiently large number. γ_i and β_i for $i = 1, 2, \cdots, N$ are taken as the random draws from the posterior density $f_{\beta\gamma}(\beta, \gamma|Y_n)$.

Therefore, the Bayesian estimators $\hat{\gamma}_{BZZ}$ and $\hat{\beta}_{BZZ}$ are given by:

$$\hat{\gamma}_{BZZ} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i, \qquad \hat{\beta}_{BZZ} = \frac{1}{N}\sum_{i=1}^{N}\beta_i,$$

where we read the subscript BZZ as the Bayesian estimator which uses the multivariate normal sampling density with mean $\hat{\gamma}_{ZZ}$ and variance Σ_{ZZ}. ZZ takes M2SE or MLE.

We consider two kinds of candidates of the sampling density for the Bayesian estimator, which are denoted by BM2SE and BMLE.

Thus, in Section 6.1.4, we compare the two Bayesian estimators (i.e, BM2SE and BMLE) with the two traditional estimators (i.e., M2SE and MLE).

6.1.4 Monte Carlo Study

Setup of the Model: In the Monte Carlo study, we consider using the artificially simulated data, in which the true data generating process (DGP) is presented in

Judge, Hill, Griffiths and Lee (1980, p.156).

The DGP is defined as:

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_t, \tag{18}$$

where u_t, $t = 1, 2, \cdots, n$, are normally and independently distributed with $\mathrm{E}(u_t) = 0$, $\mathrm{E}(u_t^2) = \sigma_t^2$ and,

$$\sigma_t^2 = \exp(\gamma_1 + \gamma_2 x_{2,t}), \qquad \text{for } t = 1, 2, \cdots, n. \tag{19}$$

As it is discussed in Judge, Hill, Griffiths and Lee (1980), the parameter values are set to be $(\beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2) = (10, 1, 1, -2, 0.25)$.

From (18) and (19), Judge, Hill, Griffiths and Lee (1980, pp.160 – 165) generated one hundred samples of y with $n = 20$.

In the Monte Carlo study, we utilize $x_{2,t}$ and $x_{3,t}$ given in Judge, Hill, Griffiths and Lee (1980, pp.156), which is shown in Table 1, and generate G samples of y_t given the X_t for $t = 1, 2, \cdots, n$.

That is, we perform G simulation runs for each estimator, where $G = 10^4$ is taken.

The simulation procedure is as follows:

(i) Given γ and $x_{2,t}$ for $t = 1, 2, \cdots, n$, generate random numbers of u_t for $t = 1, 2, \cdots, n$, based on the assumptions: $u_t \sim N(0, \sigma_t^2)$, where $(\gamma_1, \gamma_2) =$

Table 1: The Exogenous Variables $x_{1,t}$ and $x_{2,t}$

t	1	2	3	4	5	6	7	8	9	10
$x_{2,t}$	14.53	15.30	15.92	17.41	18.37	18.83	18.84	19.71	20.01	20.26
$x_{3,t}$	16.74	16.81	19.50	22.12	22.34	17.47	20.24	20.37	12.71	22.98
t	11	12	13	14	15	16	17	18	19	20
$x_{2,t}$	20.77	21.17	21.34	22.91	22.96	23.69	24.82	25.54	25.63	28.73
$x_{3,t}$	19.33	17.04	16.74	19.81	31.92	26.31	25.93	21.96	24.05	25.66

$(-2, 0.25)$ and $\sigma_t^2 = \exp(\gamma_1 + \gamma_2 x_{2,t})$ are taken.

(ii) Given β, $(x_{2,t}, x_{3,t})$ and u_t for $t = 1, 2, \cdots, n$, we obtain a set of data y_t, $t = 1, 2, \cdots, n$, from equation (18), where $(\beta_1, \beta_2, \beta_3) = (10, 1, 1)$ is assumed.

(iii) Given (y_t, X_t) for $t = 1, 2, \cdots, n$, perform M2SE, MLE, BM2SE and BMLE discussed in Sections 6.1.2 and 6.1.3 in order to obtain the estimates of $\theta = (\beta, \gamma)$, denoted by $\hat{\theta}$.

Note that $\hat{\theta}$ takes $\hat{\theta}_{M2SE}$, $\hat{\theta}_{MLE}$, $\hat{\theta}_{BM2SE}$ and $\hat{\theta}_{BMLE}$.

(iv) Repeat (i) – (iii) G times, where $G = 10^4$ is taken as mentioned above.

(v) From G estimates of θ, compute the arithmetic average (AVE), the root mean square error (RMSE), the first quartile (25%), the median (50%), the third quartile (75%) and the interquartile range (IR) for each estimator.

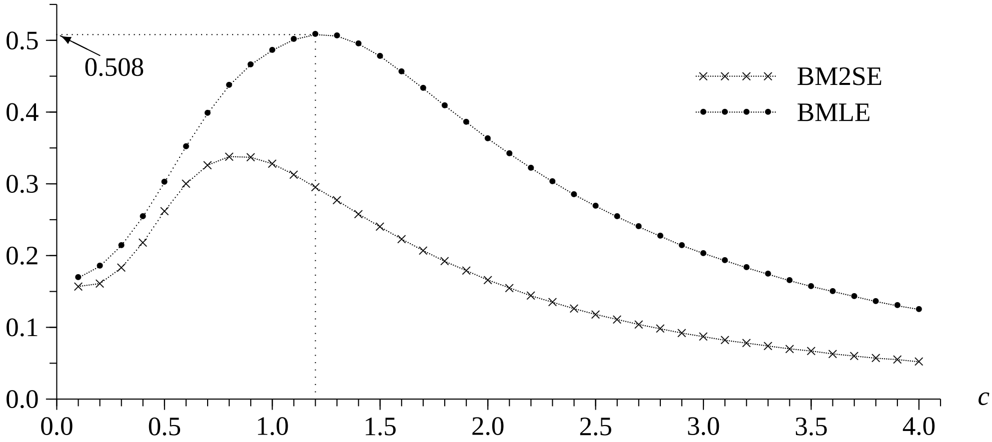
AVE and RMSE are obtained as follows:

$$\mathrm{AVE} = \frac{1}{G}\sum_{g=1}^{G}\hat{\theta}_j^{(g)}, \qquad \mathrm{RMSE} = \Big(\frac{1}{G}\sum_{g=1}^{G}(\hat{\theta}_j^{(g)} - \theta_j)^2\Big)^{1/2},$$

for $j = 1, 2, \cdots, 5$, where θ_j denotes the jth element of θ and $\hat{\theta}_j^{(g)}$ represents the j-element of $\hat{\theta}$ in the gth simulation run.

As mentioned above, $\hat{\theta}$ denotes the estimate of θ, where $\hat{\theta}$ takes $\hat{\theta}_{M2SE}$, $\hat{\theta}_{MLE}$, $\hat{\theta}_{BM2SE}$ and $\hat{\theta}_{BMLE}$.

Figure 2: Acceptance Rates in Average: $M = 5000$ and $N = 10^4$



Choice of (γ^+, Σ^+) and c: For the Bayesian approach, depending on (γ^+, Σ^+) we have BM2SE and BMLE, which denote the Bayesian estimators using the multivariate normal sampling density whose mean and covariance matrix are calibrated on the basis of M2SE or MLE.

We consider the following sampling density: $f_*(\gamma) = N(\gamma^+, c^2\Sigma^+)$, where c denotes the tuning parameter and (γ^+, Σ^+) takes $(\gamma_{M2SE}, \Sigma_{M2SE})$ or $(\gamma_{MLE}, \Sigma_{MLE})$.

Generally, for choice of the sampling density, the sampling density should not have too large variance and too small variance.

Chib and Greenberg (1995) pointed out that if standard deviation of the sampling density is too low, the Metropolis steps are too short and move too slowly within the target distribution; if it is too high, the algorithm almost always rejects and stays in the same place.

The sampling density should be chosen so that the chain travels over the support of the target density.

First, we consider choosing (γ^+, Σ^+) and c which maximizes the arithmetic average of the acceptance rates obtained from G simulation runs.

The results are in Figure 2, where $n = 20$, $M = 5000$, $N = 10^4$, $G = 10^4$ and $c = 0.1, 0.2, \cdots, 4.0$ are taken (choice of N and M is discussed in Appendix of

Section 6.1.6).

In the case of $(\gamma^+, \Sigma^+) = (\gamma_{MLE}, \Sigma_{MLE})$ and $c = 1.2$, the acceptance rate in average is 0.5078, which gives us the largest one.

It is important to reduce positive correlation between γ_i and γ_{i-1} and keep randomness.

Therefore, $(\gamma^+, \Sigma^+) = (\gamma_{MLE}, \Sigma_{MLE})$ is adopted, rather than $(\gamma^+, \Sigma^+) = (\gamma_{M2SE}, \Sigma_{M2SE})$, because BMLE has a larger acceptance probability than BM2SE for all c (see Figure 2).

However, the sampling density with the largest acceptance probability is not necessarily the best choice.

We have the result that the optimal standard error should be 1.5 – 2.5 times larger than the standard error which gives us the largest acceptance probability.

Here, $(\gamma^+, \Sigma^+) = (\gamma_{MLE}, \Sigma_{MLE})$ and $c = 2$ are taken.

When c is larger than 2, both the estimates and their standard errors become stable although here we do not show these facts.

Therefore, in this Monte Carlo study, $f_*(\gamma) = N(\gamma_{MLE}, 2^2\Sigma_{MLE})$ is chosen for the sampling density.

Hereafter, we compare BMLE with M2SE and MLE (i.e., we do not consider

BM2SE anymore).

As for computational CPU time, the case of $n = 20$, $M = 5000$, $N = 10^4$ and $G = 10^4$ takes about 76 minutes for each of $c = 0.1, 0.2, \cdots, 4.0$ and each of BM2SE and BMLE, where Dual Pentium III 1GHz CPU, Microsoft Windows 2000 Professional Operating System and Open Watcom FORTRAN 77/32 Optimizing Compiler (Version 1.0) are utilized.

Note that WATCOM Fortran 77 Compiler is downloaded from

`http://www.openwatcom.org/`.

Results and Discussion: Through Monte Carlo simulation studies, the Bayesian estimator (i.e., BMLE) is compared with the traditional estimators (i.e., M2SE and MLE).

The arithmetic mean (AVE) and the root mean square error (RMSE) have been usually used in Monte Carlo study.

Moreover, for comparison with the standard normal distribution, Skewness and Kurtosis are also computed.

Moments of the parameters are needed in the calculation of AVE, RMSE, Skewness and Kurtosis.

However, we cannot assure that these moments actually exist.

Therefore, in addition to AVE and RMSE, we also present values for quartiles, i.e., the first quartile (25%), median (50%), the third quartile (75%) and the interquartile range (IR).

Thus, for each estimator, AVE, RMSE, Skewness, Kurtosis, 25%, 50%, 75% and IR are computed from G simulation runs.

The results are given in Table 3, where BMLE is compared with M2SE and MLE.

The case of $n = 20$, $M = 5000$ and $N = 10^4$ is examined in Table 3.

A discussion on choice of M and N is given in Appendix 6.1.6, where we examine whether $M = 5000$ and $N = 10^4$ are sufficient.

Table 3: The AVE, RMSE and Quartiles: $n = 20$

		β_1	β_2	β_3	γ_1	γ_2
	True Value	10	1	1	−2	0.25
M2SE	AVE	10.064	0.995	1.002	−0.988	0.199
	RMSE	7.537	0.418	0.333	3.059	0.146
	Skewness	0.062	−0.013	−0.010	−0.101	−0.086
	Kurtosis	4.005	3.941	2.988	3.519	3.572
	25%	5.208	0.728	0.778	−2.807	0.113
	50%	10.044	0.995	1.003	−0.934	0.200
	75%	14.958	1.261	1.227	0.889	0.287
	IR	9.751	0.534	0.449	3.697	0.175

Table 3: The AVE, RMSE and Quartiles: $n = 20$ — Cont.

		β_1	β_2	β_3	γ_1	γ_2
	True Value	10	1	1	−2	0.25
MLE	AVE	10.029	0.997	1.002	−2.753	0.272
	RMSE	7.044	0.386	0.332	2.999	0.139
	Skewness	0.081	−0.023	−0.014	0.006	−0.160
	Kurtosis	4.062	3.621	2.965	4.620	4.801
	25%	5.323	0.741	0.775	−4.514	0.189
	50%	10.066	0.998	1.002	−2.710	0.273
	75%	14.641	1.249	1.229	−0.958	0.355
	IR	9.318	0.509	0.454	3.556	0.165

Table 3: The AVE, RMSE and Quartiles: $n = 20$ — Cont.

		β_1	β_2	β_3	γ_1	γ_2
	True Value	10	1	1	-2	0.25
	AVE	10.034	0.996	1.002	-2.011	0.250
	RMSE	6.799	0.380	0.328	2.492	0.117
	Skewness	0.055	-0.016	-0.013	-0.016	-0.155
BMLE	Kurtosis	3.451	3.340	2.962	3.805	3.897
	25%	5.413	0.745	0.778	-3.584	0.176
	50%	10.041	0.996	1.002	-1.993	0.252
	75%	14.538	1.246	1.226	-0.407	0.325
	IR	9.125	0.501	0.448	3.177	0.150

$c = 2.0$, $M = 5000$ and $N = 10^4$ are chosen for BMLE

First, we compare the two traditional estimators, i.e., M2SE and MLE.

Judge, Hill, Griffiths and Lee (1980, pp.141–142) indicated that 2SE of γ_1 is inconsistent although 2SE of the other parameters is consistent but asymptotically inefficient.

For M2SE, the estimate of γ_1 is modified to be consistent.

But M2SE is still asymptotically inefficient while MLE is consistent and asymptotically efficient.

Therefore, for γ, MLE should have better performance than M2SE in the sense of efficiency.

In Table 3, for all the parameters except for IR of β_3, RMSE and IR of MLE are smaller than those of M2SE.

For both M2SE and MLE, AVEs of β are close to the true parameter values.

Therefore, it might be concluded that M2SE and MLE are unbiased for β even in the case of small sample.

However, the estimates of γ are different from the true values for both M2SE and MLE.

That is, AVE and 50% of γ_1 are -0.988 and -0.934 for M2SE, and -2.753 and -2.710 for MLE, which are far from the true value -2.0.

Similarly, AVE and 50% of γ_2 are 0.199 and 0.200 for M2SE, which are different from the true value 0.25.

But 0.272 and 0.273 for MLE are slightly larger than 0.25 and they are close to 0.25.

Thus, the traditional estimators work well for the regression coefficients β but not for the heteroscedasticity parameters γ.

Next, the Bayesian estimator (i.e., BMLE) is compared with the traditional ones (i.e., M2SE and MLE).

For all the parameters of β, we can find from Table 3 that BMLE shows better performance in RMSE and IR than the traditional estimators, because RMSE and IR of BMLE are smaller than those of M2SE and MLE.

Furthermore, from AVEs of BMLE, we can see that the heteroscedasticity parameters as well as the regression coefficients are unbiased in the small sample.

Thus, Table 3 also shows the evidence that for both β and γ, AVE and 50% of BMLE are very close to the true parameter values.

The values of RMSE and IR also indicate that the estimates are concentrated around the AVE and 50%, which are vary close to the true parameter values.

For the regression coefficient β, all of the three estimators are very close to the true parameter values. However, for the heteroscedasticity parameter γ, BMLE shows a good performance but M2SE and MLE are poor.

The larger values of RMSE for the traditional counterparts may be due to “outliers” encountered with the Monte Carlo experiments.

This problem is also indicated in Zellner (1971, pp.281).

Compared with the traditional counterparts, the Bayesian approach is not characterized by extreme values for posterior modal values.

Now we compare empirical distributions for M2SE, MLE and BMLE in Figures 3 – 7.

Figure 3: Empirical Distributions of β_1

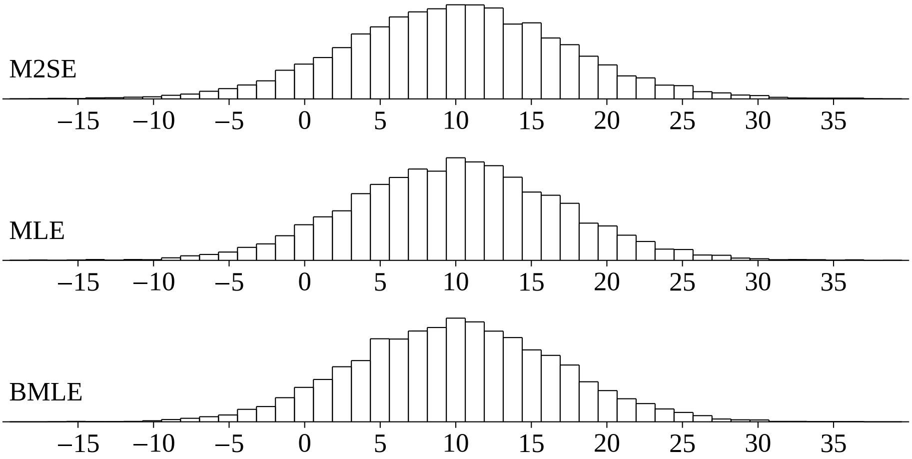


Figure 4: Empirical Distributions of β_2

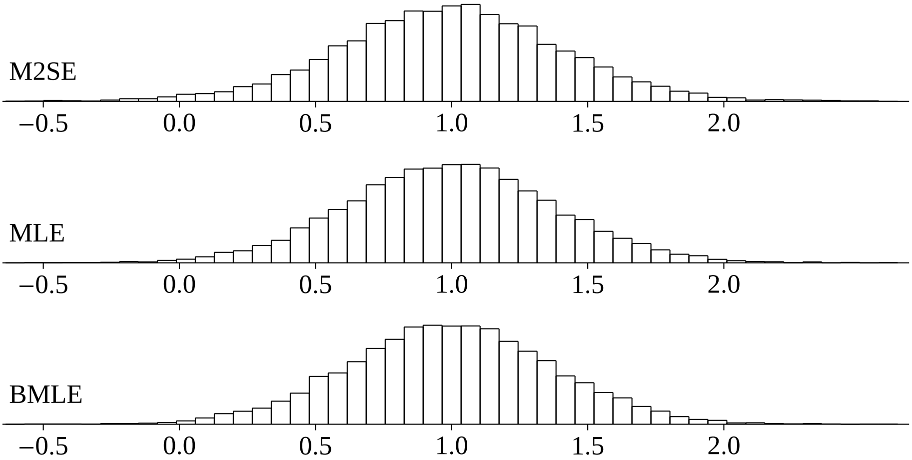


Figure 5: Empirical Distributions of β_3

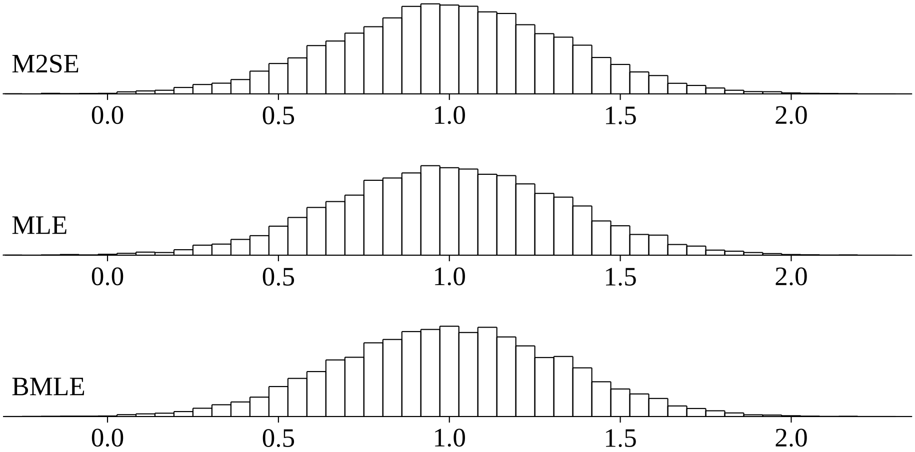


Figure 6: Empirical Distributions of γ_1

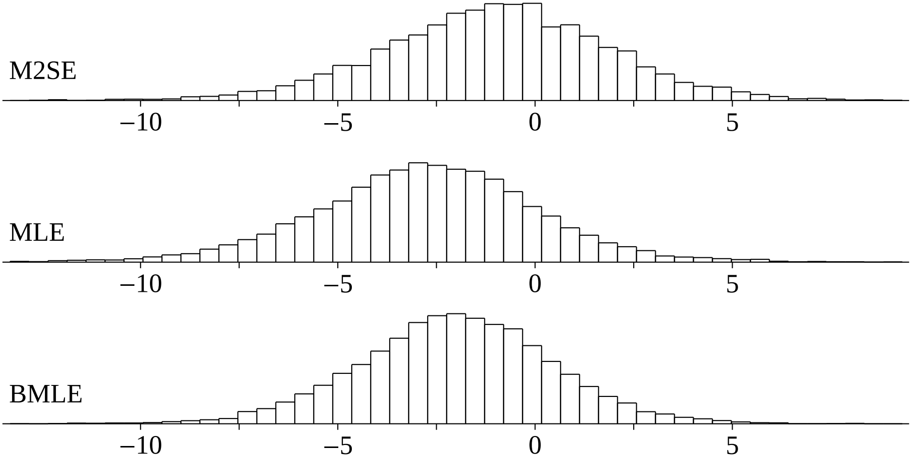
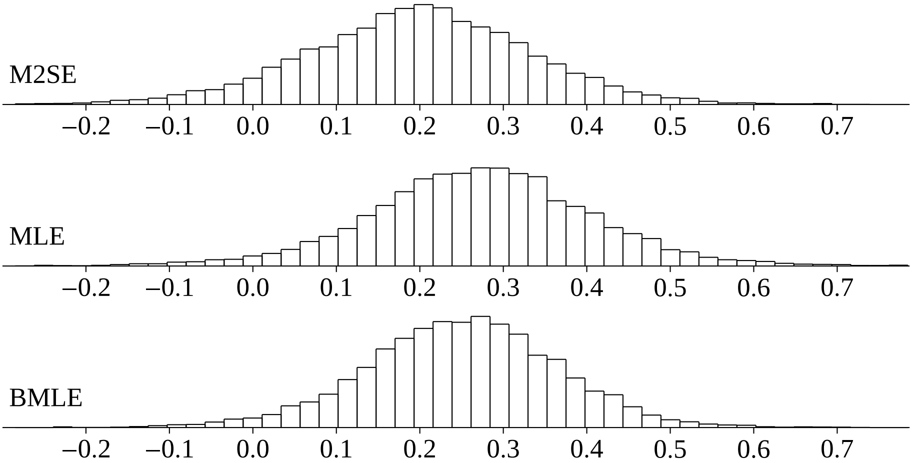


Figure 7: Empirical Distributions of γ_2



For the posterior densities of β_1 (Figure 3), β_2 (Figure 4), β_3 (Figure 5) and γ_1 (Figure 6), all of M2SE, MLE and BMLE are almost symmetric (also, see Skewness in Table 3).

For the posterior density of γ_2 (Figure 7), both MLE and BMLE are slightly skewed to the left because Skewness of γ_2 in Table 3 is negative, while M2SE is almost symmetric.

As for Kurtosis, all the empirical distributions except for β_3 have a sharp kurtosis and fat tails, compared with the normal distribution.

Especially, for the heteroscedasticity parameters γ_1 and γ_2, MLE has the largest kurtosis of the three.

For all figures, location of the empirical distributions indicates whether the estimators are unbiased or not.

For β_1 in Figure 3, β_2 in Figure 4 and β_3 in Figure 5, M2SE is biased while MLE and BMLE are distributed around the true value.

For γ_1 in Figure 6 and γ_2 in Figure 7, the empirical distributions of M2SE, MLE and BMLE are quite different.

For γ_1 in Figure 6, M2SE is located in the right-hand side of the true parameter value, MLE is in the left-hand side, and BMLE is also slightly in the left-hand side.

Moreover, for γ_2 in Figure 7, M2SE is downward-biased, MLE is overestimated, and BMLE is distributed around the true parameter value.

On the Sample Size *n*: Finally, we examine how the sample size n influences precision of the parameter estimates.

Since we utilize the exogenous variable X shown in Judge, Hill, Griffiths and Lee (1980), we cannot examine the case where n is greater than 20.

In order to see the effect of the sample size n, here the case of $n = 15$ is compared with that of $n = 20$.

The case $n = 15$ of BMLE is shown in Table 4, which should be compared with BMLE in Table 3.

As a result, all the AVEs are very close to the corresponding true parameter values.

Therefore, we can conclude from Tables 3 and 4 that the Bayesian estimator is unbiased even in the small sample such as $n = 15, 20$.

However, RMSE and IR become large as n decreases.

That is, for example, RMSEs of β_1, β_2, β_3, γ_1 and γ_2 are given by 6.799, 0.380, 0.328, 2.492 and 0.117 in Table 3, and 8.715, 0.455, 0.350, 4.449 and 0.228 in Table 4.

Thus, we can see that RMSE and IR decrease as n is large.

Table 4: BMLE: $n = 15$, $c = 2.0$, $M = 5000$ and $N = 10^4$

	β_1	β_2	β_3	γ_1	γ_2
True Value	10	1	1	−2	0.25
AVE	10.060	0.995	1.002	−2.086	0.252
RMSE	8.715	0.455	0.350	4.449	0.228
Skewness	0.014	0.033	−0.064	−0.460	0.308
Kurtosis	3.960	3.667	3.140	4.714	4.604
25%	4.420	0.702	0.772	−4.725	0.107
50%	10.053	0.995	1.004	−1.832	0.245
75%	15.505	1.284	1.237	0.821	0.391
IR	11.085	0.581	0.465	5.547	0.284

6.1.5 Summary

In Section 6.1, we have examined the multiplicative heteroscedasticity model discussed by Harvey (1976), where the two traditional estimators are compared with the Bayesian estimator.

For the Bayesian approach, we have evaluated the posterior mean by generating random draws from the posterior density, where the Markov chain Monte Carlo methods (i.e., the MH within Gibbs algorithm) are utilized.

In the MH algorithm, the sampling density has to be specified.

We examine the multivariate normal sampling density, which is the independence chain in the MH algorithm.

For mean and variance in the sampling density, we consider using the mean and variance estimated by the two traditional estimators (i.e., M2SE and MLE).

The Bayesian estimators with M2SE and MLE are called BM2SE and BMLE in Section 6.1.

Through the Monte Carlo studies, the results are summarized as follows:

(i) We compare BM2SE and BMLE with respect to the acceptance rates in the MH algorithm.

    In this case, BMLE shows higher acceptance rates than BM2SE for all c, which is shown in Figure 2.

For the sampling density, we utilize the independence chain through Section 6.1.

The high acceptance rate implies that the chain travels over the support of the target density.

For the Bayesian estimator, therefore, BMLE is preferred to BM2SE.

However, note as follows.

The sampling density which yields the highest acceptance rate is not necessarily the best choice and the tuning parameter c should be larger than the value which gives us the maximum acceptance rate.

Therefore, we have focused on BMLE with $c = 2$ (remember that BMLE with $c = 1.2$ yields the maximum acceptance rate).

(ii) For the traditional estimators (i.e., M2SE and MLE), we have obtained the result that MLE has smaller RMSE than M2SE for all the parameters, because for one reason the M2SE is asymptotically less efficient than the MLE.

Furthermore, for M2SE, the estimates of β are unbiased but those of γ are different from the true parameter values (see Table 3).

(iii) From Table 3, BMLE performs better than the two traditional estimators in the sense of RMSE and IR, because RMSE and IR of BMLE are smaller than those of the traditional ones for all the cases.

(iv) Each empirical distribution is displayed in Figures 3 – 7.

The posterior densities of almost all the estimates are distributed to be symmetric (γ_2 is slightly skewed to the left), but the posterior densities of both the regression coefficients (except for β_3) and the heteroscedasticity parameters have fat tails.

Also, see Table 3 for skewness and kurtosis.

(v) As for BMLE, the case of n = 15 is compared with n = 20.

The case n = 20 has smaller RMSE and IR than n = 15, while AVE and 50% are close to the true parameter values for β and γ.

Therefore, it might be expected that the estimates of BMLE go to the true parameter values as n is large.

6.1.6 Appendix: Are $M = 5000$ and $N = 10^4$ Sufficient?

Table 5: BMLE: $n = 20$ and $c = 2.0$

		β_1	β_2	β_3	γ_1	γ_2
	True Value	10	1	1	−2	0.25
	AVE	10.028	0.997	1.002	−2.008	0.250
	RMSE	6.807	0.380	0.328	2.495	0.117
	Skewness	0.041	−0.007	−0.012	0.017	−0.186
$M = 1000$	Kurtosis	3.542	3.358	2.963	3.950	4.042
$N = 10^4$	25%	5.413	0.745	0.778	−3.592	0.176
	50%	10.027	0.996	1.002	−1.998	0.252
	75%	14.539	1.245	1.226	−0.405	0.326
	IR	9.127	0.500	0.448	3.187	0.150

Table 5: BMLE: $n = 20$ and $c = 2.0$ — Cont.

		β_1	β_2	β_3	γ_1	γ_2
	True Value	10	1	1	−2	0.25
	AVE	10.033	0.996	1.002	−2.010	0.250
	RMSE	6.799	0.380	0.328	2.491	0.117
	Skewness	0.059	−0.016	−0.011	−0.024	−0.146
$M = 5000$	Kurtosis	3.498	3.347	2.961	3.764	3.840
$N = 5000$	25%	5.431	0.747	0.778	−3.586	0.176
	50%	10.044	0.995	1.002	−1.997	0.252
	75%	14.532	1.246	1.225	−0.406	0.326
	IR	9.101	0.499	0.447	3.180	0.149

In Section 6.1.4, only the case of $(M, N) = (5000, 10^4)$ is examined.

In this appendix, we check whether $M = 5000$ and $N = 10^4$ are sufficient.

For the burn-in period M, there are some diagnostic tests, which are discussed in Geweke (1992) and Mengersen, Robert and Guihenneuc-Jouyaux (1999).

However, since their tests are applicable in the case of one sample path, we cannot utilize them.

Because G simulation runs are implemented in Section 6.1.4 (see p.516 for the simulation procedure), we have G test statistics if we apply the tests.

It is difficult to evaluate G testing results at the same time.

Therefore, we consider using the alternative approach to see if $M = 5000$ and $N = 10^4$ are sufficient.

For choice of M and N, we consider the following two issues.

(i) Given fixed $M = 5000$, compare $N = 5000$ and $N = 10^4$.

(ii) Given fixed $N = 10^4$, compare $M = 1000$ and $M = 5000$.

(i) examines whether $N = 5000$ is sufficiently large, while (ii) checks whether $M = 1000$ is large enough. If the case of $(M, N) = (5000, 5000)$ is close to that of $(M, N) = (5000, 10^4)$, we can conclude that $N = 5000$ is sufficiently large.

Similarly, if the case of $(M, N) = (1000, 10^4)$ is not too different from that of $(M, N) = (5000, 10^4)$, it might be concluded that $M = 1000$ is also sufficient.

The results are in Table 5, where AVE, RMSE, Skewness, Kurtosis, 25%, 50%, 75% and IR are shown for each of the regression coefficients and the heteroscedasticity parameters.

BMLE in Table 3 should be compared with Table 5.

From Tables 3 and 5, the three cases, i.e., $(M, N) = (5000, 10^4)$, $(1000, 10^4)$, $(5000, 5000)$, are very close to each other.

Therefore, we can conclude that both $M = 1000$ and $N = 5000$ are large enough in the simulation study shown in Section 6.1.4.

We take the case of $M = 5000$ and $N = 10^4$ for safety in Section 6.1.4, although we obtain the results that both $M = 1000$ and $N = 5000$ are large enough.

6.2 Autocorrelation Model

In the previous section, we have considered estimating the regression model with the heteroscedastic error term, where the traditional estimators such as MLE and M2SE are compared with the Bayesian estimators.

In this section, using both the maximum likelihood estimator and the Bayes estimator, we consider the regression model with the first order autocorrelated error term, where the initial distribution of the autocorrelated error is taken into account.

As for the autocorrelated error term, the stationary case is assumed, i.e., the autocorrelation coefficient is assumed to be less than one in absolute value.

The traditional estimator (i.e., MLE) is compared with the Bayesian estimator. Utilizing the Gibbs sampler, Chib (1993) discussed the regression model with the autocorrelated error term in a Bayesian framework, where the initial condition of the autoregressive process is not taken into account.

In this section, taking into account the initial density, we compare the maximum likelihood estimator and the Bayesian estimator.

For the Bayes estimator, the Gibbs sampler and the Metropolis-Hastings algorithm are utilized to obtain random draws of the parameters.

As a result, the Bayes estimator is less biased and more efficient than the maximum likelihood estimator. Especially, for the autocorrelation coefficient, the Bayes estimate is much less biased than the maximum likelihood estimate.

Accordingly, for the standard error of the estimated regression coefficient, the Bayes estimate is more plausible than the maximum likelihood estimate.

6.2.1 Introduction

In Section 6.2, we consider the regression model with the first order autocorrelated error term, where the error term is assumed to be stationary, i.e., the autocorrelation coefficient is assumed to be less than one in absolute value.

The traditional estimator, i.e., the maximum likelihood estimator (MLE), is compared with the Bayes estimator (BE).

Utilizing the Gibbs sampler, Chib (1993) and Chib and Greenberg (1994) discussed the regression model with the autocorrelated error term in a Bayesian framework, where the initial condition of the autoregressive process is ignored.

Here, taking into account the initial density, we compare MLE and BE, where the Gibbs sampler and the Metropolis-Hastings (MH) algorithm are utilized in BE.

As for MLE, it is well known that the autocorrelation coefficient is underestimated in small sample and therefore that variance of the estimated regression coefficient is also biased.

See, for example, Andrews (1993) and Tanizaki (2000, 2001).

Under this situation, inference on the regression coefficient is not appropriate, because variance of the estimated regression coefficient depends on the estimated autocorrelation coefficient.

We show in Section 6.2 that BE is superior to MLE because BEs of both the autocorrelation coefficient and the variance of the error term are closer to the true values, compared with MLEs.

6.2.2 Setup of the Model

Let X_t be a $1 \times k$ vector of exogenous variables and β be a $k \times 1$ parameter vector. Consider the following regression model:

$$y_t = X_t\beta + u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2),$$

for $t = 1, 2, \cdots, n$, where ϵ_1, ϵ_2, $\cdots$, ϵ_n are assumed to be mutually independently distributed.

In this model, the parameter to be estimated is given by $\theta = (\beta, \rho, \sigma_\epsilon^2)$.

The unconditional density of y_t is:

$$f(y_t|\beta, \rho, \sigma_\epsilon^2) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2/(1-\rho^2)}} \exp\Big(-\frac{1}{2\sigma_\epsilon^2/(1-\rho^2)}(y_t - X_t\beta)^2\Big).$$

Let Y_t be the information set up to time t, i.e., $Y_t = \{y_t, y_{t-1}, \cdots, y_1\}$.

The conditional density of y_t given Y_{t-1} is:

$$\begin{aligned} f(y_t|Y_{t-1}, \beta, \rho, \sigma_\epsilon^2) &= f(y_t|y_{t-1}, \beta, \rho, \sigma_\epsilon^2) \\ &= \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\Big(-\frac{1}{2\sigma_\epsilon^2}((y_t - \rho y_{t-1}) - (X_t - \rho X_{t-1})\beta)^2\Big). \end{aligned}$$

Therefore, the joint density of Y_n, i.e., the likelihood function, is given by :

$$
\begin{aligned}
f(Y_n|\beta,\rho,\sigma_\epsilon^2) &= f(y_1|\beta,\rho,\sigma_\epsilon^2)\prod_{t=2}^{n} f(y_t|Y_{t-1},\beta,\rho,\sigma_\epsilon^2) \\
&= (2\pi\sigma_\epsilon^2)^{-n/2}(1-\rho^2)^{1/2}\exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}(y_t^* - X_t^*\beta)^2\Big), \qquad (20)
\end{aligned}
$$

where y_t^* and X_t^* represent the following transformed variables:

$$
y_t^* = y_t^*(\rho) = \begin{cases} \sqrt{1-\rho^2}\,y_t, & \text{for } t = 1, \\ \\ y_t - \rho y_{t-1}, & \text{for } t = 2, 3, \cdots, n, \end{cases}
$$

$$X_t^* = X_t^*(\rho) = \begin{cases} \sqrt{1-\rho^2}X_t, & \text{for } t = 1, \\ X_t - \rho X_{t-1}, & \text{for } t = 2, 3, \cdots, n, \end{cases}$$

which depend on the autocorrelation coefficient ρ.

Maximum Likelihood Estimator: We have shown above that the likelihood function is given by equation (20).

Maximizing equation (20) with respect to β and σ_ϵ^2, we obtain the following expressions:

$$\hat{\beta} \equiv \hat{\beta}(\rho) = (\sum_{t=1}^{n} X_t^{*\prime} X_t^*)^{-1} \sum_{t=1}^{n} X_t^{*\prime} y_t^*,$$

$$\hat{\sigma}_{\epsilon}^{2} \equiv \hat{\sigma}_{\epsilon}^{2}(\rho) = \frac{1}{n}\sum_{t=1}^{n}(y_t^* - X_t^*\hat{\beta})^2. \tag{21}$$

By substituting $\hat{\beta}$ and $\hat{\sigma}_{\epsilon}^{2}$ into β and σ_{ϵ}^{2} in equation (20), we have the concentrated likelihood function:

$$f(Y_n|\hat{\beta}, \rho, \hat{\sigma}_{\epsilon}^{2}) = \left(2\pi\hat{\sigma}_{\epsilon}^{2}(\rho)\right)^{-n/2}(1-\rho^2)^{1/2}\exp(-\frac{n}{2}), \tag{22}$$

which is a function of ρ.

Equation (22) has to be maximized with respect to ρ.

In the next section, we obtain the maximum likelihood estimate of ρ by a simple grid search, in which the concentrated likelihood function (22) is maximized by changing the parameter value of ρ by 0.0001 in the interval between -0.9999 and 0.9999.

Once the solution of ρ, denoted by $\hat{\rho}$, is obtained, $\hat{\beta}(\hat{\rho})$ and $\hat{\sigma}_\epsilon^2(\hat{\rho})$ lead to the maximum likelihood estimates of β and σ_ϵ^2.

Hereafter, $\hat{\beta}$, $\hat{\sigma}_\epsilon^2$ and $\hat{\rho}$ are taken as the maximum likelihood estimates of β, σ_ϵ^2 and ρ, i.e., $\hat{\beta}(\hat{\rho})$ and $\hat{\sigma}_\epsilon^2(\hat{\rho})$ are simply written as $\hat{\beta}$ and $\hat{\sigma}_\epsilon^2$.

Variance of the estimate of $\theta = (\beta', \sigma^2, \rho)'$ is asymptotically given by: $\mathrm{V}(\hat{\theta}) = I^{-1}(\theta)$, where $I(\theta)$ denotes the information matrix, which is represented as:

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2 \log f(Y_n|\theta)}{\partial\theta\partial\theta'}\right).$$

Therefore, variance of $\hat{\beta}$ is given by $\mathrm{V}(\hat{\beta}) = \sigma^2(\sum_{t=1}^{n} X_t^{*\prime} X_t^{*})^{-1}$ in large sample, where ρ in X_t^* is replaced by $\hat{\rho}$, i.e., $X_t^* = X_t^*(\hat{\rho})$.

For example, suppose that X_t^* has a tendency to rise over time t and that we have $\rho > 0$.

If ρ is underestimated, then $\mathrm{V}(\hat{\beta})$ is also underestimated, which yields incorrect inference on the regression coefficient β.

Thus, unless ρ is properly estimated, the estimate of $\mathrm{V}(\hat{\beta})$ is also biased.

In large sample, $\hat{\rho}$ is a consistent estimator of ρ and therefore $\mathrm{V}(\hat{\beta})$ is not biased.

However, in small sample, since it is known that $\hat{\rho}$ is underestimated (see, for example, Andrews (1993), Tanizaki (2000, 2001)), clearly $\mathrm{V}(\hat{\beta})$ is also underestimated.

In addition to $\hat{\rho}$, the estimate of σ^2 also influences inference of β, because we have $\mathrm{V}(\hat{\beta}) = \sigma^2(\sum_{t=1}^{n} X_t^{*\prime} X_t^*)^{-1}$ as mentioned above.

If σ^2 is underestimated, the estimated variance of β is also underestimated.

$\hat{\sigma}^2$ is a consistent estimator of σ^2 in large sample, but it is appropriate to consider that $\hat{\sigma}^2$ is biased in small sample, because $\hat{\sigma}^2$ is a function of $\hat{\rho}$ as in (21).

Therefore, the biased estimate of ρ gives us the serious problem on inference of β.

Bayesian Estimator: We assume that the prior density functions of β, ρ and σ_ϵ^2 are the following noninformative priors:

$$f_\beta(\beta) \propto \text{constant}, \qquad \text{for } -\infty < \beta < \infty, \tag{23}$$

$$f_\rho(\rho) \propto \text{constant}, \qquad \text{for } -1 < \rho < 1, \tag{24}$$

$$f_{\sigma_\epsilon}(\sigma_\epsilon^2) \propto \frac{1}{\sigma_\epsilon^2}, \qquad \text{for } 0 < \sigma_\epsilon^2 < \infty. \tag{25}$$

In equation (24), theoretically we should have $-1 < \rho < 1$.

As for the prior density of σ_ϵ^2, since we consider that $\log\sigma_\epsilon^2$ has the flat prior for $-\infty < \log\sigma_\epsilon^2 < \infty$, we obtain $f_{\sigma_\epsilon}(\sigma_\epsilon^2) \propto 1/\sigma_\epsilon^2$.

Note that in Section 6.1 the first element of the heteroscedasticity parameter γ is also assumed to be diffuse, where it is formulated as the logarithm of variance of the error term, i.e., $\log\sigma_\epsilon^2$.

Combining the four densities (20) and (23) – (25), the posterior density function of β, ρ and σ_ϵ^2, denoted by $f_{\beta\rho\sigma_\epsilon}(\beta,\rho,\sigma_\epsilon^2|Y_n)$, is represented as follows:

$$\begin{aligned} & f_{\beta\rho\sigma_\epsilon}(\beta,\rho,\sigma_\epsilon^2|Y_n) \\ & \quad \propto f(Y_n|\beta,\rho,\sigma_\epsilon^2)\, f_\beta(\beta)\, f_\rho(\rho)\, f_{\sigma_\epsilon}(\sigma_\epsilon^2) \\ & \quad \propto (\sigma_\epsilon^2)^{-(n/2+1)}(1-\rho^2)^{1/2}\exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}(y_t^*-X_t^*\beta)^2\Big). \end{aligned} \tag{26}$$

We want to have random draws of β, ρ and σ_ϵ^2 given Y_n.

However, it is not easy to generate random draws of β, ρ and σ_ϵ^2 from $f_{\beta\rho\sigma_\epsilon}(\beta, \rho, \sigma_\epsilon^2|Y_n)$.

Therefore, we perform the Gibbs sampler in this problem.

According to the Gibbs sampler, we can sample from the posterior density function (26), using the three conditional distributions $f_{\beta|\rho\sigma_\epsilon}(\beta|\rho, \sigma_\epsilon^2, Y_n)$, $f_{\rho|\beta\sigma_\epsilon}(\rho|\beta, \sigma_\epsilon^2, Y_n)$ and $f_{\sigma_\epsilon|\beta\rho}(\sigma_\epsilon^2|\beta, \rho, Y_n)$, which are proportional to $f_{\beta\rho\sigma}(\beta, \rho, \sigma^2|Y_n)$ and are obtained as follows:

- $f_{\beta|\rho\sigma_\epsilon}(\beta|\rho, \sigma_\epsilon^2, Y_n)$ is given by:

$$f_{\beta|\rho\sigma_\epsilon}(\beta|\rho, \sigma_\epsilon^2, Y_n)$$

$$
\begin{aligned}
&\propto f_{\beta\rho\sigma_\epsilon}(\beta,\rho,\sigma_\epsilon^2|Y_n) \propto \exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}(y_t^* - X_t^*\beta)^2\Big) \\
&= \exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}\Big((y_t^* - X_t^*\hat{\beta}) - X_t(\beta-\hat{\beta})\Big)^2\Big) \\
&= \exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}(y_t^* - X_t^*\hat{\beta})^2 - \frac{1}{2\sigma_\epsilon^2}(\beta-\hat{\beta})'(\sum_{t=1}^{n}X_t^{*\prime}X_t^*)(\beta-\hat{\beta})\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\beta-\hat{\beta})'(\frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{n}X_t^{*\prime}X_t^*)(\beta-\hat{\beta})\Big), \qquad (27)
\end{aligned}
$$

which indicates that $\beta \sim N(\hat{\beta}, \sigma_\epsilon^2(\sum_{t=1}^n X_t^{*\prime}X_t^*)^{-1})$, where $\hat{\beta}$ represents the OLS estimate, i.e., $\hat{\beta} = (\sum_{t=1}^n X_t^{*\prime}X_t^*)^{-1}(\sum_{t=1}^n X_t^{*\prime}y_t^*)$.

Thus, (27) implies that β can be sampled from the multivariate normal distribution with mean $\hat{\beta}$ and variance $\sigma_\epsilon^2(\sum_{t=1}^n X_t^{*\prime} X_t^*)^{-1}$.

- $f_{\rho|\beta\sigma_\epsilon}(\rho|\beta, \sigma_\epsilon^2, Y_n)$ is obtained as:

$$\begin{aligned} f_{\rho|\beta\sigma_\epsilon}(\rho|\beta, \sigma_\epsilon^2, Y_n) &\propto f_{\beta\rho\sigma_\epsilon}(\beta, \rho, \sigma_\epsilon^2|Y_n) \\ &\propto (1-\rho^2)^{1/2} \exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^n \big(y_t^* - X_t^*\beta\big)^2\Big), \end{aligned} \tag{28}$$

for $-1 < \rho < 1$, which cannot be represented in a known distribution.

Note that $y_t^* = y_t^*(\rho)$ and $X_t^* = X_t^*(\rho)$.

Sampling from (28) is implemented by the MH algorithm.

A detail discussion on sampling will be given later.

- $f_{\sigma_\epsilon|\beta\rho}(\sigma_\epsilon^2|\beta,\rho,Y_n)$ is represented as:

$$
\begin{aligned}
f_{\sigma_\epsilon|\beta\rho}(\sigma_\epsilon^2|\beta,\rho,Y_n) &\propto f_{\beta\rho\sigma_\epsilon}(\beta,\rho,\sigma_\epsilon^2|Y_n) \\
&\propto \frac{1}{(\sigma_\epsilon^2)^{n/2+1}} \exp\Big(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{n}(y_t^* - X_t^*\beta)^2\Big),
\end{aligned}
\tag{29}
$$

which is written as follows: $\sigma_\epsilon^2 \sim IG(n/2,\ 2/\sum_{t=1}^{n}\epsilon_t^2)$, or equivalently, $1/\sigma_\epsilon^2 \sim G(n/2,\ 2/\sum_{t=1}^{n}\epsilon_t^2)$, where $\epsilon_t = y_t^* - X_t^*\beta$.

Thus, in order to generate random draws of β, ρ and σ_ϵ^2 from the posterior density $f_{\beta\rho\sigma_\epsilon}(\beta,\rho,\sigma_\epsilon^2|Y_n)$, the following procedures have to be taken:

(i) Let β_i, ρ_i and $\sigma^2_{\epsilon,i}$ be the ith random draws of β, ρ and σ^2_ϵ.

Take the initial values of $(\beta, \rho, \sigma^2_\epsilon)$ as $(\beta_{-M}, \rho_{-M}, \sigma^2_{\epsilon,-M})$.

(ii) From equation (27), generate β_i given ρ_{i-1}, $\sigma^2_{\epsilon,i-1}$ and Y_n, using $\beta \sim N(\hat{\beta}, \sigma^2_{\epsilon,i-1}(\sum_{t=1}^n X_t^{*\prime} X_t^*)^{-1})$, where $\hat{\beta} = (\sum_{t=1}^n X_t^{*\prime} X_t^*)^{-1}(\sum_{t=1}^n X_t^{*\prime} y_t^*)$, $y_t^* = y_t^*(\rho_{i-1})$ and $X_t^* = X_t^*(\rho_{i-1})$.

(iii) From equation (28), generate ρ_i given β_i, $\sigma^2_{\epsilon,i-1}$ and Y_n.

Since it is not easy to generate random draws from (27), the Metropolis-Hastings algorithm is utilized, which is implemented as follows:

(a) Generate ρ^* from the uniform distribution between -1 and 1, which implies that the sampling density of ρ is given by $f_*(\rho|\rho_{i-1}) = 1/2$ for $-1 < \rho < 1$.

Compute the acceptance probability $\omega(\rho_{i-1}, \rho^*)$, which is defined as:

$$\begin{aligned}
\omega(\rho_{i-1}, \rho^*) &= \min\left(\frac{f_{\rho|\beta\sigma_\epsilon}(\rho^*|\beta_i, \sigma^2_{\epsilon,i-1}, Y_n)/f_*(\rho^*|\rho_{i-1})}{f_{\rho|\beta\sigma_\epsilon}(\rho_{i-1}|\beta_i, \sigma^2_{\epsilon,i-1}, Y_n)/f_*(\rho_{i-1}|\rho^*)},\ 1 \right) \\
&= \min\left(\frac{f_{\rho|\beta\sigma_\epsilon}(\rho^*|\beta_i, \sigma^2_{\epsilon,i-1}, Y_n)}{f_{\rho|\beta\sigma_\epsilon}(\rho_{i-1}|\beta_i, \sigma^2_{\epsilon,i-1}, Y_n)},\ 1 \right).
\end{aligned}$$

(b) Set $\rho_i = \rho^*$ with probability $\omega(\rho_{i-1}, \rho^*)$ and $\rho_i = \rho_{i-1}$ otherwise.

(iv) From equation (29), generate $\sigma^2_{\epsilon,i}$ given β_i, ρ_i and Y_n, using $1/\sigma^2_\epsilon \sim G(n/2, 2/\sum_{t=1}^n u_t^2)$, where $u_t = y_t^* - X_t^*\beta$, $y_t^* = y_t^*(\rho_i)$ and $X_t^* = X_t^*(\rho_i)$.

(v) Repeat Steps (ii) – (iv) for $i = -M+1, -M+2, \cdots, N$, where M indicates the burn-in period.

Repetition of Steps (ii) – (iv) corresponds to the Gibbs sampler.

For sufficiently large M, we have the following results:

$$\frac{1}{N}\sum_{i=1}^{N} g(\beta_i) \longrightarrow \mathrm{E}(g(\beta)),$$

$$\frac{1}{N}\sum_{i=1}^{N} g(\rho_i) \longrightarrow \mathrm{E}(g(\rho)),$$

$$\frac{1}{N}\sum_{i=1}^{N} g(\sigma_{\epsilon,i}^2) \longrightarrow \mathrm{E}(g(\sigma_\epsilon^2)),$$

where $g(\cdot)$ is a function, typically $g(x) = x$ or $g(x) = x^2$.

We define the Bayesian estimates of β, ρ and σ_ϵ^2 as $\widetilde{\beta} \equiv (1/N)\sum_{i=1}^N \beta_i$, $\widetilde{\rho} \equiv (1/N)\sum_{i=1}^N \rho_i$ and $\widetilde{\sigma}_\epsilon^2 \equiv (1/N)\sum_{i=1}^N \sigma_{\epsilon,i}^2$, respectively.

Thus, using both the Gibbs sampler and the MH algorithm, we have shown that we can sample from $f_{\beta\rho\sigma_\epsilon}(\beta, \rho, \sigma_\epsilon^2|Y_n)$.

See, for example, Bernardo and Smith (1994), Carlin and Louis (1996), Chen, Shao and Ibrahim (2000), Gamerman (1997), Robert and Casella (1999) and Smith and Roberts (1993) for the Gibbs sampler and the MH algorithm.

6.2.3 Monte Carlo Experiments

For the exogenous variables, again we take the data used in Section 6.1, in which the true data generating process (DGP) is presented in Judge, Hill, Griffiths and Lee (1980, p.156).

As in equation (18), the DGP is defined as:

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \tag{30}$$

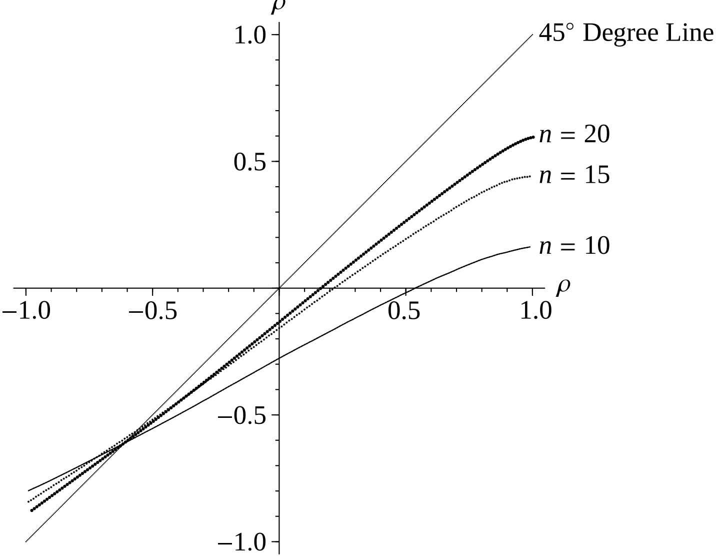
where ϵ_t, $t = 1, 2, \cdots, n$, are normally and independently distributed with $\mathrm{E}(\epsilon_t) = 0$ and $\mathrm{E}(\epsilon_t^2) = \sigma_\epsilon^2$.

As in Judge, Hill, Griffiths and Lee (1980), the parameter values are set to be (β_1, β_2, β_3) $= (10, 1, 1)$.

We utilize $x_{2,t}$ and $x_{3,t}$ given in Judge, Hill, Griffiths and Lee (1980, pp.156), which is shown in Table 1, and generate G samples of y_t given the X_t for $t = 1, 2, \cdots, n$. That is, we perform G simulation runs for each estimator, where $G = 10^4$ is taken.

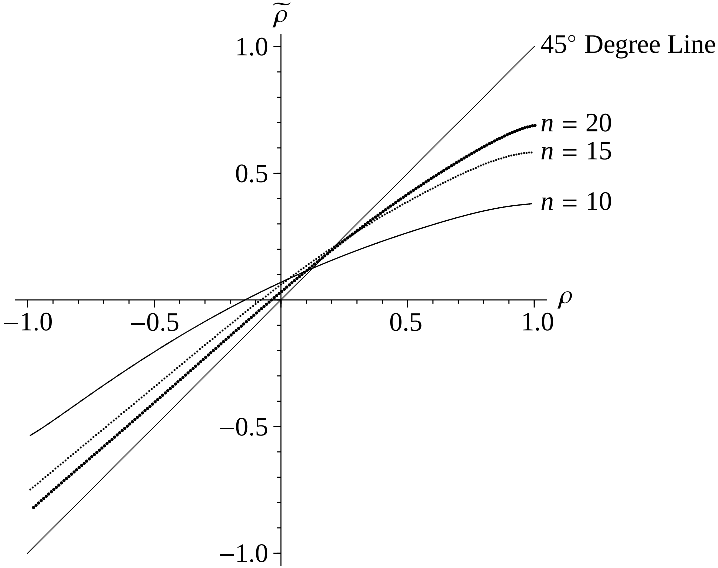
The simulation procedure is as follows:

(i) Given ρ, generate random numbers of u_t for $t = 1, 2, \cdots, n$, based on the



ρ
1.0
0.5
−0.5
−1.0
−1.0
−0.5
0.5
1.0
ρ
45° Degree Line
n = 20
n = 15
n = 10

$M = 5000$ and $N = 10$



True Value	10	1	1	0.9	1
AVE	10.012	0.999	1.000	0.559	0.752
SER	3.025	0.171	0.053	0.240	0.276
RMSE	3.025	0.171	0.053	0.417	0.372
Skewness	0.034	−0.045	−0.008	−1.002	0.736
Kurtosis	2.979	3.093	3.046	4.013	3.812
5%	5.096	0.718	0.914	0.095	0.363
10%	6.120	0.785	0.933	0.227	0.426
25%	7.935	0.883	0.965	0.426	0.550
50%	10.004	0.999	1.001	0.604	0.723
75%	12.051	1.115	1.036	0.740	0.913
90%	13.913	1.217	1.068	0.825	1.120
95%	15.036	1.274	1.087	0.863	1.255

True Value	10	1	1	0.9	1
AVE	10.010	0.999	1.000	0.661	1.051
SER	2.782	0.160	0.051	0.188	0.380
RMSE	2.782	0.160	0.051	0.304	0.384
Skewness	0.008	−0.029	−0.022	−1.389	0.725
Kurtosis	3.018	3.049	2.942	5.391	3.783
5%	5.498	0.736	0.915	0.285	0.515
10%	6.411	0.798	0.934	0.405	0.601
25%	8.108	0.891	0.966	0.572	0.776
50%	10.018	1.000	1.001	0.707	1.011
75%	11.888	1.107	1.036	0.799	1.275
90%	13.578	1.205	1.067	0.852	1.555
95%	14.588	1.258	1.085	0.875	1.750

True Value	10	1	1	0.9	1
AVE	10.011	0.999	1.000	0.661	1.051
SER	2.785	0.160	0.051	0.189	0.380
RMSE	2.785	0.160	0.052	0.305	0.384
Skewness	0.004	−0.027	−0.022	−1.390	0.723
Kurtosis	3.028	3.056	2.938	5.403	3.776
5%	5.500	0.736	0.915	0.285	0.514
10%	6.402	0.797	0.934	0.405	0.603
25%	8.117	0.891	0.966	0.572	0.775
50%	10.015	1.000	1.001	0.707	1.011
75%	11.898	1.107	1.036	0.799	1.277
90%	13.612	1.205	1.066	0.852	1.559
95%	14.600	1.257	1.085	0.876	1.747

True Value	10	1	1	0.9	1
AVE	10.010	0.999	1.000	0.661	1.051
SER	2.783	0.160	0.051	0.188	0.380
RMSE	2.783	0.160	0.051	0.304	0.384
Skewness	0.008	−0.029	−0.021	−1.391	0.723
Kurtosis	3.031	3.055	2.938	5.404	3.774
5%	5.495	0.736	0.915	0.284	0.514
10%	6.412	0.797	0.935	0.404	0.602
25%	8.116	0.891	0.966	0.573	0.774
50%	10.014	1.000	1.001	0.706	1.011
75%	11.897	1.107	1.036	0.799	1.275
90%	13.587	1.204	1.067	0.852	1.558
95%	14.588	1.257	1.085	0.876	1.746

assumptions: $u_t = \rho u_{t-1} + \epsilon_t$ and $\epsilon_t \sim N(0, 1)$.

(ii) Given β, $(x_{2,t}, x_{3,t})$ and u_t for $t = 1, 2, \cdots, n$, we obtain a set of data y_t, $t = 1, 2, \cdots, n$, from equation (30), where $(\beta_1, \beta_2, \beta_3) = (10, 1, 1)$ is assumed.

(iii) Given (y_t, X_t) for $t = 1, 2, \cdots, n$, obtain the estimates of $\theta = (\beta, \rho, \sigma_\epsilon^2)$ by the maximum likelihood estimation (MLE) and the Bayesian estimation (BE) discussed in Sections 6.2.2, which are denoted by $\hat{\theta}$ and $\widetilde{\theta}$, respectively.

(iv) Repeat (i) – (iii) G times, where $G = 10^4$ is taken.

(v) From G estimates of θ, compute the arithmetic average (AVE), the standard error (SER), the root mean square error (RMSE), the skewness (Skewness), the kurtosis (Kurtosis), and the 5, 10, 25, 50, 75, 90 and 95 percent points (5%, 10%, 25%, 50%, 75%, 90% and 95%) for each estimator.

For the maximum likelihood estimator (MLE), we compute:

$$\text{AVE} = \frac{1}{G}\sum_{g=1}^{G}\hat{\theta}_j^{(g)}, \qquad \text{RMSE} = \Big(\frac{1}{G}\sum_{g=1}^{G}(\hat{\theta}_j^{(g)} - \theta_j)^2\Big)^{1/2},$$

for $j = 1, 2, \cdots, 5$, where θ_j denotes the jth element of θ and $\hat{\theta}_j^{(g)}$ represents the jth element of $\hat{\theta}$ in the gth simulation run.

For the Bayesian estimator (BE), $\hat{\theta}$ in the above equations is replaced by $\widetilde{\theta}$, and AVE and RMSE are obtained.

(vi) Repeat (i) – (v) for $\rho = -0.99, -0.98, \cdots, 0.99$.

Thus, in Section 6.2.3, we compare the Bayesian estimator (BE) with the maximum likelihood estimator (MLE) through Monte Carlo studies.

In Figures 8 and 9, we focus on the estimates of the autocorrelation coefficient ρ.

In Figure 8 we draw the relationship between ρ and $\hat{\rho}$, where $\hat{\rho}$ denotes the arithmetic average of the 10^4 MLEs, while in Figure 9 we display the relationship between ρ and $\widetilde{\rho}$, where $\widetilde{\rho}$ indicates the arithmetic average of the 10^4 BEs.

In the two figures the cases of $n = 10, 15, 20$ are shown, and $(M, N) = (5000, 10^4)$ is taken in Figure 9 (we will discuss later about M and N).

If the relationship between ρ and $\hat{\rho}$ (or $\widetilde{\rho}$) lies on the 45° degree line, we can conclude that MLE (or BE) of ρ is unbiased.

However, from the two figures, both estimators are biased.

Take an example of $\rho = 0.9$ in Figures 8 and 9.

When the true value is $\rho = 0.9$, the arithmetic averages of 10^4 MLEs are given by 0.142 for $n = 10$, 0.422 for $n = 15$ and 0.559 for $n = 20$ (see Figure 8), while those of 10^4 BEs are 0.369 for $n = 10$, 0.568 for $n = 15$ and 0.661 for $n = 20$ (see Figure 9).

As n increases the estimators are less biased, because it is shown that MLE gives us the consistent estimators.

Comparing BE and MLE, BE is less biased than MLE in the small sample, because BE is closer to the 45° degree line than MLE.

Especially, as ρ goes to one, the difference between BE and MLE becomes quite large.

Tables 2 – 5 represent the basic statistics such as arithmetic average, standard error, root mean square error, skewness, kurtosis and percent points, which are computed from $G = 10^4$ simulation runs, where the case of $n = 20$ and $\rho = 0.9$ is examined. Table 2 is based on the MLEs while Tables 3 – 5 are obtained from the BEs.

Figure 10: Empirical Distributions of β_1

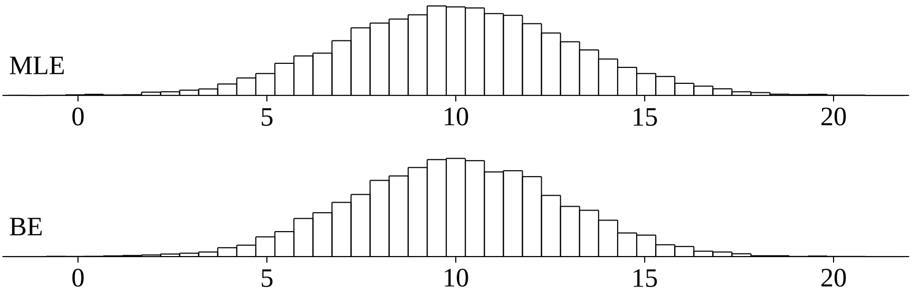


Figure 11: Empirical Distributions of β_2

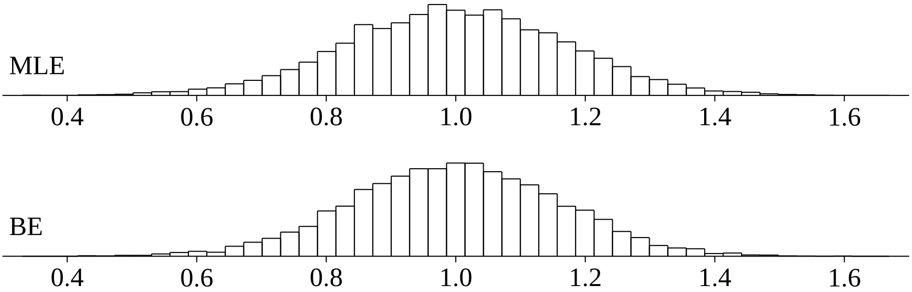


Figure 12: Empirical Distributions of β_3

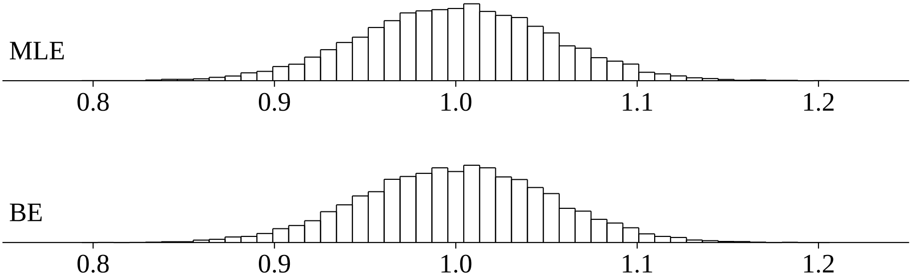


Figure 13: Empirical Distributions of ρ

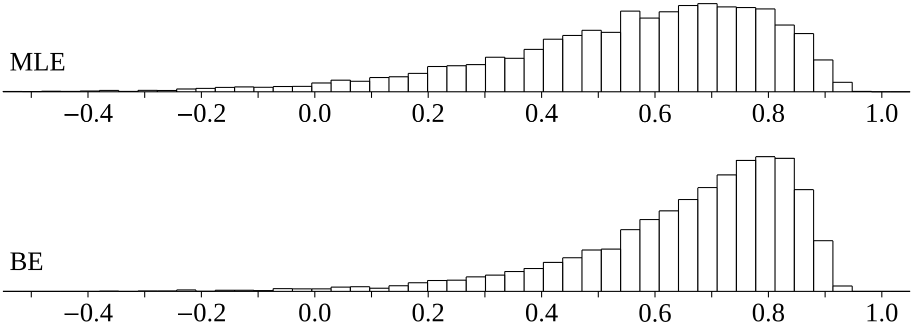
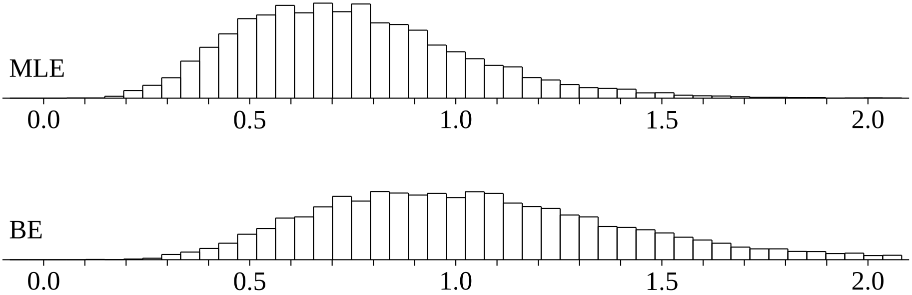


Figure 14: Empirical Distributions of σ_ϵ^2



To check whether M and N are enough large, Tables 3 – 5 are shown for BE. Comparison between Tables 3 and 4 shows whether $N = 5000$ is large enough and we can see from Tables 3 and 5 whether the burn-in period $M = 1000$ is large enough.

We can conclude that $N = 5000$ is enough if Table 3 is very close to Table 4 and that $M = 1000$ is enough if Table 3 is close to Table 5.

The difference between Tables 3 and 4 is at most 0.034 (see 90% in β_1) and that between Tables 3 and 5 is less than or equal to 0.013 (see Kurtosis in β_1).

Thus, all the three tables are very close to each other.

Therefore, we can conclude that $(M, N) = (1000, 5000)$ is enough.

For safety, hereafter we focus on the case of $(M, N) = (5000, 10^4)$.

We compare Tables 2 and 3.

Both MLE and BE give us the unbiased estimators of regression coefficients β_1, β_2 and β_3, because the arithmetic averages from the 10^4 estimates of β_1, β_2 and β_3, (i.e., AVE in the tables) are very close to the true parameter values, which are set to be $(\beta_1, \beta_2, \beta_3) = (10, 1, 1)$.

However, in the SER and RMSE criteria, BE is better than MLE, because SER and RMSE of BE are smaller than those of MLE. From Skewness and Kurtosis in the two tables, we can see that the empirical distributions of MLE and BE of $(\beta_1, \beta_2, \beta_3)$ are very close to the normal distribution. Remember that the skewness and kurtosis of the normal distribution are given by zero and three, respectively.

As for σ_ϵ^2, AVE of BE is closer to the true value than that of MLE, because AVE of MLE is 0.752 (see Table 2) and that of BE is 1.051 (see Table 3).

However, in the SER and RMSE criteria, MLE is superior to BE, since SER and RMSE of MLE are given by 0.276 and 0.372 (see Table 2) while those of BE are 0.380 and 0.384 (see Table 3).

The empirical distribution obtained from 10^4 estimates of σ_ϵ^2 is skewed to the right

(Skewness is positive for both MLE and BE) and has a larger kurtosis than the normal distribution because Kurtosis is greater than three for both tables.

For ρ, AVE of MLE is 0.559 (Table 2) and that of BE is given by 0.661 (Table 3).

As it is also seen in Figures 8 and 9, BE is less biased than MLE from the AVE criterion.

Moreover, SER and RMSE of MLE are 0.240 and 0.417, while those of BE are 0.188 and 0.304.

Therefore, BE is more efficient than MLE.

Thus, in the AVE, SER and RMSE criteria, BE is superior to MLE with respect to

ρ.

The empirical distributions of MLE and BE of ρ are skewed to the left because Skewness is negative, which value is given by -1.002 in Table 2 and -1.389 in Table 3.

We can see that MLE is less skewed than BE.

For Kurtosis, both MLE and BE of ρ are greater than three and therefore the empirical distributions of the estimates of ρ have fat tails, compared with the normal distribution.

Since Kurtosis in Table 3 is 5.391 and that in Table 2 is 4.013, the empirical distribution of BE has more kurtosis than that of MLE.

Figures 10 – 14 correspond to the empirical distributions for each parameter, which are constructed from the G estimates used in Tables 2 and 3.

As we can see from Skewness and Kurtosis in Tables 2 and 3, $\hat{\beta}_i$ and $\widetilde{\beta}_i$, $i = 1, 2, 3$, are very similar to normal distributions in Figures 10 – 12.

For β_i, $i = 1, 2, 3$, the empirical distributions of MLE have the almost same centers as those of BE, but the empirical distributions of MLE are more widely distributed than those of BE.

We can also observe these facts from AVEs and SERs in Tables 2 and 3.

In Figure 13, the empirical distribution of $\hat{\rho}$ is quite different from that of $\widetilde{\rho}$.

$\widetilde{\rho}$ is more skewed to the left than $\hat{\rho}$ and $\widetilde{\rho}$ has a larger kurtosis than $\hat{\rho}$.

Since the true value of ρ is 0.9, BE is distributed at the nearer place to the true value than MLE.

Figure 14 displays the empirical distributions of σ_ϵ^2. MLE $\hat{\sigma}_\epsilon^2$ is biased and under-estimated, but it has a smaller variance than BE $\widetilde{\sigma}_\epsilon^2$.

In addition, we can see that BE $\widetilde{\sigma}_\epsilon^2$ is distributed around the true value.

6.2.4 Summary

In Section 6.2, we have compared MLE with BE, using the regression model with the autocorrelated error term.

Chib (1993) applied the Gibbs sampler to the autocorrelation model, where the initial density of the error term is ignored.

Under this setup, the posterior distribution of ρ reduces to the normal distribution.

Therefore, random draws of ρ given β, σ_ϵ^2 and (y_t, X_t) can be easily generated.

However, when the initial density of the error term is taken into account, the posterior distribution of ρ is not normal and it cannot be represented in an explicit functional form.

Accordingly, in Section 6.2, the Metropolis-Hastings algorithm have been applied to generate random draws of ρ from its posterior density.

The obtained results are summarized as follows.

Given $\beta' = (10, 1, 1)$ and $\sigma^2 = 1$, in Figure 8 we have the relationship between ρ and $\hat{\rho}$, and $\widetilde{\rho}$ corresponding to ρ is drawn in Figure 9.

In the two figures, we can observe:

(i) both MLE and BE approach the true parameter value as n is large, and

(ii) BE is closer to the 45° degree line than MLE and accordingly BE is superior to

MLE.

Moreover, we have compared MLE with BE in Tables 2 and 3, where $\beta' = (10, 1, 1)$, $\rho = 0.9$ and $\sigma^2 = 1$ are taken as the true values.

As for the regression coefficient β, both MLE and BE gives us the unbiased estimators.

However, we have obtained the result that BE of β is more efficient than MLE. For estimation of σ^2,

BE is less biased than MLE.

In addition, BE of the autocorrelation coefficient ρ is also less biased than MLE.

Therefore, as for inference on β, BE is superior to MLE, because it is plausible to consider that the estimated variance of $\hat{\beta}$ is biased much more than that of $\widetilde{\beta}$. Remember that variance of $\hat{\beta}$ depends on both ρ and σ^2.

Thus, from the simulation studies, we can conclude that BE performs much better than MLE.

References

Amemiya, T., 1985, *Advanced Econometrics*, Cambridge:Harvard University Press.

Andrews, D.W.K., 1993, "Exactly Median-Unbiased Estimation of First Order Autoregressive / Unit Root Models," *Econometrica*, Vol.61, No.1, pp.139 – 165.

Bernardo, J.M. and Smith, A.F.M., 1994, *Bayesian Theory*, John Wiley & Sons.

Boscardin, W.J. and Gelman, A., 1996, "Bayesian Computation for parametric Models of Heteroscedasticity in the Linear Model," in *Advances in Econometrics, Vol.11 (Part A)*, edited by Hill, R.C., pp.87 – 109, Connecticut:JAI Press Inc.

Carlin, B.P. and Louis, T.A., 1996, *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall.

Chen, M.H., Shao, Q.M. and Ibrahim, J.G., 2000, *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag.

Chib, S., 1993, "Bayes Regression with Autoregressive Errors: A Gibbs Sampling Approach," *Journal of Econometrics*, Vol.58, No.3, pp.275 – 294.

Chib, S. and Greenberg, E., 1994, "Bayes Inference in Regression Models with ARMA(p, q) Errors," *Journal of Econometrics*, Vol.64, No.1&2, pp.183 – 206.

Chib, S. and Greenberg, E., 1995, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, Vol.49, No.4, pp.327 – 335.

Gamerman, D., 1997, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall.

Geweke, J., 1992, “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments,” in *Bayesian Statistics, Vol.4*, edited by Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M., pp.169 – 193 (with discussion), Oxford University Press.

Greene, W.H., 1997, *Econometric Analysis* (Third Edition), Prentice-Hall.

Harvey, A.C., 1976, “Estimating Regression Models with Multiplicative Heteroscedasticity,” *Econometrica*, Vol.44, No.3, pp.461 – 465.

Hogg, R.V. and Craig, A.T., 1995, *Introduction to Mathematical Statistics* (Fifth Edition), Prentice Hall.

Judge, G., Hill, C., Griffiths, W. and Lee, T., 1980, *The Theory and Practice of Econometrics*, John Wiley & Sons.

Mengersen, K.L., Robert, C.P. and Guihenneuc-Jouyaux, C., 1999, "MCMC Convergence Diagnostics: A Reviewww," in *Bayesian Statistics*, *Vol.6*, edited by Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M., pp.514 – 440 (with discussion), Oxford University Press.

O'Hagan, A., 1994, *Kendall's Advanced Theory of Statistics*, Vol.2B (Bayesian

Inference), Edward Arnold.

Ohtani, K., 1982, “Small Sample Properties of the Two-step and Three-step Estimators in a Heteroscedastic Linear Regression Model and the Bayesian Alternative,” *Economics Letters*, Vol.10, pp.293 – 298.

Robert, C.P. and Casella, G., 1999, *Monte Carlo Statistical Methods*, Springer-Verlag.

Smith, A.F.M. and Roberts, G.O., 1993, “Bayesian Computation via Gibbs Sampler and Related Markov Chain Monte Carlo Methods,” *Journal of the Royal Statistical Society*, Ser.B, Vol.55, No.1, pp.3 – 23.

Tanizaki, H., 2000, “Bias Correction of OLSE in the Regression Model with Lagged Dependent Variables,” *Computational Statistics and Data Analysis*, Vol.34, No.4, pp.495 – 511.

Tanizaki, H., 2001, “On Least-Squares Bias in the AR(p) Models: Bias Correction Using the Bootstrap Methods,” Unpublished Manuscript.

Tanizaki, H. and Zhang, X., 2001, “Posterior Analysis of the Multiplicative Heteroscedasticity Model,” *Communications in Statistics, Theory and Methods*, Vol.30, No.2, pp.855 – 874.

Tierney, L., 1994, “Markov Chains for Exploring Posterior Distributions,” *The An-*

nals of Statistics, Vol.22, No.4, pp.1701 – 1762.

Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons.

6.3 Marginal Likelihood, Convergence Diagnostic and so on

6.3.1 Marginal Likelihood (周辺尤度)

Model Selection $\Longrightarrow$ Marginal Likelihood

$$f_y(y) = \int f_{y|\theta}(y|\theta)\, f_\theta(\theta)\mathrm{d}\theta$$

Evaluation of Marginal Likelihood $\Longrightarrow$ Proper Prior

(i) Importance Sampling: Use of Prior Distribution

$$f_y(y) = \mathrm{E}_\theta(f_{y|\theta}(y|\theta)) \approx \frac{1}{N}\sum_{i=1}^{N} f_{y|\theta}(y|\theta_i),$$

where θ_i is the ith random draw generated from the prior distribution $f_\theta(\theta)$.

(ii) Importance Sampling: Use of the Appropriate Importance Distribution

$$f_y(y) = \int \frac{f_{y|\theta}(y|\theta)\, f_\theta(\theta)}{g(\theta)} g(\theta)\mathrm{d}\theta = \mathrm{E}\Big(\frac{f_{y|\theta}(y|\theta)\, f_\theta(\theta)}{g(\theta)} \Big)$$
$$\approx \frac{1}{N} \sum_{i=1}^{N} \frac{f_{y|\theta}(y|\theta_i)\, f_\theta(\theta_i)}{g(\theta_i)},$$

where θ_i is the ith random draw generated from the appropriately chosen importance distribution $g(\theta)$.

(iii) Harmonic Mean $\Longrightarrow$ Gelfand and Dey (1994) and Newton and Raftery (1994)

$$\frac{1}{f_y(y)} = \int \frac{g(\theta)}{f_y(y)} \mathrm{d}\theta = \int \frac{g(\theta)}{f_y(y)\, f_{\theta|y}(\theta|y)} f_{\theta|y}(\theta|y) \mathrm{d}\theta$$

$$= \int \frac{g(\theta)}{f_{y|\theta}(y|\theta)\, f_\theta(\theta)} f_{\theta|y}(\theta|y) \mathrm{d}\theta \approx \frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta_i)}{f_{y|\theta}(y|\theta_i)\, f_\theta(\theta_i)},$$

where θ_i is the ith random draw generated from the posterir distribution $f_{\theta|y}(\theta|y)$.

Thus, the marginal distribution is evaluated by:

$$f_y(y) \approx \left(\frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta_i)}{f_{y|\theta}(y|\theta_i)\, f_\theta(\theta_i)} \right)^{-1}, \qquad \Longrightarrow \qquad \text{Gelfand and Dey (1994).}$$

When $g(\theta) = f_\theta(\theta)$ is taken, the marginal distribution is given by:

$$f_y(y) \approx \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{f_{y|\theta}(y|\theta_i)} \right)^{-1}, \qquad \Longrightarrow \qquad \text{Newton and Raftery (1994).}$$

(iv) Chib (1995) and Chib and Jeliazkov (2001)

$$f_y(y) = \frac{f_{y|\theta}(y|\theta)\, f_\theta(\theta)}{f_{\theta|y}(\theta|y)}$$

$$\log f_y(y) = \log f_{y|\theta}(y|\hat{\theta}) + \log f_\theta(\hat{\theta}) - \log f_{\theta|y}(\hat{\theta}|y),$$

where $\hat{\theta}$ denotes the Bayes estimates.

We need to evaluate $\log f_{\theta|y}(\hat{\theta}|y)$, using the Gibbs sampler or the MH algorithm.

6.3.2 Convergence Diagnostic (収束判定)

We need to check whether the **burn-in period** is enough and whether MCMC converges to the **invariant distribution (不変分布)**.

Geweke (1992)

Divide the sample path into three periods, excluding the burn-in period..

Test whether the first period is different from the third period.

Suppose that we have the MCMC sequence, i.e., $\theta_{-M+1}, \cdots, \theta_0, \theta_1, \cdots, \theta_N$.

The burn-in period is denoted by $\theta_{-M+1}, \cdots, \theta_0$.

$\theta_1, \cdots, \theta_N$ are divided by three periods.

The first period is given by $\theta_1, \cdots, \theta_{N_1}$.

The second period is given by $\theta_{N_1+1}, \cdots, \theta_{N_2}$.

The third period is given by $\theta_{N_2+1}, \cdots, \theta_N$.

Consider a function $g(\cdot)$.

Define $\quad \overline{g}_1 = \dfrac{1}{N_1}\displaystyle\sum_{i=1}^{N_1} g(\theta_i) \quad$ and $\quad \overline{g}_3 = \dfrac{1}{N_3}\displaystyle\sum_{i=N_1+N_2+1}^{N} g(\theta_i) \quad$ for $N_3 = N - N_2 - N_1$.

Estimate $\quad \dfrac{1}{N_1}\mathrm{V}(\displaystyle\sum_{i=1}^{N_1} g(\theta_i)) \quad$ and $\quad \dfrac{1}{N_3}\mathrm{V}(\displaystyle\sum_{i=N_1+N_2+1}^{N} g(\theta_i)),$

which are denoted by s_1^2 and s_3^2, respectively.

By the central limit theorem,

$$\frac{\overline{g}_1 - \mathrm{E}(\overline{g}_1)}{s_1/\sqrt{N_1}} \longrightarrow N(0,1) \qquad \text{and} \qquad \frac{\overline{g}_3 - \mathrm{E}(\overline{g}_3)}{s_3/\sqrt{N_3}} \longrightarrow N(0,1).$$

Therefore, under the null hypothesis $H_0: \mathrm{E}(\overline{g}_1) = \mathrm{E}(\overline{g}_3)$,

$$\frac{\overline{g}_1 - \overline{g}_3}{\sqrt{s_1^2/N_1 + s_3^2/N_3}} \longrightarrow N(0,1).$$

The case of $g(\theta_i) = \theta_i \implies$ Testing whether the two means (i.e., first-moments) are equal.

The case of $g(\theta_i) = \theta_i^2 \implies$ Testing whether the two second-moments are equal.

Computation of s_1^2 and s_3^2 has to be careful, because $g(\theta_1)$, $\cdots$, $g(\theta_N)$ are serially correlated.

$\Longrightarrow$ Long-run variance.

Take an example of s_1^2, which is an estimate of $\dfrac{1}{N_1}\mathrm{V}(\sum_{i=1}^{N_1} g(\theta_i))$.

$$
\begin{aligned}
\frac{1}{N_1}\mathrm{V}(\sum_{i=1}^{N_1} g(\theta_i)) &= \frac{1}{N_1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\mathrm{Cov}(g(\theta_i), g(\theta_j)) \\
&= \frac{1}{N_1}(N_1\gamma(0) + 2(N_1-1)\gamma(1) + 2(N_1-2)\gamma(2) + \cdots + 2\gamma(N_1-1)) \\
&= \gamma(0) + 2\sum_{\tau=1}^{N_1-1} k(\frac{\tau}{N_1})\gamma(\tau), \quad \Longrightarrow \quad \text{Bartlett Kernel (Newy-West Est.)}
\end{aligned}
$$

where $\gamma(\tau) = \mathrm{Cov}(g(\theta_i), g(\theta_{i+\tau}))$.

We may choose the other kernels (for example, Parzen kernel or second-order spectrum kernel; see p.166-167) for $k(x)$.

Thus, s_1^2 is estimated by:

$$s_1^2 = \hat{\gamma}(0) + 2\sum_{\tau=1}^{q} k(\frac{\tau}{q+1})\hat{\gamma}(\tau),$$

for $q \leq N_1 - 1$. $\Longrightarrow$ Choice of q and $k(\cdot)$.