3 Bayesian Methods for Estimating Structural Equation Models

3.1 Introduction

In Chapter 2, we have presented some basic SEMs, and discussed how these models can be used in practice. In substantive applications of these basic SEMs or their generalizations (to be discussed in subsequent chapters) for coping with complex situations, it is important to introduce sound statistical methods that give accurate statistical results. The traditional approach for analyzing SEMs is the covariance structure analysis approach. In this approach, the statistical theory as well as the computational algorithms are developed on the basis of the sample covariance matrix $\mathbf{S}$ and its asymptotic distribution. Under some standard assumptions, for example the random observations are i.i.d. following a normal distribution, this approach works fine. As a result, almost all classical commercial software in SEMs were developed on the basis of this approach with the sample covariance matrix $\mathbf{S}$. Unfortunately, under slightly more complex situations that are common in substantive research, the covariance structure analysis approach based on $\mathbf{S}$ is not effective and may encounter theoretical and computational problems. For instance, due to the presence of the nonlinear terms of explanatory latent variables, outcome latent variables and the related observed variables in $\mathbf{y}_i$ are not normally distributed. Hence, $\mathbf{S}$ is not suitable for modeling the nonlinear relationships; and the application of the covariance structure analysis approach via some unnatural methods, for example the product indicator method, produces inferior results. See Lee (2007) for more discussions on the disadvantages in using the covariance structure analysis approach to analyze subtle SEMs and data structures that are common in substantive research. In this chapter, we will introduce an attractive Bayesian approach which can be effectively applied to analyze not only the standard SEMs but also their useful generalizations that are developed in recent years.

The basic nice feature of a Bayesian approach is its flexibility in utilizing useful prior information for achieving better results. In many practical problems, statisticians may have good prior information from some sources, for example the knowledge of the experts, and analyses of similar data and/or past data. For situations without accurate prior information, some types of noninformative prior distributions can be used in a Bayesian approach. In these cases, the accuracy of the Bayesian estimates is close to that of the maximum likelihood (ML) estimates.

It is well known that the statistical properties of the ML approach are asymptotic. Hence, they are valid for situations with large sample sizes. In the context of some basic SEMs, many studies (see, for example, Boomsma, 1982; Chou, Bentler and Satorra, 1991; Hu, Bentler and Kano, 1992; Hoogland and Boomsma, 1998) showed that the statistical properties of the ML approach are not robust to small sample sizes. On the contrary, as pointed out by many important articles in Bayesian analyses of SEMs (see Scheines, Hoijtink and Boomsma, 1999; Ansari and Jedidi, 2000; Dunson, 2000; Lee and Song, 2004), the sampling-based Bayesian methods depend less on asymptotic theory, and hence have the potential to produce reliable results even with small samples.

Recently, Bayesian methods are developed with various Markov chain Monte Carlo (MCMC) algorithms. Usually, a sufficiently large number of observations are simulated from the joint posterior distribution through these MCMC algorithms. Means as well as quantiles of this joint posterior distribution can be estimated from the simulated observations. These quantities are useful for making statistical inferences. For example, the Bayesian estimates of the unknown parameters and the latent variables can be obtained from the corresponding sample means of observations simulated from the posterior distribution. Moreover, from these estimates, the estimated residuals which are useful for assessing the goodness-of-fit of the proposed model and for detecting outliers, can be obtained. Finally, various model comparison statistics that are closely related to the Bayesian approach, such as the Bayes factor, give more flexible and natural tools for model comparison than the classical likelihood ratio test (see Kass and Raftery, 1995; Lee, 2007). We will give a detailed discussion on the model comparison in Chapter 4.

Basic statistical inferences of SEMs include estimation of the unknown parameters and latent variables, assessment of goodness-of-fit of the proposed model, and model comparison. The objective of this chapter is to provide an introduction of the Bayesian approach to conduct statistical inferences of SEMs. It is not our intention to present a full coverage on the general Bayesian theory. Readers may refer to other excellent books, such as Box and Tiao (1973), and Gelman *et al.* (2003) for more details. Section 3.2 of this chapter presents the basic ideas of the Bayesian approach in estimation, including the discussion of the prior distribution. Posterior analyses through applications of some MCMC methods are considered in Section 3.3. An application of the MCMC methods is presented in Section 3.4. Section 3.5 describes how to apply the software WinBUGS to obtain Bayesian estimation and to conduct simulation studies. Some technical details are given in the appendices.

3.2 Basic Concepts of the Bayesian Estimation and Prior Distributions

The Bayesian approach is well recognized in the statistics literature as an attractive approach in analyzing a wide variety of models (Berger, 1985; Congdon, 2003). To introduce this approach for SEMs, we let M be an arbitrary SEM with a vector of unknown parameters $\boldsymbol{\theta}$, and let $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ be the observed data set of raw observations with a sample size n. In a non-Bayesian approach, $\boldsymbol{\theta}$ is not considered as random. In a Bayesian approach, $\boldsymbol{\theta}$ is considered to be random with a distribution (called prior distribution) and an associated (prior) density function, say, $p(\boldsymbol{\theta}|M)$. We give here a simple example to show the rationale for regarding an unknown parameter as random. Suppose that we wish to estimate the mean of systolic blood pressure, say μ, it is not necessary to assume that μ is fixed with a certain value; instead, μ is allowed to vary randomly, for example with a higher (or lower) probability at some values. Hence, it is more reasonable to treat μ as random with a prior distribution and a prior density $p(\mu)$. See Berger (1985) and the references therein for theoretical and practical rationales for treating $\boldsymbol{\theta}$ as random.

For simplicity, we use $p(\boldsymbol{\theta})$ to denote $p(\boldsymbol{\theta}|M)$. Bayesian estimation is based on the observed data $\mathbf{Y}$ and the prior distribution of $\boldsymbol{\theta}$. Let $p(\mathbf{Y}, \boldsymbol{\theta}|M)$ be the probability density function of the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\theta}$ under M. The behavior of $\boldsymbol{\theta}$ under the given data $\mathbf{Y}$ is fully described by the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{Y}$. This conditional distribution is called the posterior distribution of $\boldsymbol{\theta}$. Let $p(\boldsymbol{\theta}|\mathbf{Y}, M)$ be the density function of the posterior distribution, which is called the posterior density function. The posterior distribution of $\boldsymbol{\theta}$ or its density plays the most important role in the Bayesian analysis of the model. Based on a well-known identity in probability, we have $p(\mathbf{Y}, \boldsymbol{\theta}|M) = p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{Y}, M)p(\mathbf{Y}|M)$. As $p(\mathbf{Y}|M)$ does not depend on $\boldsymbol{\theta}$, and can be regarded as a constant with fixed $\mathbf{Y}$, we have

$$\begin{aligned} &p(\boldsymbol{\theta}|\mathbf{Y}, M) \propto p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}), \qquad \text{or} \\ &\log p(\boldsymbol{\theta}|\mathbf{Y}, M) = \log p(\mathbf{Y}|\boldsymbol{\theta}, M) + \log p(\boldsymbol{\theta}) + \text{constant}. \end{aligned} \tag{3.1}$$

Note that $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ can be regarded as the likelihood function because it is the probability density of $\mathbf{y}_1, \cdots, \mathbf{y}_n$ conditional on the parameter vector $\boldsymbol{\theta}$. It follows from (3.1) that the posterior density function incorporates the sample information through the likelihood function $p(\mathbf{Y}|\boldsymbol{\theta}, M)$, and the prior information through the prior density function $p(\boldsymbol{\theta})$. Note also that $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ depends on the sample size, whereas $p(\boldsymbol{\theta})$ does not. When the sample size becomes arbitrarily large, $\log(\mathbf{Y}|\boldsymbol{\theta}, M)$ could be very large and hence $\log p(\mathbf{Y}|\boldsymbol{\theta}, M)$ dominates $\log p(\boldsymbol{\theta})$. In this situation, the prior distribution of $\boldsymbol{\theta}$ plays a less important role, and the logarithm of posterior density function $\log p(\boldsymbol{\theta}|\mathbf{Y}, M)$ is close to the log-likelihood function $\log p(\mathbf{Y}|\boldsymbol{\theta}, M)$. Hence, asymptotically Bayesian and ML approaches are equivalent, and the Bayesian estimates have the same optimal properties as the ML estimates. When the sample sizes are small or moderate, the prior distribution of $\boldsymbol{\theta}$ plays a more substantial role in Bayesian estimation. Hence, in substantive research problems where the sample sizes are small or moderate, prior information of the parameter vector $\boldsymbol{\theta}$ incorporated into the Bayesian analysis is useful for achieving better results (see below for the utilization of useful prior information in the analysis). For many problems in biomedical and behavioral sciences, researchers may have good prior information from the subject experts, from analyses of similar or past data, or from some other sources. As more accurate results can be obtained by incorporating appropriate prior information in the analysis through the prior distribution of $\boldsymbol{\theta}$, the selection of $p(\boldsymbol{\theta})$ is an important issue in Bayesian analysis. In the following sections and chapters, the symbol M will be suppressed if the context is clear; for example, $p(\boldsymbol{\theta}|\mathbf{Y})$ will denote the posterior density of $\boldsymbol{\theta}$ under M, and $[\boldsymbol{\theta}|\mathbf{Y}]$ will denote the posterior distribution of $\boldsymbol{\theta}$ under M.

3.2.1 Prior Distributions

The prior distribution of $\boldsymbol{\theta}$ represents the distribution of possible parameter values, from which the parameter $\boldsymbol{\theta}$ has been drawn. Basically, there are two kinds of prior distributions, namely the noninformative and the informative prior distributions. Noninformative prior distributions associate with situations when the prior distributions have no population basis. They are used when we have little prior information, and hence the prior distributions play a minimal role in the posterior distribution of $\boldsymbol{\theta}$. The associated prior densities are chosen to be vague, diffuse, flat, or noninformative, for example a density that is proportional to a constant or has a huge variance. In this case, the Bayesian estimation is unaffected by information external to the observed data. For informative prior distribution, we may have useful prior knowledge about this distribution, either from closely related data or from subjective knowledge of experts. Usually, an informative prior distribution has its own parameters, which are called hyperparameters.

A commonly used informative prior distribution in the general Bayesian approach for statistical problems is the conjugate prior distribution. We consider the univariate binomial model to motivate this kind of prior distribution. Considered as a function of θ, the likelihood of an observation y is of the form

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

If the prior density of θ is of the same form, it can be seen from (3.1) that the posterior density will also be of this form. More specifically, consider the following prior density of θ:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{3.2}$$

which is a beta distribution with hyperparameters α and β. Then,

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}, \end{aligned} \tag{3.3}$$

which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$. We see that $p(\theta)$ and $p(\theta|y)$ are of the same form. The property that the posterior distribution follows the same parametric form as the prior distribution is called conjugacy, and the prior distribution is called a conjugate prior distribution (Gelman *et al.*, 2003). One advantage of this kind of prior distribution is providing a manageable posterior distribution for developing the MCMC algorithm for statistical inference.

If the hyperparameters in the conjugate prior distributions are unknown, then they may be treated as unknown parameters and thus have their own prior distributions in a full Bayesian analysis. These hyperprior distributions again have their own hyperparameters. As a result, the problem will become very tedious. Hence, in developing the Bayesian methods for analyzing SEMs, we usually assign fixed known values to the hyperparameters in the conjugate prior distributions.

3.2.2 Conjugate Prior Distributions in Bayesian Analyses of SEMs

In the field of SEMs, almost all existing work in Bayesian analysis used conjugate prior distributions with the given hyperparameter values; see Lee (2007) and the references therein. It has been shown that these distributions work well for many SEMs. Therefore, in this book, we will use the conjugate prior distributions in our Bayesian analyses. In general, it has been shown that for an univariate normal distribution, the conjugate prior distributions of the unknown mean and variance are normal and inverted Gamma, respectively (see Gelman *et al.*, 2003; Lee, 2007). This fact motivates the selection of conjugate prior distributions for the parameters in SEMs, which are basically the regression coefficients related to the mean vector of a multivariate normal distribution, and variance and covariance matrix related to the residual errors and latent vector, respectively.

Without loss of generality, we illustrate the selection of prior distributions in the context of a nonlinear SEM with fixed covariates in the structural equation. More specifically, we first consider the following measurement equation and structural equation of the model:

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{3.4}$$

$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{3.5}$$

where $\mathbf{y}_i$ is a $p \times 1$ vector of observed variables, $\boldsymbol{\mu}$ is a vector of intercepts, $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$ is a vector of latent variables which is partitioned into a $q_1 \times 1$ vector of outcome latent variables $\boldsymbol{\eta}_i$ and a $q_2 \times 1$ vector of explanatory latent variables $\boldsymbol{\xi}_i$, $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\delta}_i$ are residual errors, $\mathbf{d}_i$ is an $r \times 1$ vector of fixed covariates, $\boldsymbol{\Lambda}$, $\mathbf{B}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Gamma}$ are parameter matrices of unknown regression coefficients, and $\mathbf{F}(\cdot)$ is a given vector of differentiable functions of $\boldsymbol{\xi}_i$. Similar to the model described in Chapter 2, the distributions of $\boldsymbol{\xi}_i$, $\boldsymbol{\epsilon}_i$, and $\boldsymbol{\delta}_i$ are $N[\mathbf{0}, \boldsymbol{\Phi}]$, $N[\mathbf{0}, \boldsymbol{\Psi}_\epsilon]$, and $N[\mathbf{0}, \boldsymbol{\Psi}_\delta]$, respectively; and the assumptions as given in Chapter 2 are satisfied. In this model, the unknown parameters are $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, $\mathbf{B}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Gamma}$ which are related to the mean vectors of $\mathbf{y}_i$ and $\boldsymbol{\eta}_i$; and $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}_\epsilon$, and $\boldsymbol{\Psi}_\delta$ which are the covariance matrices. Now consider the prior distributions of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}_\epsilon$ that are involved in the measurement equation. Let $\boldsymbol{\Lambda}_k^T$ be the kth row of $\boldsymbol{\Lambda}$, and $\psi_{\epsilon k}$ be the kth diagonal element of $\boldsymbol{\Psi}_\epsilon$. It can be shown (see Lee, 2007) that the conjugate type prior distributions of $\boldsymbol{\mu}$ and $(\mathbf{\Lambda}_k, \psi_{\epsilon k})$ are

$$\begin{aligned} &\psi_{\epsilon k} \overset{D}{=} \text{Inverted } Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}] \text{ or equivalently } \psi_{\epsilon k}^{-1} \overset{D}{=} Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}], \\ &\boldsymbol{\mu} \overset{D}{=} N[\boldsymbol{\mu}_0, \mathbf{\Sigma}_0], \text{ and } [\mathbf{\Lambda}_k|\psi_{\epsilon k}] \overset{D}{=} N\left[\mathbf{\Lambda}_{0k}, \psi_{\epsilon k}\mathbf{H}_{0yk}\right], \end{aligned} \tag{3.6}$$

where $\alpha_{0\epsilon k}, \beta_{0\epsilon k}$, and elements in $\boldsymbol{\mu}_0, \mathbf{\Lambda}_{0k}, \mathbf{\Sigma}_0$, and $\mathbf{H}_{0yk}$ are hyperparameters, and $\mathbf{\Sigma}_0$ and $\mathbf{H}_{0yk}$ are positive definite matrices. For simplicity of notation, we rewrite the structural equation (3.5) as:

$$\boldsymbol{\eta}_i = \mathbf{\Lambda}_\omega \mathbf{G}(\boldsymbol{\omega}_i) + \boldsymbol{\delta}_i, \tag{3.7}$$

where $\mathbf{\Lambda}_\omega = (\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma})$ and $\mathbf{G}(\boldsymbol{\omega}_i) = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \mathbf{F}(\boldsymbol{\xi}_i)^T)^T$. Let $\mathbf{\Lambda}_{\omega k}^T$ be the kth row of $\mathbf{\Lambda}_\omega$, and $\psi_{\delta k}$ be the kth diagonal element of $\mathbf{\Psi}_\delta$. Based on similar reasoning as before, the conjugate type prior distributions of $\mathbf{\Phi}$ and $(\mathbf{\Lambda}_{\omega k}, \psi_{\delta k})$ are:

$$\begin{aligned} &\mathbf{\Phi} \overset{D}{=} IW_{q_2}[\mathbf{R}_0^{-1}, \rho_0], \text{ or equivalently } \mathbf{\Phi}^{-1} \overset{D}{=} W_{q_2}[\mathbf{R}_0, \rho_0], \\ &\psi_{\delta k} \overset{D}{=} \text{Inverted } Gamma[\alpha_{0\delta k}, \beta_{0\delta k}] \text{ or equivalently } \psi_{\delta k}^{-1} \overset{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}], \\ &[\mathbf{\Lambda}_{\omega k}|\psi_{\delta k}] \overset{D}{=} N[\mathbf{\Lambda}_{0\omega k}, \psi_{\delta k}\mathbf{H}_{0\omega k}], \end{aligned} \tag{3.8}$$

where $W_{q_2}[\mathbf{R}_0, \rho_0]$ is a q_2-dimensional Wishart distribution with hyperparameters ρ_0 and a positive definite matrix $\mathbf{R}_0$, $IW_{q_2}[\mathbf{R}_0^{-1}, \rho_0]$ is a q_2-dimensional inverted Wishart distribution with hyperparameters ρ_0 and a positive definite matrix $\mathbf{R}_0^{-1}$, $\alpha_{0\delta k}, \beta_{0\delta k}$, and elements in $\mathbf{\Lambda}_{0\omega k}$ and $\mathbf{H}_{0\omega k}$ are hyperparameters, and $\mathbf{H}_{0\omega k}$ is a positive definite matrix. Note that the prior distribution of $\mathbf{\Phi}^{-1}$ (or $\mathbf{\Phi}$) is a multivariate extension of the prior distribution of $\psi_{\delta k}^{-1}$ (or $\psi_{\delta k}$). For clarity, the Gamma, inverted Gamma, Wishart, and inverted Wishart distributions as well as their characteristics are given in Appendix 3.1.

In specifying conjugate prior distributions, we assign values to their hyperparameters. These preassigned values (prior inputs) represent the available prior knowledge. In general, if we have confidence to have good prior information about a parameter, then it is advantageous to select the corresponding prior distribution with a small variance; otherwise the prior distribution with a larger variance should be selected. We use the prior distributions given in (3.6) to illustrate this. If we have confidence that the true $\mathbf{\Lambda}_k$ is not too far away from the preassigned hyperparameter value $\mathbf{\Lambda}_{0k}$, then $\mathbf{H}_{0yk}$ should be taken as a matrix with small variances (such as $0.5\mathbf{I}$). The choice of $\alpha_{0\epsilon k}$ and $\beta_{0\epsilon k}$ is based on the same general rationale and the nature of $\psi_{\epsilon k}$ in the model. First we note that the distribution of ϵ_k is $N[0, \psi_{\epsilon k}]$. Hence, if we think that the variation of ϵ_k is small (that is, $\mathbf{\Lambda}_k^T \boldsymbol{\omega}_i$ is a good predictor of y_{ik}), then the prior distribution of $\psi_{\epsilon k}$ should have a small mean value as well as a small variance. Otherwise, the prior distribution of $\psi_{\epsilon k}$ should have a large mean value and/or a large variance. This gives some idea in choosing the hyperparameters $\alpha_{0\epsilon k}$ and $\beta_{0\epsilon k}$ in the inverted Gamma distribution. Note that for the inverted Gamma distribution, the mean is equal to $\beta_{0\epsilon k}/(\alpha_{0\epsilon k} - 1)$, and the variance is equal to $\beta_{0\epsilon k}^2/\{(\alpha_{0\epsilon k} - 1)^2(\alpha_{0\epsilon k} - 2)\}$. Hence, we may take $\alpha_{0\epsilon k} = 9$ and $\beta_{0\epsilon k} = 4$ for a situation where we have confidence that $\mathbf{\Lambda}_k^T \boldsymbol{\omega}_i$ is a good predictor of y_{ik} in the measurement equation. Under this choice, the mean of $\psi_{\epsilon k}$ is $4/8 = 0.5$, and the variance of $\psi_{\epsilon k}$ is $4^2/\{(9-1)^2(9-2)\} = 1/28$. For a situation with little confidence, we may take $\alpha_{0k} = 6$ and $\beta_{0\epsilon k} = 10$, then the mean of $\psi_{\epsilon k}$ is 2.0 and the variance is 1.0. The above ideas for choosing preassigned hyperparameter values can be similarly used in specifying $\mathbf{\Lambda}_{0\omega k}$, $\alpha_{0\delta k}$, and $\beta_{0\delta k}$ in the conjugate prior distributions of $\mathbf{\Lambda}_{\omega k}$ and $\psi_{\delta k}$; see (3.8). Now, we consider the choice of $\mathbf{R}_0$ and ρ_0 in the prior distribution of $\mathbf{\Phi}$. It follows from Muirhead (1982, pp.97) that the mean of $\mathbf{\Phi}$ is $\mathbf{R}_0^{-1}/(\rho_0 - q_2 - 1)$. Hence, if we have confidence that $\mathbf{\Phi}$ is not too far away from a known matrix $\mathbf{\Phi}_0$, we can choose $\mathbf{R}_0^{-1}$ and ρ_0 such that $\mathbf{R}_0^{-1} = (\rho_0 - q_2 - 1)\mathbf{\Phi}_0$. Other values of $\mathbf{R}_0^{-1}$ and ρ_0 may be considered for situations without good prior information.

Now, we discuss some methods to get $\mathbf{\Lambda}_{0k}$, $\mathbf{\Lambda}_{0\omega k}$, and $\mathbf{\Phi}_0$. As mentioned before, these hyperparameter values may be obtained from subjective knowledge of the field experts, and/or analysis of past or closely related data. If this kind of information is not available and the sample size is small, we may consider using the following noninformative prior distributions:

$$\begin{aligned} &p(\mathbf{\Lambda}, \mathbf{\Psi}_\epsilon) \propto p(\psi_{\epsilon 1}, \cdots, \psi_{\epsilon p}) \propto \prod_{k=1}^{p} \psi_{\epsilon k}^{-1}, \\ &p(\mathbf{\Lambda}_\omega, \mathbf{\Psi}_\delta) \propto p(\psi_{\delta 1}, \cdots, \psi_{\delta q_1}) \propto \prod_{k=1}^{q_1} \psi_{\delta k}^{-1}, \\ &p(\mathbf{\Phi}) \propto |\mathbf{\Phi}|^{-(q_2+1)/2}. \end{aligned} \tag{3.9}$$

In (3.9), the prior distributions of the unknown parameters in $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_\omega$ are implicitly taken to be proportional to a constant. Note that no hyperparameters are involved in these noninformative prior distributions. Bayesian analysis on the basis of the above noninformative prior distributions is basically close to the Bayesian analysis with conjugate prior distributions given by (3.6) and (3.8) with very large variances. If the sample size is large, one possible method to get $\mathbf{\Lambda}_{0k}$, $\mathbf{\Lambda}_{0\omega k}$, and $\mathbf{\Phi}_0$ is to use a portion of the data, say one-third or less, to conduct an auxiliary Bayesian estimation with noninformative priors to get initial Bayesian estimates. The remaining data are then used to conduct the actual Bayesian analysis with the initial Bayesian estimates as hyperparameter values in relation to $\mathbf{\Lambda}_{0k}$, $\mathbf{\Lambda}_{0\omega k}$, and $\mathbf{\Phi}_0$. For situations with moderate sample sizes, Bayesian analysis may be done by applying data dependent prior inputs that are obtained from an initial estimation with the whole data set. Although the above methods are reasonable, we emphasize that we are not routinely recommending them for every practical application. In general, the issue of choosing prior inputs should be carefully approached on a problem-by-problem basis. Moreover, under situations without useful prior information, a sensitivity analysis should be conducted to see whether the results are robust to prior inputs. This can be done by perturbing the given hyperparameter values or considering some ad hoc prior inputs.

3.3 Posterior Analysis Using MCMC Methods

Bayesian estimate of $\boldsymbol{\theta}$ is usually defined as the mean or the mode of the posterior distribution $[\boldsymbol{\theta}|\mathbf{Y}]$. In this book, we are mainly interested in estimating the unknown parameters via the mean of the posterior distribution. Theoretically, it could be obtained via integration. For most situations, the integration does not have a closed form. However, if we can simulate a sufficiently large number of observations from $[\boldsymbol{\theta}|\mathbf{Y}]$ (or $p(\boldsymbol{\theta}|\mathbf{Y})$), we can approximate the mean and other useful statistics through the simulated observations. Hence, to solve the problem, it suffices to develop efficient and dependable methods for drawing observations from the posterior distribution. For most nonstandard SEMs, the posterior distribution $[\boldsymbol{\theta}|\mathbf{Y}]$ is complicated. It is difficult to derive this distribution and simulate observations from it. A major breakthrough for posterior simulation is the idea of data augmentation proposed by Tanner and Wong (1987). The strategy is to treat latent quantities as hypothetical missing data and to augment the observed data with them so that the posterior distribution based on the complete data set is relatively easy to analyze. This strategy has been widely applied in analyzing many statistical models (see, for example, Rubin, 1991; Albert and Chib, 1993; Dunson, 2000; among many others). It is particularly useful for SEMs which involve latent variables, see Lee (2007). The feature that makes SEMs different from the common regression model and the simultaneous equation model is the existence of random latent variables. In many situations, the presence of latent variables causes the major difficulties in the analysis of the model. However, if the random latent variables are given, SEMs will become familiar regression models that can be handled without much difficulty.

Hence, the abovementioned strategy based on data augmentation provides a useful approach to cope with the problem that is induced by latent variables. By augmenting the observed variables in complicated SEMs with the latent variables that are treated as hypothetical missing data, we can obtain the Bayesian solution based on the complete data set. More specifically, instead of working on the intractable posterior density $p(\boldsymbol{\theta}|\mathbf{Y})$, we will work on $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, where $\boldsymbol{\Omega}$ is the set of latent variables in the model. For most cases, $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ is still not in closed form and it is difficult to directly deal with it. However, on the basis of the complete data set $(\boldsymbol{\Omega}, \mathbf{Y})$, the conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ is usually standard, and the conditional distribution $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$ can also be derived from the definition of the model without much difficulty. As a result, we can apply some MCMC methods to simulate observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$ by drawing observations iteratively from their full conditional densities $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$. Following the terminology in MCMC methods, we may call $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$ and $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$ conditional distributions if the context is clear. Note that as $\boldsymbol{\Omega}$ is given in $p(\boldsymbol{\theta}|\boldsymbol{\Omega}, \mathbf{Y})$, the derivation of this conditional distribution is possible. A useful algorithm to achieve this goal is the following Gibbs sampler (Geman and Geman, 1984).

In the model M, suppose the parameter vector $\boldsymbol{\theta}$ and the latent matrix $\boldsymbol{\Omega}$ are respectively decomposed into the following components: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_a)$ and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_b)$. The Gibbs sampler is a Markov chain Monte Carlo algorithm which performs an alternating conditional sampling at each of its iteration. It cycles through the components of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$, drawing each component conditional on the values of all the other components. More specifically, at the jth iteration with current values

$\boldsymbol{\theta}^{(j)} = (\boldsymbol{\theta}_1^{(j)}, \cdots, \boldsymbol{\theta}_a^{(j)})$ and $\boldsymbol{\Omega}^{(j)} = (\boldsymbol{\Omega}_1^{(j)}, \cdots, \boldsymbol{\Omega}_b^{(j)})$, it simulates in turn,

$$\begin{aligned}
\boldsymbol{\theta}_1^{(j+1)} &\text{ from } \ p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(j)}, \cdots, \boldsymbol{\theta}_a^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\
\boldsymbol{\theta}_2^{(j+1)} &\text{ from } \ p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(j+1)}, \cdots, \boldsymbol{\theta}_a^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\
&\vdots \qquad\qquad \vdots \\
\boldsymbol{\theta}_a^{(j+1)} &\text{ from } \ p(\boldsymbol{\theta}_a|\boldsymbol{\theta}_1^{(j+1)}, \cdots, \boldsymbol{\theta}_{a-1}^{(j+1)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\
\boldsymbol{\Omega}_1^{(j+1)} &\text{ from } \ p(\boldsymbol{\Omega}_1|\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_2^{(j)}, \cdots, \boldsymbol{\Omega}_b^{(j)}, \mathbf{Y}), \\
\boldsymbol{\Omega}_2^{(j+1)} &\text{ from } \ p(\boldsymbol{\Omega}_2|\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_1^{(j+1)}, \cdots, \boldsymbol{\Omega}_b^{(j)}, \mathbf{Y}), \\
&\vdots \qquad\qquad \vdots \\
\boldsymbol{\Omega}_b^{(j+1)} &\text{ from } \ p(\boldsymbol{\Omega}_b|\boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_1^{(j+1)}, \cdots, \boldsymbol{\Omega}_{b-1}^{(j+1)}, \mathbf{Y}).
\end{aligned} \tag{3.10}$$

There are $a + b$ steps in the jth iteration of the Gibbs sampler. At each step, each component in $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ is updated conditionally on the latest values of the other components. We may simulate the components in $\boldsymbol{\Omega}$ first, then the components in $\boldsymbol{\theta}$; or vice versa. Most of the full conditional distributions in (3.10) are the normal, Gamma, and inverted Wishart distributions. Simulating observations from them is straightforward and fast. For nonstandard conditional distributions, the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970) may be used for efficient simulation. A brief description of the MH algorithm is given in Appendix 3.2.

It has been shown (Geman and Geman, 1984) that under mild regularity conditions, the joint distribution of $(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Omega}^{(j)})$ converges to the desired posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$ after a sufficiently large number of iterations, say J. It should be noted that if the iterations have not proceeded long enough, the simulated observations may not be representative of the posterior distribution. Moreover, even the algorithm has reached approximate convergence, observations obtained at the early iterations should be discarded because they still do not belong to the desired posterior distribution. The required number of iterations for achieving convergence of the Gibbs sampler, that is the burn-in iterations J, can be determined by plots of the simulated sequences of the individual parameters. At convergence, parallel sequences generated with different starting values should mix well together. Examples of sequences from which convergence looks reasonable, and sequences that have not reached convergence are presented in Figure 3.1. A minor problem with iterative simulation draws is their within-sequence correlation. In general, statistical inference from correlated observations is less precise than that from the same number of independent observations. To obtain a less correlated sample, observations may be collected in cycles with indices $J+s, J+2s, \cdots, J+Ts$ for some spacing s (Gelfand and Smith, 1990). However, in most practical applications a small s will be sufficient for many statistical analyses such as getting estimates of the parameters; see Albert and Chib (1993). In the numerical illustrations of the remaining chapters, we will use $s = 1$.

Figure 3.1 here

Statistical inference of the model can then be conducted on the basis of a simulated sample of observations from $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y})$, namely, $\{(\boldsymbol{\theta}^{(t)}, \boldsymbol{\Omega}^{(t)}) : t = 1, \cdots, T^*\}$. The Bayesian estimate of $\boldsymbol{\theta}$ as well as the numerical standard error estimate can be obtained from

$$\hat{\boldsymbol{\theta}} = T^{*-1} \sum_{t=1}^{T^*} \boldsymbol{\theta}^{(t)}, \tag{3.11}$$

$$\widehat{\mathrm{Var}}(\boldsymbol{\theta}|\mathbf{Y}) = (T^* - 1)^{-1} \sum_{t=1}^{T^*} (\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})^T. \tag{3.12}$$

It has been shown (Geyer, 1992) that $\hat{\boldsymbol{\theta}}$ tends to $E(\boldsymbol{\theta}|\mathbf{Y})$ as T^* tends to infinity. Other statistical inference on $\boldsymbol{\theta}$ can be carried out based on the simulated sample, $\{\boldsymbol{\theta}^{(t)} : t =$

$1, \ldots, T^*\}$. For instance, the 2.5% and 97.5% quantiles of the sampled distribution of an individual parameter can give a 95% posterior credible interval and convey skewness in its marginal posterior density. The total number of draws, T^*, that is required for accurate statistical analysis depends on the complexity of the posterior distribution. For most simple SEMs, 3,000 draws after convergence are sufficient. Different choices of sufficiently large T^* would produce close estimates, although they may not be exactly equal.

As the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Y}$ describes the distributional behaviors of $\boldsymbol{\theta}$ with the given data, the dispersion of $\boldsymbol{\theta}$ can be assessed through $\mathrm{Var}(\boldsymbol{\theta}|\mathbf{Y})$, with an estimate given by (3.12), based on the sample covariance matrix of the simulated observations. Let θ_k be the kth element of $\boldsymbol{\theta}$. The positive square root of the kth diagonal element in $\widehat{\mathrm{Var}}(\boldsymbol{\theta}|\mathbf{Y})$ can be taken as the estimate of the standard deviation of θ_k. Although this estimate is commonly taken as the standard error estimate and provides some information about the variation of $\hat{\theta}_k$, it may not be appropriate to construct a 'z-score' for hypothesis testing. In general Bayesian analysis, the issue of hypothesis testing is formulated as a model comparison problem, and is handled by some model comparison statistics such as the Bayes factor. See Chapter 4 for a more detailed discussion.

For any individual $\mathbf{y}_i$, let $\boldsymbol{\omega}_i$ be the vector of latent variables, and $E(\boldsymbol{\omega}_i|\mathbf{y}_i)$ be the posterior mean. A Bayesian estimate $\hat{\boldsymbol{\omega}}_i$ can be obtained through $\{\boldsymbol{\Omega}^{(t)},\ t = 1, \cdots, T^*\}$ as follows:

$$\hat{\boldsymbol{\omega}}_i = T^{*-1} \sum_{t=1}^{T^*} \boldsymbol{\omega}_i^{(t)}, \tag{3.13}$$

where $\boldsymbol{\omega}_i^{(t)}$ is the ith column of $\boldsymbol{\Omega}^{(t)}$. This gives a direct Bayesian estimate that is not expressed in terms of the structural parameter estimates. Hence, in contrast to the classical methods in estimating latent variables, no sampling errors of the estimates are involved in the Bayesian method. It can be shown (Geyer, 1992) that $\hat{\boldsymbol{\omega}}_i$ is a consistent estimate of $E(\boldsymbol{\omega}_i|\mathbf{y}_i)$. These estimates $\hat{\boldsymbol{\omega}}_i$ can be used for outlier and residual analyses, and the assessment of goodness-of-fit of the measurement equation or the structural equation, particularly in the analysis of complicated SEMs. See examples in Lee (2007) or other chapters of this book. It should be noted that as the data information for estimating $\hat{\boldsymbol{\omega}}_i$ is only given by the single observation $\mathbf{y}_i$, $\hat{\boldsymbol{\omega}}_i$ is not an accurate estimate of the true latent variable $\boldsymbol{\omega}_{i0}$, see the simulation study reported in Lee and Shi (2000) on the estimation of factor scores in a factor analysis model. However, the empirical distribution of the Bayesian estimates $\{\hat{\boldsymbol{\omega}}_1, \cdots, \hat{\boldsymbol{\omega}}_n\}$ is close to the distribution of the true factor scores $\{\hat{\boldsymbol{\omega}}_{10}, \cdots, \hat{\boldsymbol{\omega}}_{n0}\}$; see Shi and Lee (1998).

3.4 An Application of MCMC Methods to Some SEMs

In this section, we illustrate the implementation of MCMC methods through their applications to some SEMs described in Chapter 2. First, we consider the following linear SEM with fixed covariates. Its measurement equation for a $p \times 1$ observed random vector $\mathbf{y}_i$ measured on an individual i is given by:

$$\mathbf{y}_i = \mathbf{A}\mathbf{c}_i + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \cdots, n, \tag{3.14}$$

in which $\mathbf{A}$ and $\boldsymbol{\Lambda}$ are unknown parameter matrices, $\mathbf{c}_i$ is an $r_1 \times 1$ vector of fixed covariates, $\boldsymbol{\omega}_i$ is a $q \times 1$ latent random vector, and $\boldsymbol{\epsilon}_i$ is a random vector of residual errors with distribution $N[\mathbf{0}, \boldsymbol{\Psi}_\epsilon]$, where $\boldsymbol{\Psi}_\epsilon$ is diagonal and $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\omega}_i$. The structural equation is defined as

$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \tag{3.15}$$

where $\mathbf{d}_i$ is an $r_2 \times 1$ vector of fixed covariates, $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$, $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ are $q_1 \times 1$ and $q_2 \times 1$ latent vectors, respectively; $\mathbf{B}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Gamma}$ are unknown parameter matrices, $\boldsymbol{\xi}_i$

and $\boldsymbol{\delta}_i$ are independently distributed as $N[\mathbf{0}, \boldsymbol{\Phi}]$ and $N[\mathbf{0}, \boldsymbol{\Psi}_\delta]$, respectively, where $\boldsymbol{\Psi}_\delta$ is a diagonal covariance matrix. To simplify notation, Equation (3.15) is rewritten as

$$\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_\omega \mathbf{v}_i + \boldsymbol{\delta}_i, \tag{3.16}$$

where $\boldsymbol{\Lambda}_\omega = (\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma})$ and $\mathbf{v}_i = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$. See Section 2.3.1 for more detailed discussions of this model, such as the required assumptions and identification conditions.

Let $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n), \mathbf{C} = (\mathbf{c}_1, \cdots, \mathbf{c}_n)$, and $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_n)$ be the data matrices; and let $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n)$ be the matrix of latent vectors, and $\boldsymbol{\theta}$ be the structural parameter vector that contains all the unknown parameters in $\{\mathbf{A}, \boldsymbol{\Lambda}, \mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Psi}_\delta\} = \{\mathbf{A}, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_\omega, \boldsymbol{\Phi}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Psi}_\delta\}$. Our main objective is to use MCMC methods to obtain the Bayesian estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$. To achieve our goal, a sequence of random observations from the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$ will be generated via the Gibbs sampler which is implemented as follows: At the jth iteration with current value $\boldsymbol{\theta}^{(j)}$;

Step a: Generate a random variate $\boldsymbol{\Omega}^{(j+1)}$ from the conditional distribution $[\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}^{(j)}]$.

Step b: Generate a random variate $\boldsymbol{\theta}^{(j+1)}$ from the conditional distribution $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}^{(j+1)}]$, and return to 'Step a' if necessary.

Here $\boldsymbol{\theta}$ has six components that correspond to unknown parameters in $\mathbf{A}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}_\omega$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}_\epsilon$, and $\boldsymbol{\Psi}_\delta$, while $\boldsymbol{\Omega}$ has only one component. Conjugate prior distributions for parameters in various components of $\boldsymbol{\theta}$ can be similarly obtained as before; see Equations (3.6) and (3.8). For readers who are interested in developing their own computer programs, full conditional distributions for implementing Step a and Step b of the Gibbs sampler are presented in Appendix 3.3. These full conditional distributions are the familiar normal, Gamma, and inverted Wishart distributions. Simulating observations from them is fast and straightforward. For applied researchers, we will discuss the use of the freely available software WinBUGS (Spiegelhalter *et al.*, 2003) for obtaining the Bayesian results in Section 3.5.

As we discussed at the beginning of Section 3.3, the main difference between SEMs and the familiar regression model is the presence of latent variables in SEMs. Since latent variables are random rather than observed, classical techniques in regression cannot be applied in estimating parameters in SEMs. The idea of data augmentation is used to solve the problem. We augment $\mathbf{\Omega}$, the matrix containing all latent variables, with the observed data $\mathbf{Y}$ and work on the joint posterior distribution $[\boldsymbol{\theta}, \mathbf{\Omega}|\mathbf{Y}]$. In Step b of the Gibbs sampler, we need to simulate $\boldsymbol{\theta}$ from $[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{\Omega}]$, the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{Y}$ and $\mathbf{\Omega}$. It is important to note that once $\mathbf{\Omega}$ is given rather than random, the SEM becomes the familiar regression model. Consequently, the conditional distribution $[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{\Omega}]$ can be derived and the implementation of Gibbs sampler is possible.

The above strategy based on data augmentation is very useful for developing Bayesian methods in the analysis of various complex SEMs with complicated data structures; see detailed discussions in subsequent chapters. Here, we present an application of this strategy to the analysis of nonlinear SEMs for illustration.

Consider a generalization of linear SEMs with fixed covariates to nonlinear SEMs with fixed covariates by extending the structural equation (3.15) to a nonlinear structural equation as follows:

$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \mathbf{\Pi}\boldsymbol{\eta}_i + \mathbf{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{3.17}$$

where $\mathbf{F}(\boldsymbol{\xi}_i)$ is a vector-valued nonlinear function of $\boldsymbol{\xi}_i$, and the definitions of other random vectors and parameter matrices are the same as before. The distributions of the nonlinear terms of $\boldsymbol{\xi}_i$ in $\mathbf{F}(\boldsymbol{\xi}_i)$ are not normal and hence induce serious difficulties in applying the traditional method such as the covariance structure analysis approach, and the existing commercial software in SEMs. In contrast, the nonlinear terms of $\boldsymbol{\xi}_i$ can be easily handled using the Bayesian approach with data augmentation. First note that the Gibbs sampler is similarly implemented with Step a and Step b as before, although $[\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y}]$ and $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$ are slightly different. The nonlinear terms of $\boldsymbol{\xi}_i$ induce no difficulties in deriving these conditional distributions. In fact, $[\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y}]$ can be derived on the basis of the distribution of the latent variables and the definition of the model. For $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$, as $\boldsymbol{\Omega}$ is given, the nonlinear SEM again becomes the familiar regression model. The conditional distributions for the implementation of the MCMC methods in nonlinear SEMs are presented in Appendix 3.4. We note that the differences between the conditional distributions corresponding to linear and nonlinear SEMs are minor. Hence, we regard nonlinear SEMs as basic SEMs.

3.5 Bayesian Estimation via WinBUGS

The freely available software WinBUGS (**W**indows version of **B**ayesian inference **U**sing **G**ibbs **S**ampling) is useful for producing reliable Bayesian statistics for a wide range of statistical models. WinBUGS is developed using MCMC techniques, such as the Gibbs sampler (Geman and Geman, 1984) and the MH algorithm (Metropolis *et al.*, 1953; Hastings, 1970). It has been shown that under broad conditions, this software can provide simulated samples from the joint posterior distribution of the unknown quantities, such as parameters and latent variables in the model. As discussed in previous sections, Bayesian estimates of the unknown parameters and latent variables in the model can be obtained from these samples for conducting statistical inferences.

The advanced version of the program is WinBUGS 1.4 running under Windows, which is developed by the Medical Research Council (MRC) Biostatistics Unit (Cambridge, UK) and the Department of Epidemiology and Public Health of the Imperial College School of Medicine at St. Mary's Hospital (London). It can be downloaded from the website http://www.mrc-bsu.cam.ac.uk/bugs/. The WinBUGS manual (Spiegelhalter *et al.*, 2003) is available online, which gives brief instructions on WinBUGS. See also Lawson, Browne and Vidal Rodeiro (2003, Chapter 4) for supplementary descriptions.

We illustrate the use of WinBUGS through the analysis of an artificial example that is based on the following nonlinear SEM with a linear covariate (see Lee, Song and Tang, 2007; Lee, 2007). For easy application of the program, we use the following scalar representation of the model. Let $y_{ij} \overset{D}{=} N[\mu^*_{ij}, \psi_j]$, where

$$\begin{aligned} &\mu^*_{i1} = \mu_1 + \eta_i, \quad \mu^*_{ij} = \mu_j + \lambda_{j1}\eta_i, \quad j = 2,\ 3, \\ &\mu^*_{i4} = \mu_4 + \xi_{i1},\ \mu^*_{ij} = \mu_j + \lambda_{j2}\xi_{i1},\ j = 5,\ 6,\ 7,\ \text{and} \\ &\mu^*_{i8} = \mu_8 + \xi_{i2},\ \mu^*_{ij} = \mu_j + \lambda_{j3}\xi_{i2},\ j = 9,\ 10, \end{aligned} \tag{3.18}$$

where μ_j's are intercepts, the η's and ξ's are the latent variables. The structural equation is reformulated by defining the conditional distribution of η_i given ξ_{i1} and ξ_{i2} as $N[\nu_i, \psi_\delta]$, where

$$\nu_i = b_1 d_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_3\xi_{i1}\xi_{i2} + \gamma_4\xi^2_{i1} + \gamma_5\xi^2_{i2}. \tag{3.19}$$

in which d_i is a fixed covariate coming from a t distribution with 5 degrees of freedom. The true population values of the unknown parameters in the model were taken to be:

$$\begin{aligned} &\mu_1 = \cdots = \mu_{10} = 0.0,\ \lambda_{21} = \lambda_{52} = \lambda_{93} = 0.9,\ \lambda_{31} = \lambda_{62} = \lambda_{10,3} = 0.7, \\ &\lambda_{72} = 0.5,\ \psi_{\epsilon 1} = \psi_{\epsilon 2} = \psi_{\epsilon 3} = 0.3,\ \psi_{\epsilon 4} = \cdots = \psi_{\epsilon 7} = 0.5,\ \psi_{\epsilon 8} = \psi_{\epsilon 9} = \psi_{\epsilon 10} = 0.4, \\ &b_1 = 0.5,\ \gamma_1 = \gamma_2 = 0.4,\ \gamma_3 = 0.3,\ \gamma_4 = 0.2,\ \gamma_5 = 0.5,\ \text{and} \\ &\phi_{11} = \phi_{22} = 1.0,\ \phi_{12} = 0.3,\ \psi_\delta = 0.36. \end{aligned} \tag{3.20}$$

Based on the model formulation and these true parameter values, a random sample of continuous observations $\{\mathbf{y}_i,\ i = 1, \cdots, 500\}$ was generated, which gave the observed data set $\mathbf{Y}$. The following hyperparameter values were taken for the conjugate prior distributions in Equations (3.6) and (3.8):

$$\begin{aligned}
&\boldsymbol{\mu}_0 = (0.0, \cdots, 0.0)^T,\ \boldsymbol{\Sigma}_0 = \mathbf{I}_{10},\ \alpha_{0\epsilon k} = \alpha_{0\delta} = 9,\ \beta_{0\epsilon k} = \beta_{0\delta} = 4, \\
&\text{elements in } \boldsymbol{\Lambda}_{0k} \text{ and } \boldsymbol{\Lambda}_{0\omega k} \text{ are taken to be the true values,} \qquad (3.21) \\
&\mathbf{H}_{0yk} = \mathbf{I}_{10},\ \mathbf{H}_{0\omega k} = \mathbf{I}_6,\ \rho_0 = 4,\ \mathbf{R}_0 = \boldsymbol{\Phi}_0^{-1},
\end{aligned}$$

where $\boldsymbol{\Phi}_0$ is the matrix with true values of ϕ_{11}, ϕ_{22}, and ϕ_{12}. These hyperparameter values represent accurate prior inputs. The WinBUGS code and data are respectively given in the following websites:

http://www.sta.cuhk.edu.hk/song-lee/book-chapter3(section3.5)/WinBUGS-code,
http://www.sta.cuhk.edu.hk/song-lee/book-chapter3(section3.5)/WinBUGS-data.
(PLEASE CHANGE TO A WEB-SITE HOUSED IN JOHN-WILEY).

We observed that the WinBUGS program converged in less than 4,000 iterations. Plots of some simulated sequences of observations for monitoring convergence are presented in Figure 3.2. Based on Equations (3.11) and (3.12), Bayesian estimates of the parameters and their standard error estimates as obtained from 6,000 iterations after the 4,000 burn-in iterations are presented in Table 3.1. We observe that the Bayesian estimates (EST) are close to the true values, and that the standard error estimates (SE) are reasonable. WinBUGS also produces estimates of the latent variables $\{\hat{\boldsymbol{\omega}}_i = (\hat{\eta}_i, \hat{\xi}_{i1}, \hat{\xi}_{i2})^T,\ i = 1, \cdots, n\}$. Histograms that correspond to the sets of latent variable estimates $\hat{\xi}_{i1}$ and $\hat{\xi}_{i2}$ are displayed in Figure 3.3. We observe from these histograms that the corresponding empirical distributions are close to the normal distributions. The elements in the sample covariance matrix of $\{\hat{\boldsymbol{\xi}}_i,\ i = 1,\ \cdots, n\}$ are $s_{11} = 0.902$, $s_{12} = 0.311$, and $s_{22} = 0.910$, and hence this sample covariance matrix is close to the true covariance matrix of $\boldsymbol{\xi}_i$; see (3.20). The residuals can be estimated via

$\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\omega}}_i = (\hat{\eta}_i, \hat{\xi}_{i1}, \hat{\xi}_{i2})^T$ for $i = 1, \cdots, n$ as follows:

$$\begin{aligned}
&\hat{\epsilon}_{i1} = y_{i1} - \hat{\mu}_1 - \hat{\eta}_i, \quad \hat{\epsilon}_{ij} = y_{ij} - \hat{\mu}_j - \hat{\lambda}_{j1}\hat{\eta}_i, \quad j = 2,\ 3,\\
&\hat{\epsilon}_{i4} = y_{i4} - \hat{\mu}_4 - \hat{\xi}_{i1}, \quad \hat{\epsilon}_{ij} = y_{ij} - \hat{\mu}_j - \hat{\lambda}_{j2}\hat{\xi}_{i1}, \quad j = 5,\ 6,\ 7,\\
&\hat{\epsilon}_{i8} = y_{i8} - \hat{\mu}_8 - \hat{\xi}_{i2}, \quad \hat{\epsilon}_{ij} = y_{ij} - \hat{\mu}_j - \hat{\lambda}_{j3}\hat{\xi}_{i2}, \quad j = 9,\ 10,\\
&\hat{\delta}_i = \hat{\eta}_i - \hat{b}_1 d_i - \hat{\gamma}_1\hat{\xi}_{i1} - \hat{\gamma}_2\hat{\xi}_{i2} - \hat{\gamma}_3\hat{\xi}_{i1}\hat{\xi}_{i2} - \hat{\gamma}_4\hat{\xi}_{i1}^2 - \hat{\gamma}_5\hat{\xi}_{i2}^2.
\end{aligned}$$

Some estimated residual plots, $\hat{\epsilon}_{i2}$, $\hat{\epsilon}_{i3}$, $\hat{\epsilon}_{i8}$, and $\hat{\delta}_i$, against the case number are presented in Figure 3.4. The plots of estimated residuals $\hat{\delta}_i$ versus $\hat{\xi}_{i1}$ and $\hat{\xi}_{i2}$; and $\hat{\epsilon}_{i2}$ versus $\hat{\xi}_{i1}$, $\hat{\xi}_{i2}$, and $\hat{\eta}_i$ are presented in Figures 3.5 and 3.6, respectively. Other residual plots are similar. The interpretation of these residual plots is similar to that in regression models. We observe that the plots lie within two parallel horizontal lines that are centered at zero, and no linear or quadratic trends are detected. This roughly indicates that the proposed measurement equation and structural equation are adequate. Moreover, based on $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Omega}}$, we can compute the estimate of the proportion of the variance of $\mathbf{y}$ that can be explained by the measurement equation, using exactly the same method as in analyzing a regression model. Similarly, the proportion of the variance of $\boldsymbol{\eta}$ that can be explained by the structural equation can also be estimated.

Figures 3.2 to 3.6 here

WinBUGS is rather flexible in the analysis of SEMs. In this example, it is applied to analyze nonlinear SEMs with covariates. In the program set up (see the above mentioned website), it only requires a single program statement for the structural equation given by (3.19). In fact, even with more complicated quadratic or interaction terms of the explanatory latent variables and fixed covariates, one program statement is sufficient.

Hence, nonlinear SEMs with covariates can be easily analyzed via WinBUGS. This is the reason for us to regard nonlinear SEMs as basic SEMs.

WinBUGS is an interactive program, and it is not convenient to directly use it to do a simulation study. However, WinBUGS can be run in batch mode using scripts, and the R package R2WinBUGS (Sturtz, Ligges and Gelman, 2005) uses this feature and provides tools to directly call WinBUGS after the manipulation in R. Furthermore, it is possible to work on the results after importing them back into R. The implementation of R2WinBUGS is mainly based on the R function 'bugs($\cdots$)', which takes data and initial values as input. It automatically writes a WinBUGS script, calls the model, and saves the simulation for easy access in R.

To illustrate the applications of WinBUGS together with R2WinBUGS, we present a simulation study based on the settings of the artificial example described above, see Equations (3.18), (3.19), and (3.20). The sample size was again taken to be 500, and the conjugate prior distributions with hyperparameter values as given in (3.21) were used. Based on 100 replications, the simulation results reported in Table 3.2 were obtained. In the simulation study, we first use R to generate the data sets, input these data sets in WinBUGS to obtain the Bayesian estimates from the WinBUGS outputs. The Bayesian estimates and the associated results are then stored and analyzed by R. The WinBUGS and the R codes for the simulation study are presented in Appendices 3.5 and 3.6, respectively.

Appendix 3.1: The Gamma, Inverted Gamma, Wishart, and Inverted Wishart Distributions and Their Characteristics

Let θ and $\mathbf{W}$ denote unknown parameter and unknown covariance matrix, respectively; and let $p(\cdot)$, $E(\cdot)$, and $Var(\cdot)$ denote the density function, the expectation, and the variance, respectively.

1. Gamma distribution: $\theta \overset{D}{=} Gamma[\alpha, \beta]$

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta},$$
$$E(\theta) = \alpha/\beta, \quad Var(\theta) = \alpha/\beta^2.$$

2. Inverted Gamma distribution: $\theta \overset{D}{=} Inverted\ Gamma[\alpha, \beta]$

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta},$$
$$E(\theta) = \frac{\beta}{\alpha-1}, \quad Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-1)}.$$

3. Relation between Gamma and inverted Gamma distributions

$$\text{If} \quad \theta \overset{D}{=} Inverted\ Gamma[\alpha, \beta], \quad \text{then} \quad \theta^{-1} \overset{D}{=} Gamma[\alpha, \beta].$$

4. Wishart distribution: $\mathbf{W} \overset{D}{=} Wishart_q[\mathbf{R}_0, \rho_0]$

$$p(\mathbf{W}) = \left[2^{\rho_0 q/2} \pi^{q(q-1)/4} \prod_{i=1}^{q} \Gamma\left(\frac{\rho_0 + 1 - i}{2}\right) \right]^{-1}$$
$$\times |\mathbf{R}_0|^{-\rho_0/2} \times |\mathbf{W}|^{(\rho_0-q-1)/2} \times \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{R}_0^{-1}\mathbf{W}) \right\},$$
$$E(\mathbf{W}) = \rho_0 \mathbf{R}_0.$$

5. Inverted Wishart distribution: $\mathbf{W} \overset{D}{=} IW_q[\mathbf{R}_0^{-1}, \rho_0]$

$$p(\mathbf{W}) = \left[2^{\rho_0 q/2} \pi^{q(q-1)/4} \prod_{i=1}^{q} \Gamma\left(\frac{\rho_0 + 1 - i}{2}\right) \right]^{-1}$$

$$\times|\mathbf{R}_0|^{-\rho_0/2} \times |\mathbf{W}|^{-(\rho_0+q+1)/2} \times \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{R}_0^{-1}\mathbf{W}^{-1})\right\},$$

$$E(\mathbf{W}) = \frac{\mathbf{R}_0^{-1}}{\rho_0 - q - 1}.$$

6. Relation between Wishart and inverted Wishart distributions

$$\text{If} \;\; \mathbf{W} \overset{D}{=} IW[\mathbf{R}_0^{-1}, \rho_0], \quad \text{then} \;\; \mathbf{W}^{-1} \overset{D}{=} W[\mathbf{R}_0, \rho_0].$$

Appendix 3.2: The Metropolis-Hastings Algorithm

Suppose we wish to simulate observations say $\{X_j, \; j = 1, 2, \cdots\}$ from a conditional distribution with target density $p(\cdot)$. At the jth iteration of the Metropolis-Hastings (MH) algorithm with a current value X_j, the next X_{j+1} is chosen by first sampling a candidate point Y from a proposal distribution $q(\cdot|X_j)$ which is easy to sample. This candidate point Y is accepted as X_{j+1} with probability

$$\min\left(1, \frac{p(Y)q(X_j|Y)}{p(X_j)q(Y|X_j)}\right).$$

If the candidate point Y is rejected, then $X_{j+1} = X_j$ and the chain does not move.

The proposal distribution $q(\cdot|\cdot)$ can have any form and the stationary distribution of the Markov chain will be the target distribution with density $p(\cdot)$. In most analyses of SEMs considered in this book, we will take $q(\cdot|X)$ to be a normal distribution with mean X and a fixed covariance matrix.

Appendix 3.3: Conditional Distributions $[\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}]$ and $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$

Conditional Distribution $[\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}]$

We first note that for $i = 1, \cdots, n$, $\boldsymbol{\omega}_i$ are conditionally independent given $\boldsymbol{\theta}$, and $\mathbf{y}_i$ are also conditionally independent given $(\boldsymbol{\omega}_i, \boldsymbol{\theta})$. Hence,

$$p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\boldsymbol{\theta}) \; p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}). \tag{3.A1}$$

It implies that the conditional distributions of $\boldsymbol{\omega}_i$ given $(\mathbf{y}_i, \boldsymbol{\theta})$ are mutually independent for different i, and $p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}) \propto p(\boldsymbol{\omega}_i|\boldsymbol{\theta})p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta})$. Let $\mathbf{\Pi}_0 = \mathbf{I} - \mathbf{\Pi}$ and the covariance matrix of $\boldsymbol{\omega}_i$ be

$$\mathbf{\Sigma}_\omega = \begin{bmatrix} \mathbf{\Pi}_0^{-1}(\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \mathbf{\Psi}_\delta)\mathbf{\Pi}_0^{-T} & \mathbf{\Pi}_0^{-1}\mathbf{\Gamma}\mathbf{\Phi} \\ \mathbf{\Phi}\mathbf{\Gamma}^T\mathbf{\Pi}_0^{-T} & \mathbf{\Phi} \end{bmatrix}.$$

It can be shown that

$$[\boldsymbol{\omega}_i|\boldsymbol{\theta}] \overset{D}{=} N\left[\begin{pmatrix} \mathbf{\Pi}_0^{-1}\mathbf{B}\mathbf{d}_i \\ \mathbf{0} \end{pmatrix}, \mathbf{\Sigma}_\omega\right],$$

and$[\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}] \overset{D}{=} N[\mathbf{A}\mathbf{c}_i + \mathbf{\Lambda}\boldsymbol{\omega}_i, \mathbf{\Psi}_\epsilon]$. Thus,

$$[\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}] \overset{D}{=} N\left[\mathbf{\Sigma}^{*-1}\mathbf{\Lambda}^T\mathbf{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \mathbf{A}\mathbf{c}_i) + \mathbf{\Sigma}^{*^{-1}}\mathbf{\Sigma}_\omega^{-1}\begin{pmatrix} \mathbf{\Pi}_0^{-1}\mathbf{B}\mathbf{d}_i \\ \mathbf{0} \end{pmatrix}, \mathbf{\Sigma}^{*-1}\right] \qquad (3.A2)$$

where $\mathbf{\Sigma}^* = \mathbf{\Sigma}_\omega^{-1} + \mathbf{\Lambda}^T\mathbf{\Psi}_\epsilon^{-1}\mathbf{\Lambda}$. We see that the conditional distribution $[\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}]$ is a normal distribution.

Conditional Distribution $[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{\Omega}]$

The conditional distribution of $\boldsymbol{\theta}$ given $(\mathbf{Y}, \mathbf{\Omega})$ is proportional to $p(\boldsymbol{\theta})p(\mathbf{Y}, \mathbf{\Omega}|\boldsymbol{\theta})$. We note that as $\mathbf{\Omega}$ is given, the equations defined in (3.14) and (3.15) are linear models with fixed covariates. Let $\boldsymbol{\theta}_y$ be the unknown parameters in $\mathbf{A}$, $\mathbf{\Lambda}$, and $\mathbf{\Psi}_\epsilon$ associated with the measurement equation, and $\boldsymbol{\theta}_\omega$ be the unknown parameters in $\mathbf{B}$, $\mathbf{\Pi}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}_\delta$ associated with the structural equation. It is assumed that the prior distribution of $\boldsymbol{\theta}_y$ is independent of the prior distribution of $\boldsymbol{\theta}_\omega$, that is, $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)$. Moreover, as $p(\mathbf{Y}|\mathbf{\Omega}, \boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{\Omega}, \boldsymbol{\theta}_y)$ and $p(\mathbf{\Omega}|\boldsymbol{\theta}) = p(\mathbf{\Omega}|\boldsymbol{\theta}_\omega)$, it can be shown that the marginal conditional densities of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\omega$ given $(\mathbf{Y}, \mathbf{\Omega})$ are proportional to $p(\mathbf{Y}|\mathbf{\Omega}, \boldsymbol{\theta}_y)p(\boldsymbol{\theta}_y)$ and $p(\mathbf{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega)$, respectively. Hence, these conditional densities can be treated separately.

Consider first the marginal conditional distribution of $\boldsymbol{\theta}_y$. Let $\mathbf{\Lambda}_y = (\mathbf{A}, \mathbf{\Lambda})$ with general elements λ_{ykj}, $j = 1, \cdots, r_1 + q$, $k = 1, \cdots, p$, and $\mathbf{u}_i = (\mathbf{c}_i^T, \boldsymbol{\omega}_i^T)^T$. It follows that $\mathbf{y}_i = \boldsymbol{\Lambda}_y \mathbf{u}_i + \boldsymbol{\epsilon}_i$. This simple transformation reformulates the model with fixed covariate $\mathbf{c}_i$ to the original factor analysis model. The positions of the fixed elements in $\boldsymbol{\Lambda}_y$ are identified via an index matrix $\mathbf{L}_y$ with the following elements:

$$l_{ykj} = \begin{cases} 0, & \text{if } \lambda_{ykj} \text{ is fixed,} \\ 1, & \text{if } \lambda_{ykj} \text{ is free;} \end{cases} \qquad \text{for} \quad j = 1, \cdots, r_1 + q \quad \text{and} \quad k = 1, \cdots, p.$$

Let $\psi_{\epsilon k}$ be the kth diagonal element of $\boldsymbol{\Psi}_\epsilon$, and $\boldsymbol{\Lambda}_{yk}^T$ be the row vector that contains the unknown parameters in the kth row of $\boldsymbol{\Lambda}_y$. The following commonly used conjugate type prior distributions are used. For any $k \neq h$, we assume that the prior distribution of $(\psi_{\epsilon k}, \boldsymbol{\Lambda}_{yk})$ is independent of $(\psi_{\epsilon h}, \boldsymbol{\Lambda}_{yh})$, and

$$\psi_{\epsilon k}^{-1} \overset{D}{=} Gamma[\alpha_{0\epsilon k},\ \beta_{0\epsilon k}], \quad \text{and} \quad [\boldsymbol{\Lambda}_{yk} | \psi_{\epsilon k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0yk}, \psi_{\epsilon k} \mathbf{H}_{0yk}], \ k = 1, \cdots, p, \quad (3.A3)$$

where $\alpha_{0\epsilon k}, \beta_{0\epsilon k}, \boldsymbol{\Lambda}_{0yk}^T = (\mathbf{A}_{0k}^T, \boldsymbol{\Lambda}_{0k}^T)$, and the positive definite matrix $\mathbf{H}_{0yk}$ are hyperparameters whose values are assumed to be given from the prior information of previous studies or other sources.

Let $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$ and $\mathbf{U}_k$ be the submatrix of $\mathbf{U}$ such that all the rows corresponding to $l_{ykj} = 0$ are deleted; and let $\mathbf{Y}_k^{*T} = (y_{1k}^*, \cdots, y_{nk}^*)$ with

$$y_{ik}^* = y_{ik} - \sum_{j=1}^{r_1+q} \lambda_{ykj} u_{ij} (1 - l_{ykj}).$$

where u_{ij} is the jth element of $\mathbf{u}_i$. Then, for $k = 1, \cdots, p$, it can be shown (see Lee, 2007, Chapter 4, Appendix 4.3) that

$$[\psi_{\epsilon k}^{-1} | \mathbf{Y}, \boldsymbol{\Omega}] \overset{D}{=} Gamma[n/2 + \alpha_{0\epsilon k}, \beta_{\epsilon k}], \ [\boldsymbol{\Lambda}_{yk} | \mathbf{Y}, \boldsymbol{\Omega}, \psi_{\epsilon k}^{-1}] \overset{D}{=} N[\mathbf{a}_{yk}, \psi_{\epsilon k} \mathbf{A}_{yk}], \quad (3.A4)$$

where $\mathbf{A}_{yk} = (\mathbf{H}_{0yk}^{-1} + \mathbf{U}_k \mathbf{U}_k^T)^{-1}$, $\mathbf{a}_{yk} = \mathbf{A}_{yk}(\mathbf{H}_{0yk}^{-1} \boldsymbol{\Lambda}_{0yk} + \mathbf{U}_k \mathbf{Y}_k^*)$, and

$$\beta_{\epsilon k} = \beta_{0\epsilon k} + \frac{1}{2}(\mathbf{Y}_k^{*T} \mathbf{Y}_k^* - \mathbf{a}_{yk}^T \mathbf{A}_{yk}^{-1} \mathbf{a}_{yk} + \boldsymbol{\Lambda}_{0yk}^T \mathbf{H}_{0yk}^{-1} \boldsymbol{\Lambda}_{0yk}).$$

Since $[\mathbf{\Lambda}_{yk}, \psi_{\epsilon k}^{-1}|\mathbf{Y}, \mathbf{\Omega}]$ equals to $[\psi_{\epsilon k}^{-1}|\mathbf{Y}, \mathbf{\Omega}][\mathbf{\Lambda}_{yk}|\mathbf{Y}, \mathbf{\Omega}, \psi_{\epsilon k}^{-1}]$, it can be obtained via (3.A4). This gives the conditional distribution in relation to $\boldsymbol{\theta}_y$.

Now, consider the conditional distribution of $\boldsymbol{\theta}_\omega$ that is proportional to $p(\mathbf{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega)$. Let $\mathbf{\Omega}_1 = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)$ and $\mathbf{\Omega}_2 = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)$. Since the distribution of $\boldsymbol{\xi}_i$ only involves $\mathbf{\Phi}$, $p(\mathbf{\Omega}_2|\boldsymbol{\theta}_\omega) = p(\mathbf{\Omega}_2|\mathbf{\Phi})$. Under the assumption that the prior distribution of $\mathbf{\Phi}$ is independent of the prior distributions of $\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma}$, and $\mathbf{\Psi}_\delta$, we have

$$p(\mathbf{\Omega}|\boldsymbol{\theta}_\omega)p(\boldsymbol{\theta}_\omega) = [p(\mathbf{\Omega}_1|\mathbf{\Omega}_2, \mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma}, \mathbf{\Psi}_\delta)p(\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma}, \mathbf{\Psi}_\delta)][p(\mathbf{\Omega}_2|\mathbf{\Phi})p(\mathbf{\Phi})].$$

Hence, the marginal conditional densities of $(\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma}, \mathbf{\Psi}_\delta)$ and $\mathbf{\Phi}$ can be treated separately.

Consider a conjugate type prior distribution for $\mathbf{\Phi}$ with $\mathbf{\Phi} \overset{D}{=} IW_{q_2}[\mathbf{R}_0^{-1}, \rho_0]$ or $\mathbf{\Phi}^{-1} \overset{D}{=} W_{q_2}[\mathbf{R}_0, \rho_0]$, with hyperparameters ρ_0 and $\mathbf{R}_0^{-1}$ or $\mathbf{R}_0$.

To derive $p(\mathbf{\Phi}|\mathbf{\Omega}_2)$, we first note that it is proportional to $p(\mathbf{\Phi})p(\mathbf{\Omega}_2|\mathbf{\Phi})$. As $\boldsymbol{\xi}_i$ are independent, we have

$$p(\mathbf{\Phi}|\mathbf{\Omega}_2) \propto p(\mathbf{\Phi}) \prod_{i=1}^n p(\boldsymbol{\xi}_i|\boldsymbol{\theta}).$$

Moreover, since the distribution of $\boldsymbol{\xi}_i$ given $\mathbf{\Phi}$ is $N(\mathbf{0}, \mathbf{\Phi})$, we have

$$\begin{aligned} p(\mathbf{\Phi}|\mathbf{\Omega}_2) &\propto \left[|\mathbf{\Phi}|^{-(\rho_0+q_2+1)/2} \exp\Big\{-\frac{1}{2}\mathrm{tr}[\mathbf{R}_0^{-1}\mathbf{\Phi}^{-1}]\Big\}\right]\left[|\mathbf{\Phi}|^{-n/2} \exp\Big\{-\frac{1}{2}\sum_{i=1}^n \boldsymbol{\xi}_i^T\mathbf{\Phi}^{-1}\boldsymbol{\xi}_i\Big\}\right] \\ &= |\mathbf{\Phi}|^{-(n+\rho_0+q_2+1)/2} \exp\Big\{-\frac{1}{2}\,\mathrm{tr}[\mathbf{\Phi}^{-1}(\mathbf{\Omega}_2\mathbf{\Omega}_2^T + \mathbf{R}_0^{-1})]\Big\}. \end{aligned} \tag{3.A5}$$

Since the right hand side of (3.A5) is proportional to the density function of an inverted Wishart distribution (Zellner, 1971), it follows that the conditional distribution of $\mathbf{\Phi}$ given $\mathbf{\Omega}_2$ is given by

$$[\mathbf{\Phi}|\mathbf{\Omega}_2] \overset{D}{=} IW_{q_2}[(\mathbf{\Omega}_2\mathbf{\Omega}_2^T + \mathbf{R}_0^{-1}), n + \rho_0] \tag{3.A6}$$

Recall that $\boldsymbol{\eta}_i = \mathbf{\Lambda}_\omega\mathbf{v}_i + \boldsymbol{\delta}_i$, where $\mathbf{\Lambda}_\omega = (\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma})$ with general elements $\lambda_{\omega kj}$ for $k = 1, \cdots, q_1$, and $\mathbf{v}_i = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T = (\mathbf{d}_i^T, \boldsymbol{\omega}_i^T)^T$ be an $(r_2 + q_1 + q_2) \times 1$ vector.

The model $\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_\omega \mathbf{v}_i + \boldsymbol{\delta}_i$ is similar to $\mathbf{y}_i = \boldsymbol{\Lambda}_y \mathbf{u}_i + \boldsymbol{\epsilon}_i$ considered before. Hence, the derivations for the conditional distributions corresponding to $\boldsymbol{\theta}_\omega$ are similar to those corresponding to $\boldsymbol{\theta}_y$. Let $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$, $\mathbf{L}_\omega$ be the index matrix with general elements $l_{\omega kj}$ that similarly defined as $\mathbf{L}_y$ to indicate the fixed known parameters in $\boldsymbol{\Lambda}_\omega$; $\psi_{\delta k}$ be the kth diagonal element of $\boldsymbol{\Psi}_\delta$ and $\boldsymbol{\Lambda}_{\omega k}^T$ be the row vector that contains the unknown parameters in the kth row of $\boldsymbol{\Lambda}_\omega$. The prior distributions of $\boldsymbol{\Lambda}_{\omega k}$ and $\psi_{\delta k}^{-1}$ are similarly selected as the following conjugate type distributions:

$$\psi_{\delta k}^{-1} \overset{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}], \quad \text{and} \quad [\boldsymbol{\Lambda}_{\omega k}|\psi_{\delta k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0\omega k}, \psi_{\delta k}\mathbf{H}_{0\omega k}], \quad k = 1, \cdots, q_1, \quad (3.A7)$$

where $\alpha_{0\delta k}$, $\beta_{0\delta k}$, $\boldsymbol{\Lambda}_{0\omega k}$, and $\mathbf{H}_{0\omega k}$ are given hyperparameters. Moreover, it is assumed that for $h \neq k$, $(\psi_{\delta k}, \boldsymbol{\Lambda}_{\omega k})$ and $(\psi_{\delta h}, \boldsymbol{\Lambda}_{\omega h})$ are independent. Let $\mathbf{V}_k$ be the submatrix of $\mathbf{V}$ such that all the rows corresponding to $l_{\omega kj} = 0$ are deleted; and let $\boldsymbol{\Xi}_k^T = (\eta_{1k}^*, \cdots, \eta_{nk}^*)$ where

$$\eta_{ik}^* = \eta_{ik} - \sum_{j=1}^{r_2+q} \lambda_{\omega kj} v_{ij}(1 - l_{\omega kj}).$$

Then, it can be shown that:

$$[\psi_{\delta k}^{-1}|\boldsymbol{\Omega}] \overset{D}{=} Gamma[n/2 + \alpha_{0\delta k}, \beta_{\delta k}], \quad \text{and} \quad [\boldsymbol{\Lambda}_{\omega k}|\boldsymbol{\Omega}, \psi_{\delta k}^{-1}] \overset{D}{=} N[\mathbf{a}_{\omega k}, \psi_{\delta k}\mathbf{A}_{\omega k}], \quad (3.A8)$$

where $\mathbf{A}_{\omega k} = (\mathbf{H}_{0\omega k}^{-1} + \mathbf{V}_k\mathbf{V}_k^T)^{-1}, \mathbf{a}_{\omega k} = \mathbf{A}_{\omega k}(\mathbf{H}_{0\omega k}^{-1}\boldsymbol{\Lambda}_{0\omega k} + \mathbf{V}_k\boldsymbol{\Xi}_k)$, and

$$\beta_{\delta k} = \beta_{0\delta k} + \frac{1}{2}(\boldsymbol{\Xi}_k^T\boldsymbol{\Xi}_k - \mathbf{a}_{\omega k}^T\mathbf{A}_{\omega k}^{-1}\mathbf{a}_{\omega k} + \boldsymbol{\Lambda}_{0\omega k}^T\mathbf{H}_{0\omega k}^{-1}\boldsymbol{\Lambda}_{0\omega k}).$$

The conditional distribution $[\boldsymbol{\theta}_\omega|\boldsymbol{\Omega}] = [\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_\delta|\boldsymbol{\Omega}]$ can be obtained through (3.A8).

In this appendix, we use $\boldsymbol{\Lambda}_{yk}^T$ and $\boldsymbol{\Lambda}_{\omega k}^T$ to denote the row vectors that contain the unknown parameters in the kth rows of $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_\omega$, respectively. However, in the subsequent chapters we sometimes assume that all the elements in the kth rows of $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_\omega$ are unknown parameters for simplicity. Under these assumptions, $\boldsymbol{\Lambda}_{yk}^T$ and $\boldsymbol{\Lambda}_{\omega k}^T$ simply denote the kth rows of $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_\omega$, respectively.

Appendix 3.4: Conditional Distributions $[\boldsymbol{\Omega}|\mathbf{Y},\boldsymbol{\theta}]$ and $[\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\Omega}]$ in Nonlinear SEMs with Covariates

Conditional Distribution $[\boldsymbol{\Omega}|\mathbf{Y},\boldsymbol{\theta}]$

First note that the measurement equation of a nonlinear SEM with covariates is the same as given in Equation (3.14), while the structural equation is defined as in Equation (3.17). Again to simplify notation, Equation (3.17) is rewritten as

$$\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_\omega \mathbf{G}(\boldsymbol{\omega}_i) + \boldsymbol{\delta}_i, \tag{3.A9}$$

where $\boldsymbol{\Lambda}_\omega = (\mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma})$ and $\mathbf{G}(\boldsymbol{\omega}_i) = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \mathbf{F}(\boldsymbol{\xi}_i)^T)^T$. Similar reasoning can be used to derive the conditional distribution $[\boldsymbol{\Omega}|\mathbf{Y},\boldsymbol{\theta}]$. It can be shown on the basis of the definition and assumptions that

$$p(\boldsymbol{\Omega}|\mathbf{Y},\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\omega}_i,\boldsymbol{\theta}) p(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i,\boldsymbol{\theta}) p(\boldsymbol{\xi}_i|\boldsymbol{\theta}). \tag{3.A10}$$

As $\boldsymbol{\omega}_i$ are mutually independent, and $\mathbf{y}_i$ are also mutually independent given $\boldsymbol{\omega}_i$, $p(\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta})$ is proportional to

$$\begin{aligned}\exp\Big\{ & -\frac{1}{2}\boldsymbol{\xi}_i^T \boldsymbol{\Phi}^{-1}\boldsymbol{\xi}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{A}\mathbf{c}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_\epsilon^{-1}(\mathbf{y}_i - \mathbf{A}\mathbf{c}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i) \\ & -\frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_\omega \mathbf{G}(\boldsymbol{\omega}_i))^T \boldsymbol{\Psi}_\delta^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_\omega \mathbf{G}(\boldsymbol{\omega}_i))\Big\}.\end{aligned} \tag{3.A11}$$

This distribution is non-standard and complex. Hence, the MH algorithm is used to generate observations from the target density $p(\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta})$ as given in (3.A11). In this algorithm, we choose $N[\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_\omega]$ as the proposal distribution, where $\boldsymbol{\Sigma}_\omega^{-1} = \boldsymbol{\Sigma}_\delta^{-1} + \boldsymbol{\Lambda}^T \boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_\delta^{-1}$ is given by

$$\boldsymbol{\Sigma}_\delta^{-1} = \begin{bmatrix} \boldsymbol{\Pi}_0^T \boldsymbol{\Psi}_\delta^{-1}\boldsymbol{\Pi}_0 & -\boldsymbol{\Pi}_0^T \boldsymbol{\Psi}_\delta^{-1}\boldsymbol{\Gamma}\boldsymbol{\Delta} \\ -\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Psi}_\delta^{-1}\boldsymbol{\Pi}_0 & \boldsymbol{\Phi}^{-1} + \boldsymbol{\Delta}^T\boldsymbol{\Gamma}^T\boldsymbol{\Psi}_\delta^{-1}\boldsymbol{\Gamma}\boldsymbol{\Delta} \end{bmatrix}$$

where $\mathbf{\Pi}_0 = \mathbf{I} - \mathbf{\Pi}$ and $\mathbf{\Delta} = [\partial \mathbf{F}(\boldsymbol{\xi}_i)/\partial \boldsymbol{\xi}_i]^T|_{\boldsymbol{\xi}_i=\mathbf{0}}$. Let $p(\cdot|\boldsymbol{\omega}, \sigma^2\mathbf{\Sigma}_\omega)$ be the proposal density corresponding to $N[\boldsymbol{\omega}, \sigma^2\mathbf{\Sigma}_\omega]$, the MH algorithm for our problem is implemented as follows: At the rth iteration with the current value $\boldsymbol{\omega}_i^{(r)}$, a new candidate $\boldsymbol{\omega}_i$ is generated from $p(\cdot|\boldsymbol{\omega}_i^{(r)}, \sigma^2\mathbf{\Sigma}_\omega)$, and accept this new candidate with the probability

$$\min\left\{1, \frac{p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta})}{p(\boldsymbol{\omega}_i^{(r)}|\mathbf{y}_i, \boldsymbol{\theta})}\right\}.$$

The variance σ^2 is chosen such that the acceptance rate is approximately 0.25 or more, see Gelman, Roberts and Gilks (1996).

Conditional Distribution $[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{\Omega}]$

When $\mathbf{\Omega}$ is given, the structural equation (see (3.17)) is just a multiple regression equation, which is slightly different from the linear regression equation (see (3.15)) associated with the linear SEMs. Hence, the components of the conditional distribution $[\boldsymbol{\theta}|\mathbf{\Omega}, \mathbf{Y}]$ involved in the Gibbs sampler in analyzing nonlinear SEMs are very similar to those in analyzing linear SEMs. To obtain $[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{\Omega}]$ for nonlinear SEMs, we only need to replace $\boldsymbol{\xi}_i$ by $\mathbf{F}(\boldsymbol{\xi}_i)$ in the corresponding conditional distributions that are derived for linear SEMs and are presented in Appendix 3.2.

Appendix 3.5: WinBUGS Code

```
model {
    for (i in 1:N) {
        for (j in 1:10) { y[i,j]~dnorm(mu[i,j], psi[j]) }
        mu[i,1]<-u[1]+eta[i]
        mu[i,2]<-u[2]+lam[1]*eta[i]
        mu[i,3]<-u[3]+lam[2]*eta[i]
        mu[i,4]<-u[4]+xi[i,1]
        mu[i,5]<-u[5]+lam[3]*xi[i,1]
        mu[i,6]<-u[6]+lam[4]*xi[i,1]
        mu[i,7]<-u[7]+lam[5]*xi[i,1]
        mu[i,8]<-u[8]+xi[i,2]
        mu[i,9]<-u[9]+lam[6]*xi[i,2]
        mu[i,10]<-u[10]+lam[7]*xi[i,2]

        #structural equation
        eta[i]~dnorm(nu[i], psd)

        nu[i]<-b*d[i]+gam[1]*xi[i,1]+gam[2]*xi[i,2]+gam[3]*xi[i,1]*xi[i,2]
               +gam[4]*xi[i,1]*xi[i,1]+gam[5]*xi[i,2]*xi[i,2]

        xi[i,1:2]~dmnorm(zero[1:2], phi[1:2,1:2])
    }   #end of i

    #prior distribution
    lam[1]~dnorm(0.9,psi[2])    lam[2]~dnorm(0.7,psi[3])
    lam[3]~dnorm(0.9,psi[5])    lam[4]~dnorm(0.7,psi[6])
    lam[5]~dnorm(0.5,psi[7])    lam[6]~dnorm(0.9,psi[9])
    lam[7]~dnorm(0.7,psi[10])

    b~dnorm(0.5, psd)           gam[1]~dnorm(0.4,psd)
    gam[2]~dnorm(0.4,psd)       gam[3]~dnorm(0.3,psd)
    gam[4]~dnorm(0.2,psd)       gam[5]~dnorm(0.5,psd)

    for (j in 1:10) {
        psi[j]~dgamma(9,4)     sgm[j]<-1/psi[j]
        u[j]~dnorm(0,1)
    }

    psd~dgamma(9,4)    sgd<-1/psd

    phi[1:2,1:2]~dwish(R[1:2,1:2], 4)
    phx[1:2,1:2]<-inverse(phi[1:2,1:2])
} #end of model
```

Appendix 3.6: R2WinBUGS Code

```
library(mvtnorm)   #Load mvtnorm package
library(R2WinBUGS) #Load R2WinBUGS package
```

```
N=500                          #Sample size
BD=numeric(N)                  #Fixed covariate in structural equation
XI=matrix(NA, nrow=N, ncol=2) #Explanatory latent variables
Eta=numeric(N)                 #Outcome latent variables
Y=matrix(NA, nrow=N, ncol=8)  #Observed variables

#The covariance matrix of xi
phi=matrix(c(1, 0.3, 0.3, 1), nrow=2)

#Estimates and standard error estimates
Eu=matrix(NA, nrow=100, ncol=10);    SEu=matrix(NA, nrow=100, ncol=10)
Elam=matrix(NA, nrow=100, ncol=7);   SElam=matrix(NA, nrow=100, ncol=7)
Eb=numeric(100);                     SEb=numeric(100)
Egam=matrix(NA, nrow=100, ncol=5);   SEgam=matrix(NA, nrow=100, ncol=5)
Esgm=matrix(NA, nrow=100, ncol=10);  SEsgm=matrix(NA, nrow=100, ncol=10)
Esgd=numeric(100);                   SEsgd=numeric(100)
Ephx=matrix(NA, nrow=100, ncol=3);   SEphx=matrix(NA, nrow=100, ncol=3)

R=matrix(c(1.0, 0.3, 0.3, 1.0), nrow=2)

parameters=c("u", "lam", "b", "gam", "sgm", "sgd", "phx")

init1=list(u=rep(0,10), lam=rep(0,7), b=0, gam=rep(0,5), psi=rep(1,10),
           psd=1, phi=matrix(c(1, 0, 0, 1), nrow=2))

init2=list(u=rep(1,10), lam=rep(1,7), b=1, gam=rep(1,5), psi=rep(2,10),
           psd=2, phi=matrix(c(2, 0, 0, 2), nrow=2))

inits=list(init1, init2)

eps=numeric(10)

for (t in 1:100) {
    #Generate Data
    for (i in 1:N) {
```

```
    BD[i]=rt(1, 5)

    XI[i,]=rmvnorm(1, c(0,0), phi)

    delta=rnorm(1, 0, sqrt(0.36))
    Eta[i]=0.5*BD[i]+0.4*XI[i,1]+0.4*XI[i,2]+0.3*XI[i,1]*XI[i,2]
           +0.2*XI[i,1]*XI[i,1]+0.5*XI[i,2]*XI[i,2]+delta

    eps[1:3]=rnorm(3, 0, sqrt(0.3))
    eps[4:7]=rnorm(4, 0, sqrt(0.5))
    eps[8:10]=rnorm(3, 0, sqrt(0.4))
    Y[i,1]=Eta[i]+eps[1]
    Y[i,2]=0.9*Eta[i]+eps[2]
    Y[i,3]=0.7*Eta[i]+eps[3]
    Y[i,4]=XI[i,1]+eps[4]
    Y[i,5]=0.9*XI[i,1]+eps[5]
    Y[i,6]=0.7*XI[i,1]+eps[6]
    Y[i,7]=0.5*XI[i,1]+eps[7]
    Y[i,8]=XI[i,2]+eps[8]
    Y[i,9]=0.9*XI[i,2]+eps[9]
    Y[i,10]=0.7*XI[i,2]+eps[10]
}

#Run WinBUGS
data=list(N=500, zero=c(0,0), d=BD, R=R, y=Y)

model<-bugs(data,inits,parameters,
            model.file="C:/Simulation/model.txt",
            n.chains=2,n.iter=10000,n.burnin=4000,n.thin=1,
            bugs.directory="C:/Program Files/WinBUGS14/",
            working.directory="C:/Simulation/")

#Save Estimates
Eu[t,]=model$mean$u;               SEu[t,]=model$sd$u
Elam[t,]=model$mean$lam;           SElam[t,]=model$sd$lam
Eb[t]=model$mean$b;                SEb[t]=model$sd$b
```

```
    Egam[t,]=model$mean$gam;            SEgam[t,]=model$sd$gam
    Esgm[t,]=model$mean$sgm;            SEsgm[t,]=model$sd$sgm
    Esgd[t]=model$mean$sgd;             SEsgd[t]=model$sd$sgd
    Ephx[t,1]=model$mean$phx[1,1];      SEphx[t,1]=model$sd$phx[1,1]
    Ephx[t,2]=model$mean$phx[1,2];      SEphx[t,2]=model$sd$phx[1,2]
    Ephx[t,3]=model$mean$phx[2,2];      SEphx[t,3]=model$sd$phx[2,2]
}
```

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669-679.

Ansari, A. and Jedidi, K. (2000) Bayesian factor analysis for multilevel binary observations. *Psychometrika*, **65**, 475-496.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Boomsma, A. (1982) The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog and H. Wold (eds), *Systems under Indirect Observation: Causality, Structure, Prediction.* pp. 149-173. Amsterdam: North-Holland.

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

Chou, C. P., Bentler, P. M. and Satorra, A. (1991) Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, **44**, 347-357.

Congdon, P. (2003) *Applied Bayesian Modeling.* Hoboken, New York: John Wiley & Sons Inc.

Dunson, D. B. (2000) Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355-366.

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

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

Gelman, A., Roberts, G. O. and Gilks, W. R. (1996) Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 5*, pp. 599-607. Oxford: Oxford University Press.

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

Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473-483.

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

Hoogland, J. J. and Boomsma, A. (1998) Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, **26**, 329-367.

Hu, L., Bentler, P. M. and Kano, Y. (1992) Can test statistics in covariance structure analysis be trusted. *Psychological Bulletin*, **112**, 351-362.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Lawson, A. B., Browne, W. J. and Vidal Rodeiro, C. L. (2003) *Disease Mapping with WinBUGS and MLWIN.* Cluchester: John Wiley & Sons, Ltd.

Lee, S. Y. (2007) *Structural Equation Modeling: A Bayesian Approach.* UK: John Wiley & Sons, Ltd.

Lee, S. Y. and Shi, J. Q. (2000) Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Annals of the Institute of Statistical mathematics*, **52**, 722-736.

Lee, S. Y. and Song, X. Y. (2004) Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, **39**, 653-686.

Lee, S. Y., Song, X. Y. and Tang, N. S. (2007) Bayesian methods for analyzing structural equation models with covariates, interaction and quadratic latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, **14**, 404-434.

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

Muirhead, R. J. (1982) *Aspects of Multivariate Statistical Theory.* New York: John Wiley & Sons, Inc.

Rubin, D. B. (1991) EM and beyond. *Psychometrika*, **56**, 241-254.

Scheines, R., Hoijtink, H. and Boomsma, A. (1999) Bayesian estimation and testing of structural equation models. *Psychometrika*, **64**, 37-52.

Shi, J. Q. and Lee, S. Y. (1998) Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, **51**, 233-252.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Lunn, D. (2003) *WinBUGS User Manual. Version 1.4*. Cambridge, UK: MRC Biostatistics Unit.

Sturtz, S., Ligges, U. and Gelman, A. (2005) R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, **12**, 1-16.

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

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics.* New York: John Wiley.

Table 3.1: Bayesian estimates of the artificial example obtained from WinBUGS.

Par	True value	EST	SE	Par	True value	EST	SE
μ_1	0.0	0.022	0.069	$\psi_{\epsilon 1}$	0.3	0.324	0.032
μ_2	0.0	0.065	0.062	$\psi_{\epsilon 2}$	0.3	0.285	0.027
μ_3	0.0	0.040	0.052	$\psi_{\epsilon 3}$	0.3	0.284	0.022
μ_4	0.0	0.003	0.058	$\psi_{\epsilon 4}$	0.5	0.558	0.050
μ_5	0.0	0.036	0.056	$\psi_{\epsilon 5}$	0.5	0.480	0.045
μ_6	0.0	0.002	0.047	$\psi_{\epsilon 6}$	0.5	0.554	0.041
μ_7	0.0	0.004	0.042	$\psi_{\epsilon 7}$	0.5	0.509	0.035
μ_8	0.0	0.092	0.053	$\psi_{\epsilon 8}$	0.4	0.382	0.035
μ_9	0.0	0.032	0.050	$\psi_{\epsilon 9}$	0.4	0.430	0.035
μ_{10}	0.0	-0.000	0.044	$\psi_{\epsilon 10}$	0.4	0.371	0.029
λ_{21}	0.9	0.889	0.022	b_1	0.5	0.525	0.075
λ_{31}	0.7	0.700	0.019	γ_1	0.4	0.438	0.059
λ_{52}	0.9	0.987	0.053	γ_2	0.4	0.461	0.034
λ_{62}	0.7	0.711	0.046	γ_3	0.3	0.304	0.045
λ_{72}	0.5	0.556	0.040	γ_4	0.2	0.184	0.060
λ_{93}	0.9	0.900	0.042	γ_5	0.5	0.580	0.050
$\lambda_{10,3}$	0.7	0.766	0.038	ϕ_{11}	1.0	1.045	0.120
				ϕ_{12}	0.3	0.302	0.057
				ϕ_{22}	1.0	1.023	0.089
				ψ_δ	0.36	0.376	0.045

Table 3.2: Bayesian estimates of the artificial example obtained from WinBUGS based on 100 replications.

Par	AB	RMS	Par	AB	RMS
μ_1	0.009	0.068	$\psi_{\epsilon 1}$	0.008	0.027
μ_2	0.001	0.064	$\psi_{\epsilon 2}$	0.010	0.028
μ_3	0.003	0.050	$\psi_{\epsilon 3}$	0.004	0.021
μ_4	0.008	0.058	$\psi_{\epsilon 4}$	0.012	0.046
μ_5	0.000	0.055	$\psi_{\epsilon 5}$	0.000	0.047
μ_6	0.005	0.046	$\psi_{\epsilon 6}$	0.002	0.038
μ_7	0.005	0.041	$\psi_{\epsilon 7}$	0.009	0.036
μ_8	0.002	0.051	$\psi_{\epsilon 8}$	0.012	0.037
μ_9	0.001	0.048	$\psi_{\epsilon 9}$	0.001	0.032
μ_{10}	0.001	0.037	$\psi_{\epsilon 10}$	0.006	0.031
λ_{21}	0.006	0.021	b_1	0.001	0.030
λ_{31}	0.001	0.022	γ_1	0.019	0.056
λ_{52}	0.021	0.063	γ_2	0.000	0.066
λ_{62}	0.016	0.047	γ_3	0.003	0.071
λ_{72}	0.015	0.043	γ_4	0.021	0.048
λ_{93}	0.004	0.046	γ_5	0.018	0.062
$\lambda_{10,3}$	0.003	0.037	ϕ_{11}	0.046	0.107
			ϕ_{21}	0.017	0.053
			ϕ_{22}	0.088	0.040
			ψ_δ	0.013	0.040

Note: 'AB' and 'RMS' denote the averages of the absolute bias and the root mean square values, respectively.

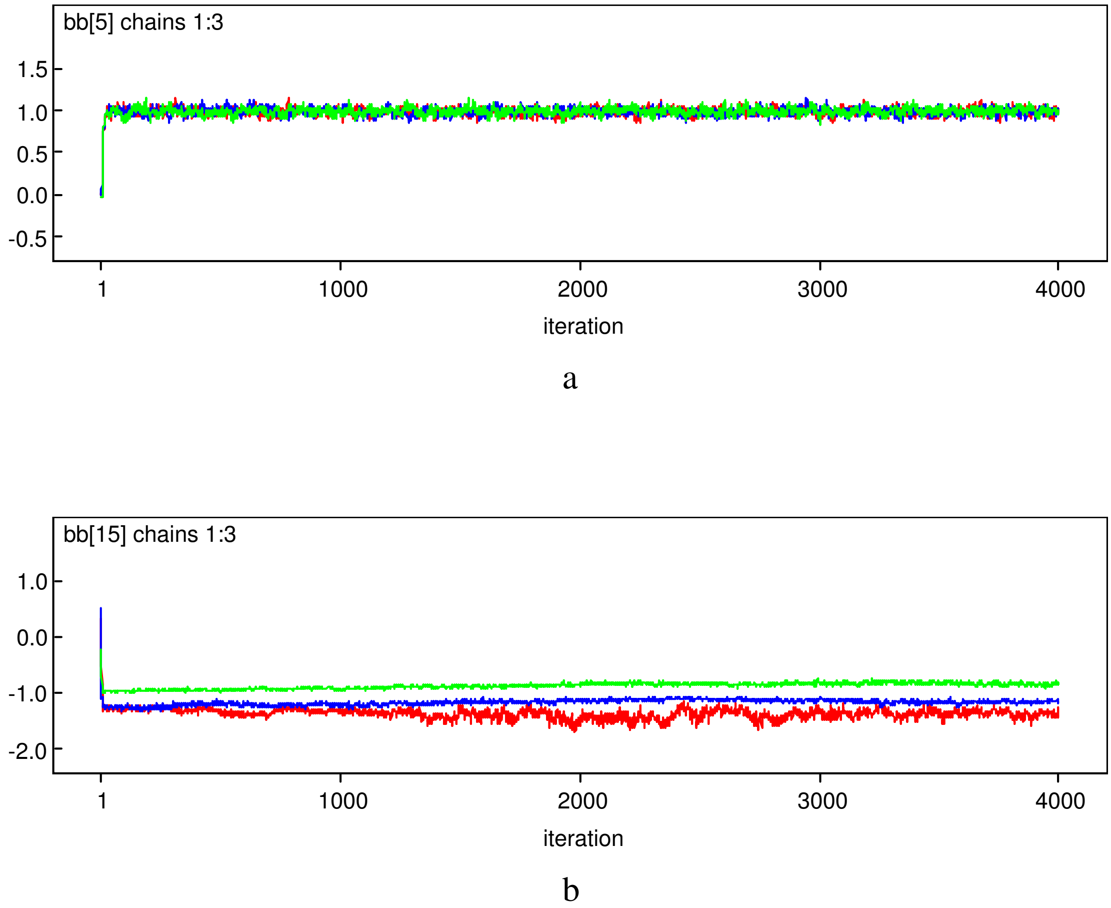


Figure 3.1: Sample traces of chains from which: (a) convergence looks reasonable; (b) convergence does not reach.

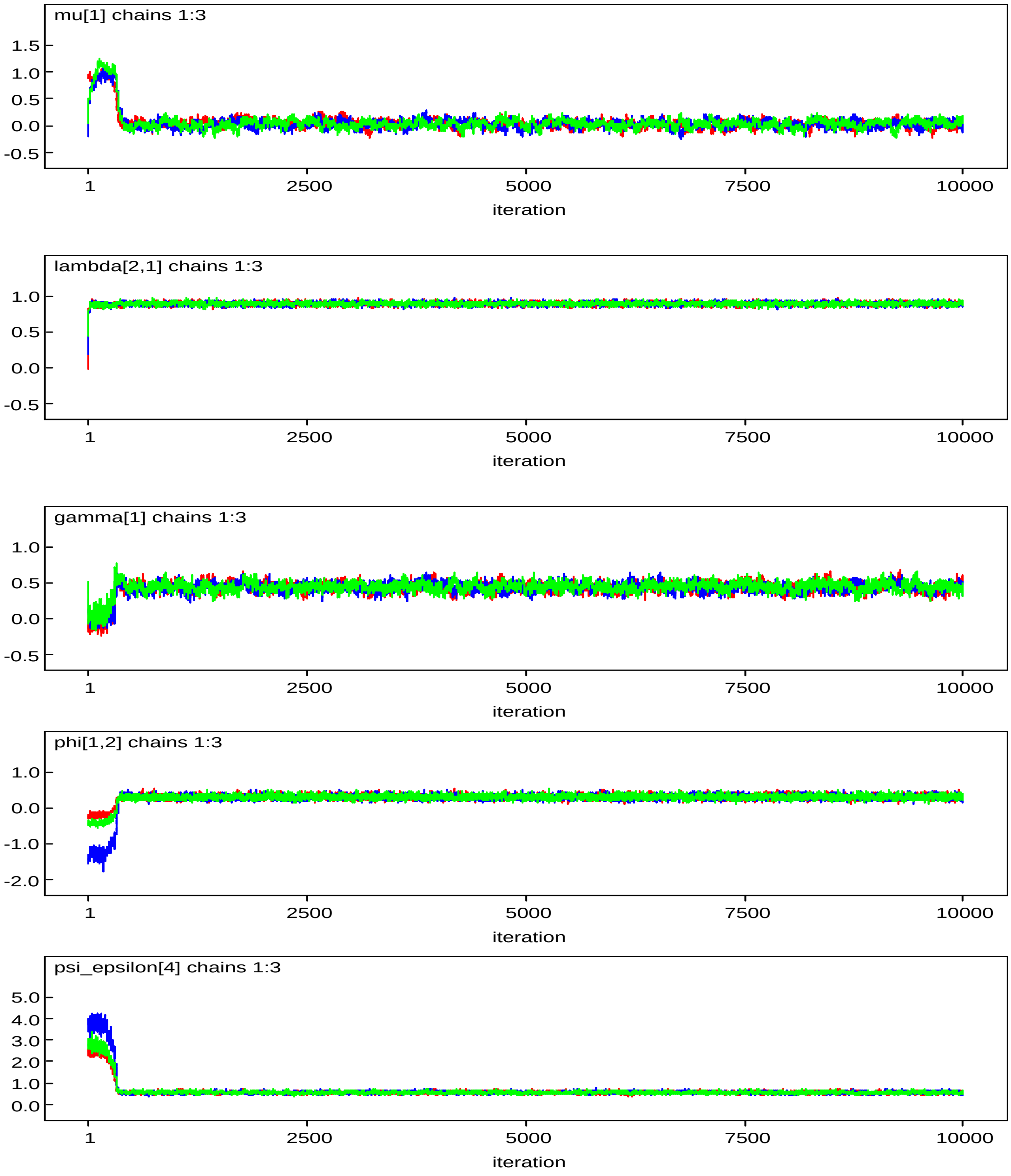


Figure 3.2: Plots from top to bottom represent three chains of observations corresponding to μ_1, λ_{21}, γ_1, ϕ_{12} and $\psi_{\epsilon 4}$, generated by different initial values.

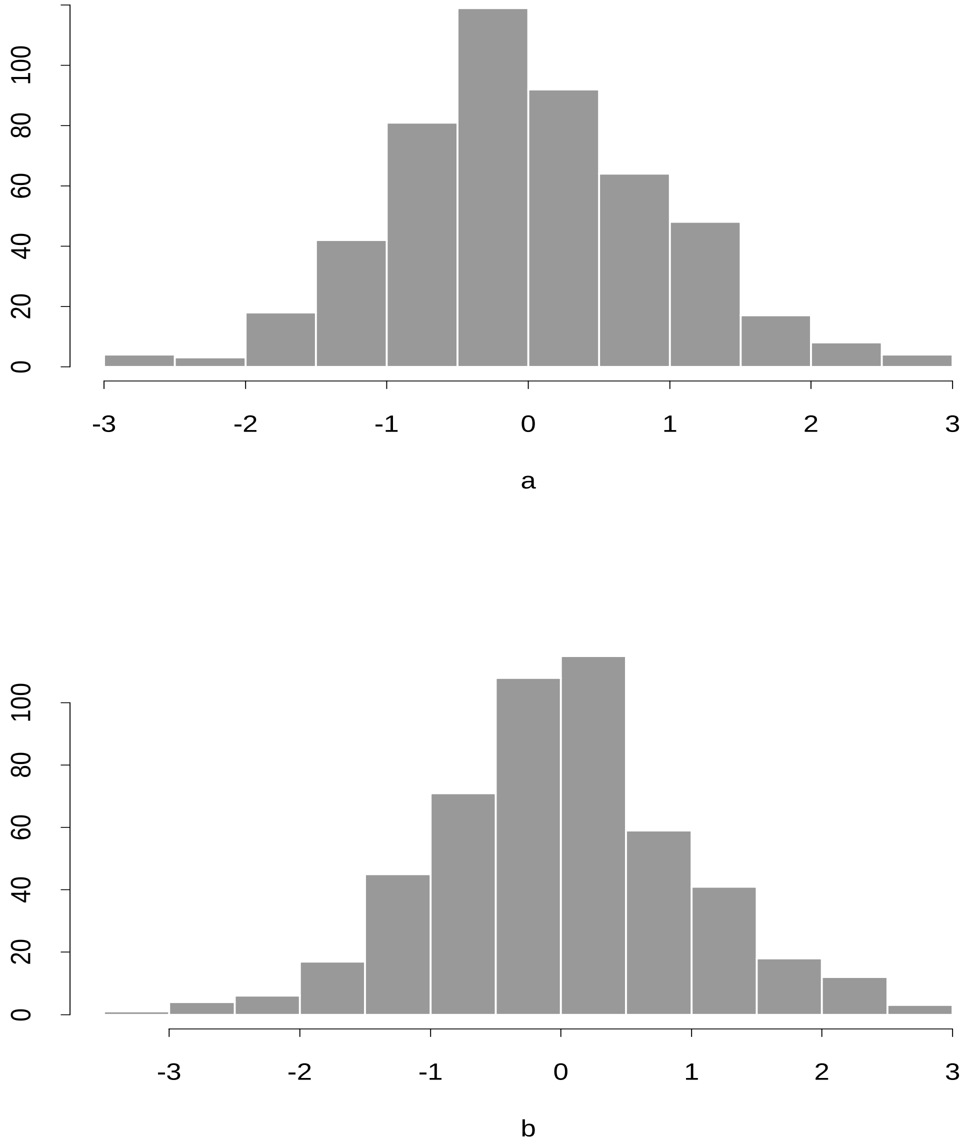
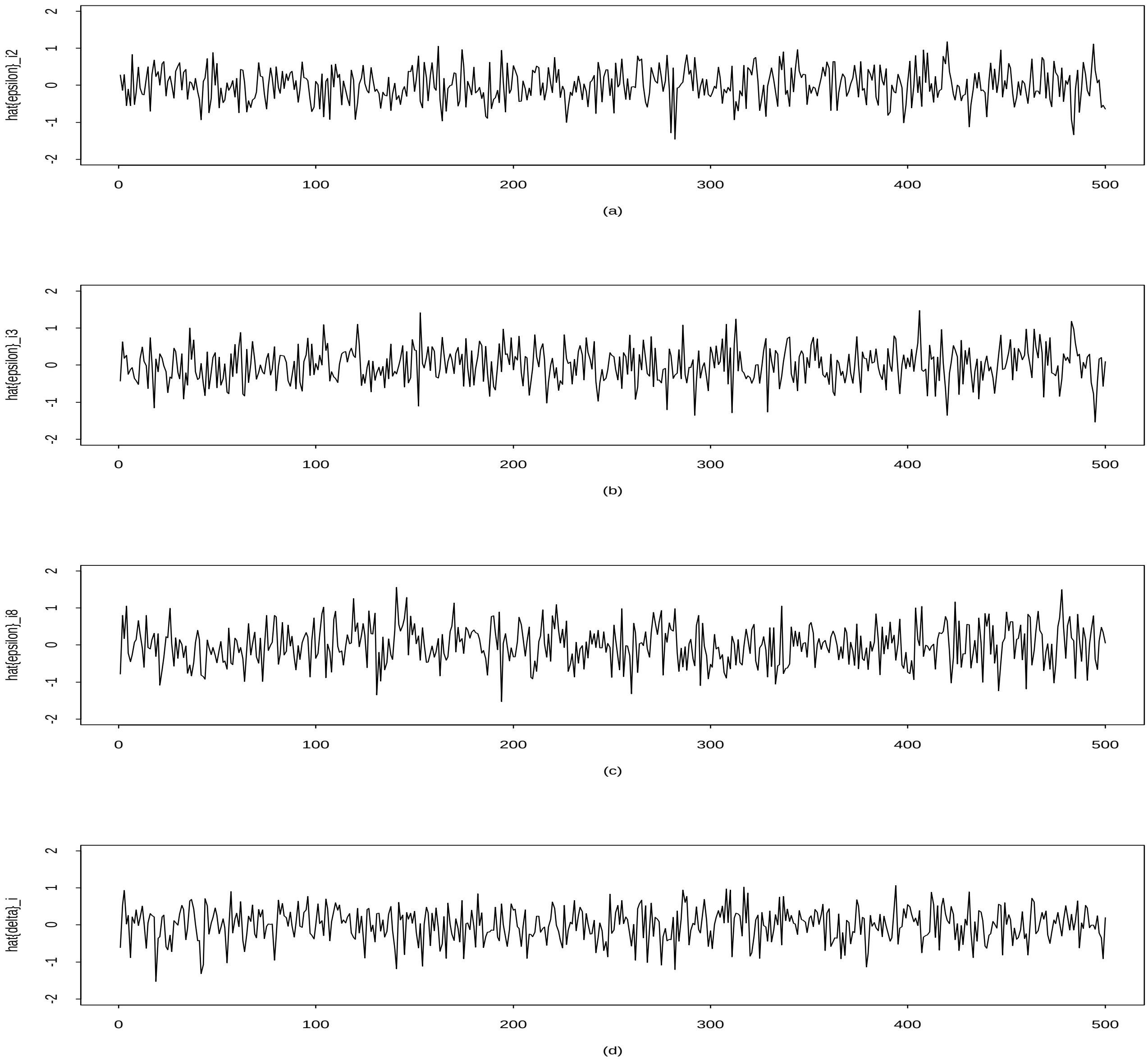


Figure 3.3: Histograms of the latent variables (a) $\hat{\xi}_{i1}$ and (b) $\hat{\xi}_{i2}$.



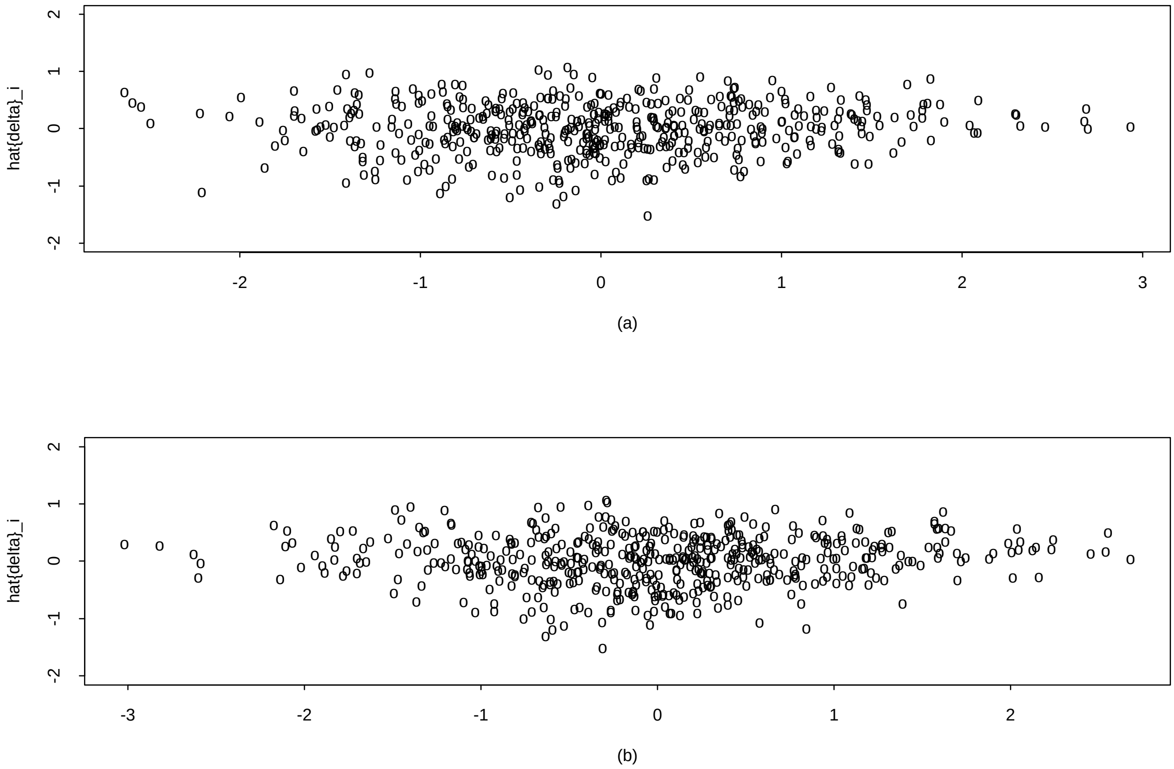
(a)

(b)

(c)

(d)

Figure 3.4: Estimated residual plots, (a) $\hat{\epsilon}_{i2}$, (b) $\hat{\epsilon}_{i3}$, (c) $\hat{\epsilon}_{i8}$, and (d) $\hat{\delta}_i$.



(a)

(b)

Figure 3.5: Plots of estimated residuals $\hat{\delta}_i$ versus (a) $\hat{\xi}_{i1}$, (b) $\hat{\xi}_{i2}$.

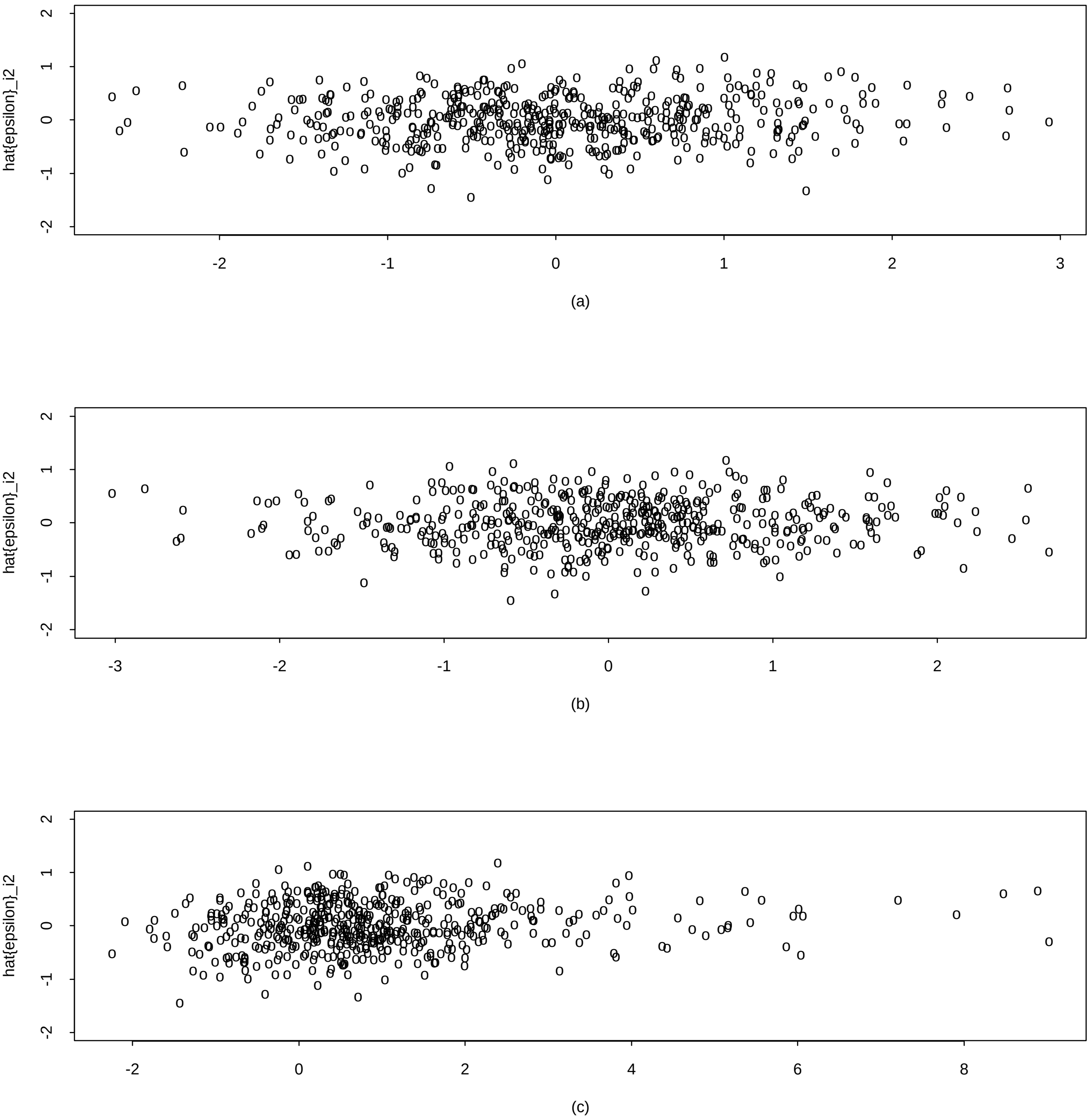


Figure 3.6: Plots of estimated residuals $\hat{\epsilon}_{i2}$ versus (a) $\hat{\xi}_{i1}$, (b) $\hat{\xi}_{i2}$, and (c) $\hat{\eta}_i$.