Chapter 3 Bayesian Methods for Estimating Structural Equation Models

The traditional approach for analyzing SEMs is the covariance structure analysis approach. Under some standard assumptions, for example, the random observations are i.i.d. following a normal distribution, this approach works fine. Hence, almost all classical commercial software in SEMs were developed based on this approach with the sample covariance matrix $\mathbf{S}$.

$\hat{\theta} = f(S, \Sigma(\theta))$ such as missing data.

However, under more complex situations (will be discussed in the subsequent chapters), the covariance structure analysis approach based on $\mathbf{S}$ is not effective and may encounter theoretical and computational problems.

In this chapter, we introduce the Bayesian approach which has nice features and can be effectively applied to analyze not only the standard SEMs but also their useful generalizations.

The objective of this chapter:

1. Introduce the basic ideas of the Bayesian approach in estimation, including the discussion of the prior distribution.
2. Introduce posterior analysis and MCMC methods.
3. Present an application of the MCMC methods.
4. Describes how to apply the software WinBUGS to obtain Bayesian estimation and to conduct simulation studies.

Notations and concepts related to the Bayesian approach:

- M — an arbitrary SEM with a vector of unknown parameters $\boldsymbol{\theta}$.
- $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ — the observed data set
- $p(\boldsymbol{\theta}|M)$ (or $p(\theta)$) — the prior density function of $\boldsymbol{\theta}$.
- $p(\mathbf{Y}, \boldsymbol{\theta}|M)$ — the probability density function of the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\theta}$ under M.
- $p(\boldsymbol{\theta}|\mathbf{Y}, M)$ — the density function of the posterior distribution of $\boldsymbol{\theta}$ under M. This function fully describes the behavior of $\boldsymbol{\theta}$ under the given data $\mathbf{Y}$.
- $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ — the likelihood function under M.

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

$$\begin{aligned} &p(\boldsymbol{\theta}|\mathbf{Y}, M) \propto p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}), \quad \text{or} \\ &\log p(\boldsymbol{\theta}|\mathbf{Y}, M) = \log p(\mathbf{Y}|\boldsymbol{\theta}, M) + \log p(\boldsymbol{\theta}) + \text{constant}. \end{aligned} \tag{1}$$

MLE MAP JLN

It follows from (1) that

1. $p(\boldsymbol{\theta}|\mathbf{Y}, M)$ incorporates the sample information through $p(\mathbf{Y}|\boldsymbol{\theta}, M)$, and the prior information through $p(\boldsymbol{\theta})$.
2. $p(\mathbf{Y}|\boldsymbol{\theta}, M)$ depends on n, whereas $p(\boldsymbol{\theta})$ does not. When n becomes arbitrarily large, $\log p(\mathbf{Y}|\boldsymbol{\theta}, M)$ could dominate $\log p(\boldsymbol{\theta})$. In this situation, $p(\boldsymbol{\theta})$ plays a less important role, and $\log p(\boldsymbol{\theta}|\mathbf{Y}, M)$ is close to $\log p(\mathbf{Y}|\boldsymbol{\theta}, M)$. Hence, asymptotically Bayesian and ML approaches are equivalent, and the Bayesian estimates have the same optimal properties as the ML estimates.

3. When the sample sizes are small or moderate, the prior distribution of $\boldsymbol{\theta}$ plays a more substantial role in Bayesian estimation. Hence, in the problems with small or moderate sample sizes, $p(\boldsymbol{\theta})$ is useful for achieving better results.

For many problems in biomedical and behavioral sciences, researchers may have good prior information from

- the subject experts,
- analysis of similar or past data, $\Rightarrow p(\theta)$
- some other sources.

Hence, 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; e.g., $p(\boldsymbol{\theta}|\mathbf{Y})$ 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.

What is the difference?

Basically, there are two kinds of prior distributions:

1. Noninformative prior distribution — we have little prior information about $\boldsymbol{\theta}$, and hence $p(\boldsymbol{\theta})$ plays a minimal role in $p(\boldsymbol{\theta}|\mathbf{Y})$. The associated prior density is 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.
2. Informative prior distributions — we may have useful prior knowledge about $\boldsymbol{\theta}$, either from closely related data or from subjective knowledge of experts. Usually, an informative prior distribution has its own parameters, which are called hyperparameters.

$\theta = [\alpha \ \beta]^T \quad \theta \sim \mathcal{N}(\theta_0, \Sigma)$ — vagueness. (if $\Sigma \to \infty$, $\theta \to$ Uniform); hyperparameters; your knowledge

$$\log p(\theta|Y,M) = \log p(Y|\theta,M) + \underbrace{\log p(\theta|M)}_{\propto 1} + \text{constant}$$
$$\propto \log p(Y|\theta,M) + \text{constant}.$$

A commonly used informative prior distribution 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

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

If the prior density of θ has the same form, it can be seen from (1) that the posterior density will also has this form. Consider the following prior density of θ:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{2}$$

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

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}, \end{aligned} \tag{3}$$

which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$.

Another example of conjugate prior distributions:

Assume that $y_1, \cdots, y_n$ are i.i.d. $\sim N[\mu, \sigma^2]$, and $\boldsymbol{\theta} = (\mu, \sigma^2)$.
Let $\mathbf{Y} = (y_1, \cdots, y_n)$, the likelihood function is

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sigma^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.$$

Consider $p(\mathbf{Y}|\boldsymbol{\theta})$ as a function of μ (σ^2 is given), the likelihood is an exponential of a quadratic form in μ.

A conjugate prior distribution of μ can be parameterized as

$$p(\mu) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\},$$

that is, $\mu \stackrel{D}{=} N[\mu_0, \sigma_0^2]$, where μ_0 and σ_0^2 are hyperparameters.

The conditional posterior density of $p(\mu|\mathbf{Y}, \sigma^2)$ is

$$\begin{aligned} p(\mu|\mathbf{Y}, \sigma^2) &\propto p(\mu)p(\mathbf{Y}|\boldsymbol{\theta}) \\ &\propto \exp\left\{ -\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 \right\} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \right\} \\ &\propto \exp\left[-\frac{1}{2}\left\{ \frac{1}{\sigma_0^2}(\mu - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \right\} \right]. \end{aligned}$$

It can be shown that $[\mu|\mathbf{Y}, \sigma^2] \overset{D}{=} N[\tilde{\mu}, \tilde{\sigma}^2]$, where

$$\tilde{\mu} = \tilde{\sigma}^2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i + \frac{\mu_0}{\sigma_0^2}\right), \qquad \tilde{\sigma}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

If we consider $p(\mathbf{Y}|\boldsymbol{\theta})$ as a function of σ^2 (μ is given), then

$$p(\mathbf{Y}|\boldsymbol{\theta}) \quad \propto \quad (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}.$$

A conjugate prior distribution of σ^2 can be parameterized as

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha_0+1)} \exp(-\beta_0/\sigma^2),$$

that is, $\sigma^2 \stackrel{D}{=} IG(\alpha_0, \beta_0)$, where $IG(\alpha_0, \beta_0)$ is the inverted Gamma distribution with hyperparameters α_0 and β_0. Thus,

$$\begin{aligned} p(\sigma^2|\mathbf{Y}, \mu) \quad &\propto \quad p(\sigma^2) p(\mathbf{Y}|\boldsymbol{\theta}) \\ &\propto \quad (\sigma^2)^{-(\frac{n}{2}+\alpha_0+1)} \exp\left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{n} (y_i - \mu)^2 + \beta_0 \right\} \right], \quad \text{or} \end{aligned}$$

$[\sigma^2|\mathbf{Y}, \mu] \stackrel{D}{=} IG(\tilde{\alpha}, \tilde{\beta})$, with $\tilde{\alpha} = \frac{n}{2} + \alpha_0$ and $\tilde{\beta} = \frac{1}{2}\left\{ \sum_{i=1}^{n} (y_i - \mu)^2 + \beta_0 \right\}$.

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.

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. In this book, we will use the conjugate prior distributions in our Bayesian analyses.

As an illustration, we consider the following model:

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{4}$$

$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{5}$$

where $\boldsymbol{\mu}$ is a vector of intercepts, $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$, $\boldsymbol{\Lambda}$, $\mathbf{B}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Gamma}$ are parameter matrices of unknown regression coefficients, and $\mathbf{F}(\cdot)$ is a given vector of differentiable functions of $\boldsymbol{\xi}_i$. 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 A1-A4 are satisfied.

In this model, the unknown parameters are

- $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, $\mathbf{B}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Gamma}$ — related to the mean vectors of $\mathbf{y}_i$ and $\boldsymbol{\eta}_i$,
- $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}_\epsilon$, and $\boldsymbol{\Psi}_\delta$ — 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 $(\boldsymbol{\Lambda}_k, \psi_{\epsilon k})$ are

$$\begin{aligned} &\psi_{\epsilon k} \overset{D}{=} IG[\alpha_{0\epsilon k}, \beta_{0\epsilon k}] \text{ or } \psi_{\epsilon k}^{-1} \overset{D}{=} Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}], \\ &\boldsymbol{\mu} \overset{D}{=} N[\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0], \text{ and } [\boldsymbol{\Lambda}_k | \psi_{\epsilon k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0k}, \psi_{\epsilon k} \mathbf{H}_{0yk}], \end{aligned} \tag{6}$$

where $\alpha_{0\epsilon k}, \beta_{0\epsilon k}$, and elements in $\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{0k}, \boldsymbol{\Sigma}_0$, and $\mathbf{H}_{0yk}$ are hyperparameters, and $\boldsymbol{\Sigma}_0$ and $\mathbf{H}_{0yk}$ are positive definite matrices.

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\eta}_i &= \mathbf{B}\mathbf{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \end{aligned}$$

We rewrite the structural equation (5) as:

$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i = \boldsymbol{\Lambda}_\omega \mathbf{G}(\boldsymbol{\omega}_i) + \boldsymbol{\delta}_i, \tag{7}$$

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

try how to derive it. □

$$\begin{aligned}
&\boldsymbol{\Phi} \overset{D}{=} IW_{q_2}[\mathbf{R}_0^{-1}, \rho_0], \text{ or equivalently } \boldsymbol{\Phi}^{-1} \overset{D}{=} W_{q_2}[\mathbf{R}_0, \rho_0], \\
&\psi_{\delta k} \overset{D}{=} IG[\alpha_{0\delta k}, \beta_{0\delta k}] \text{ or } \psi_{\delta k}^{-1} \overset{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}], \\
&[\boldsymbol{\Lambda}_{\omega k} | \psi_{\delta k}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0\omega k}, \psi_{\delta k}\mathbf{H}_{0\omega k}],
\end{aligned} \tag{8}$$

where $W_{q_2}[\mathbf{R}_0, \rho_0]$ is a q_2-dimensional Wishart distribution with hyper-parameters 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, $\boldsymbol{\Lambda}_{0\omega k}$, $\alpha_{0\delta k}, \beta_{0\delta k}$, and the positive definite matrix $\mathbf{H}_{0\omega k}$ are hyperparameters. Note that the prior distribution of $\boldsymbol{\Phi}^{-1}$ (or $\boldsymbol{\Phi}$) is a multivariate extension of the prior distribution of $\psi_{\delta k}^{-1}$ (or $\psi_{\delta k}$).

In specifying conjugate prior distributions, we assign values to their hyper-parameters. These preassigned values (prior inputs) represent the available prior knowledge. In general,

- If we have good prior information about a parameter — select the prior distribution with a small variance, e.g., 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 $0.5\mathbf{I}$).
- If we have no good information about a parameter — select the prior distribution with a larger variance.

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i,$$
$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \mathbf{\Pi}\boldsymbol{\eta}_i + \mathbf{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i,$$

The choice of $\alpha_{0\epsilon k}$ and $\beta_{0\epsilon k}$ is based on the same rationale. Note that $\epsilon_k \overset{D}{=} N[0, \psi_{\epsilon k}]$, 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. If $\psi_{\epsilon k} \overset{D}{=} \textit{Inverted Gamma}(\alpha_{0\epsilon k}, \beta_{0\epsilon k})$, then

$$\begin{aligned} E(\psi_{\epsilon k}) &= \beta_{0\epsilon k}/(\alpha_{0\epsilon k} - 1), \\ Var(\psi_{\epsilon k}) &= \beta_{0\epsilon k}^2/\{(\alpha_{0\epsilon k} - 1)^2(\alpha_{0\epsilon k} - 2)\}. \end{aligned}$$

For instance,

- if $\alpha_{0\epsilon k} = 9$ and $\beta_{0\epsilon k} = 4$, then

$$E(\psi_{\epsilon k}) = 4/8 = 0.5, \qquad Var(\psi_{\epsilon k}) = 4^2/\{(9-1)^2(9-2)\} = 1/28;$$

- if $\alpha_{0k} = 6$ and $\beta_{0\epsilon k} = 10$, then

$$E(\psi_{\epsilon k}) = 2.0, \qquad Var(\psi_{\epsilon k}) = 1.0.$$

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 (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

$$E(\mathbf{\Phi}) = \mathbf{R}_0^{-1}/(\rho_0 - q_2 - 1).$$

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:

$$
\begin{aligned}
p(\mathbf{\Lambda}, \mathbf{\Psi}_\epsilon) &\propto p(\psi_{\epsilon 1}, \cdots, \psi_{\epsilon p}) \propto \prod_{k=1}^{p} \psi_{\epsilon k}^{-1}, \\
p(\mathbf{\Lambda}_\omega, \mathbf{\Psi}_\delta) &\propto p(\psi_{\delta 1}, \cdots, \psi_{\delta q_1}) \propto \prod_{k=1}^{q_1} \psi_{\delta k}^{-1}, \\
p(\mathbf{\Phi}) &\propto |\mathbf{\Phi}|^{-(q_2+1)/2}.
\end{aligned}
\tag{9}
$$

In (9), no hyperparameters are involved, and the prior distributions of the unknown parameters in $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_\omega$ are implicitly taken to be proportional to a constant. Bayesian analysis based on (9) is close to that with the conjugate prior distributions given by (6) and (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.

$\mu \sim N(\mu_0, \sigma_0^2)$, $\mu_0 \sim N(\mu_{00}, \sigma_{00}^2)$, $\sigma_0^2 \sim IG(\alpha_{00}, \beta_{00})$

In general, 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.

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.

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.

Assign a distribution to it,

Data augmentation provides a useful approach to cope with the problem that is induced by latent variables. 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. Based on 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.

Some MCMC methods can then be used 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 $p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y})$. A useful algorithm to achieve this goal is the Gibbs sampler (Geman and Geman, 1984).

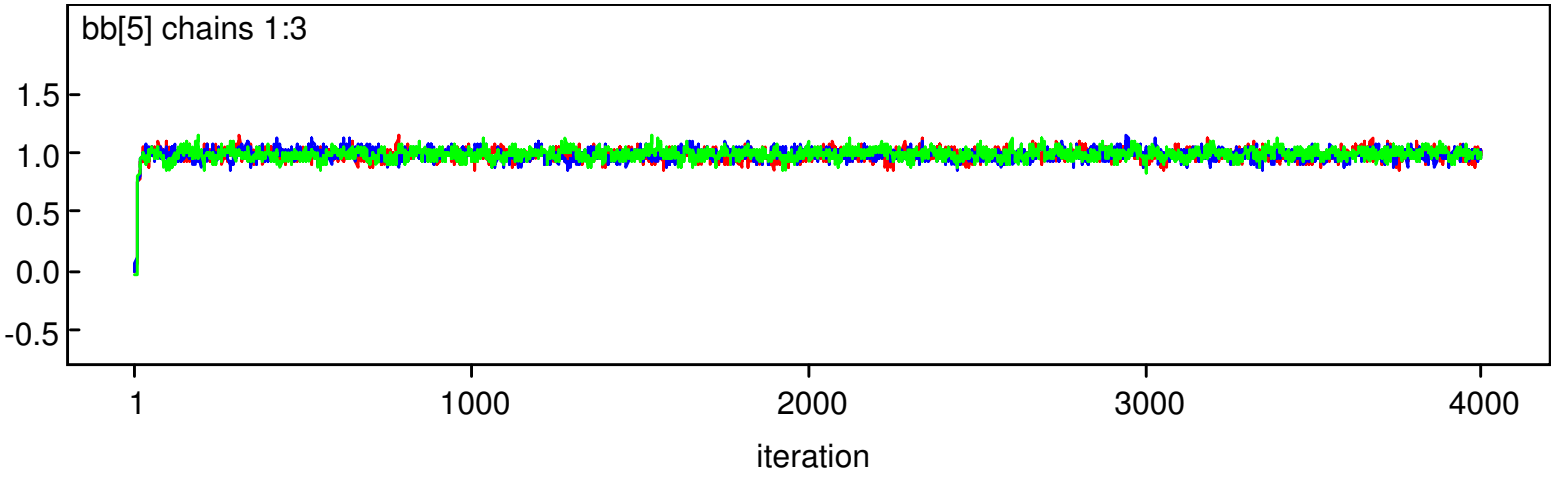
Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_a)$ and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_b)$. The Gibbs sampler performs an alternating conditional sampling at each iteration. It cycles through the components of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$, drawing each component conditional on the others. 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,

$$
\begin{aligned}
\boldsymbol{\theta}_1^{(j+1)} &\text{ from } p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(j)}, \cdots, \boldsymbol{\theta}_a^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\
\boldsymbol{\theta}_2^{(j+1)} &\text{ from } p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(j+1)}, \cdots, \boldsymbol{\theta}_a^{(j)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\
&\vdots \qquad\qquad \vdots \\
\boldsymbol{\theta}_a^{(j+1)} &\text{ from } p(\boldsymbol{\theta}_a | \boldsymbol{\theta}_1^{(j+1)}, \cdots, \boldsymbol{\theta}_{a-1}^{(j+1)}, \boldsymbol{\Omega}^{(j)}, \mathbf{Y}), \\
\boldsymbol{\Omega}_1^{(j+1)} &\text{ from } p(\boldsymbol{\Omega}_1 | \boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_2^{(j)}, \cdots, \boldsymbol{\Omega}_b^{(j)}, \mathbf{Y}), \\
\boldsymbol{\Omega}_2^{(j+1)} &\text{ from } p(\boldsymbol{\Omega}_2 | \boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_1^{(j+1)}, \cdots, \boldsymbol{\Omega}_b^{(j)}, \mathbf{Y}), \\
&\vdots \qquad\qquad \vdots \\
\boldsymbol{\Omega}_b^{(j+1)} &\text{ from } p(\boldsymbol{\Omega}_b | \boldsymbol{\theta}^{(j+1)}, \boldsymbol{\Omega}_1^{(j+1)}, \cdots, \boldsymbol{\Omega}_{b-1}^{(j+1)}, \mathbf{Y}).
\end{aligned}
\tag{10}
$$

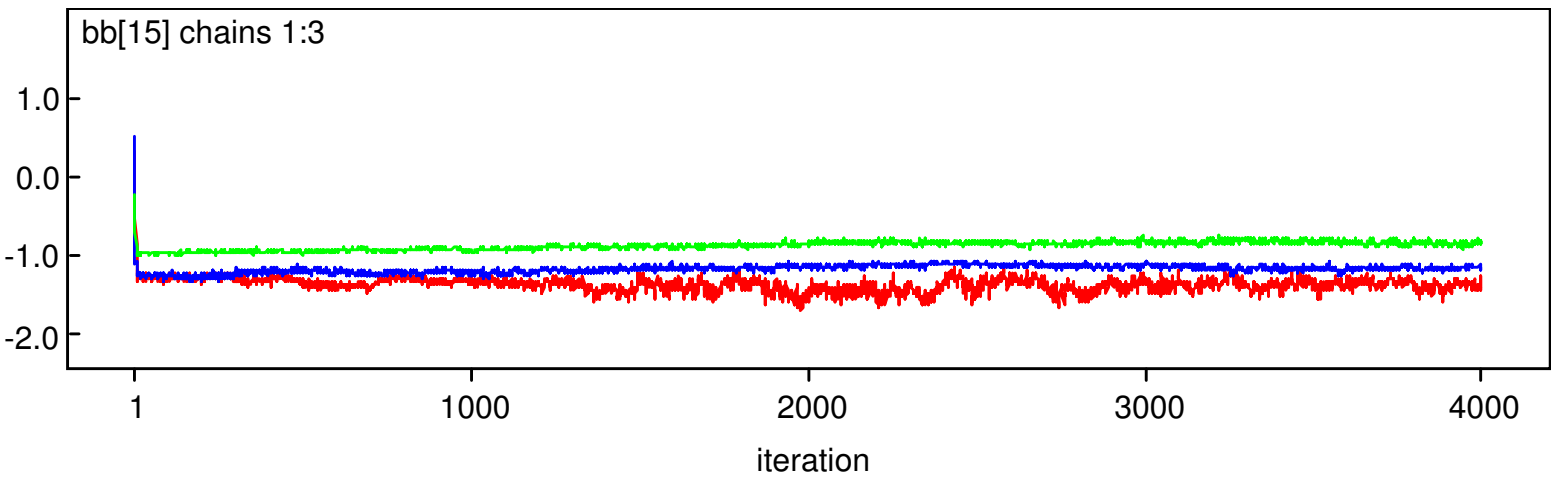
full conditional density.

Most of the full conditional distributions in (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) can be used.

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. 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 given in Figure 3.1.


bb[5] chains 1:3
1.5
1.0
0.5
0.0
-0.5
1
1000
2000
3000
4000
iteration

a


bb[15] chains 1:3
1.0
0.0
-1.0
-2.0
1
1000
2000
3000
4000
iteration

b

Another procedure to monitor convergence is using 'estimated potential scale reduction (EPSR)' value. As suggested by Gelmen (1996), convergence is achieved when the EPSR values are all less than 1.2. The computation of the EPSR values is based on several simulation sequences generated independently from different starting points. The EPSR approach monitors each parameter of interest separately.

Let n be the length of the each sequence. For each parameter estimate, say θ, let θ_{jk} $(j = 1, \cdots, n; k = 1, \cdots, K)$ be the observations from K parallel sequences of length n. The between- and within-sequence variances are computed as

$$B = \frac{n}{K-1}\sum_{k=1}^{K}(\theta_{\cdot k} - \theta_{\cdot\cdot})^2, \quad \theta_{\cdot k} = n^{-1}\sum_{j=1}^{n}\theta_{jk}, \ \theta_{\cdot\cdot} = K^{-1}\sum_{k=1}^{K}\theta_k,$$

$$W = \frac{1}{K}\sum_{k=1}^{K}s_k^2, \qquad s_k^2 = (n-1)^{-1}\sum_{j=1}^{n}(\theta_{jk} - \theta_{\cdot k})^2.$$

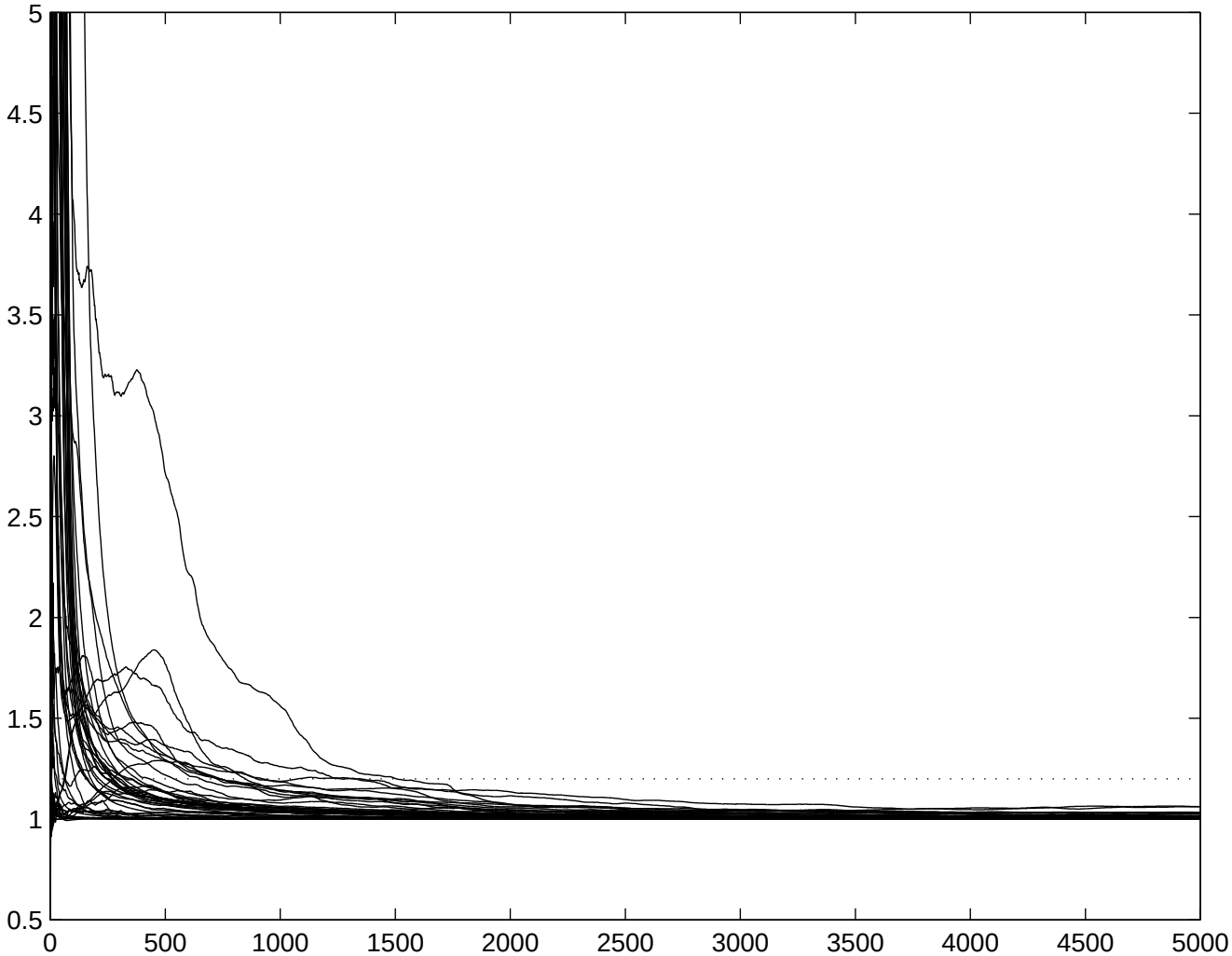
The estimate of $\text{var}(\theta|\mathbf{Y})$, the marginal posterior variance of the estimate, is then obtained by a weighted average of B and W as follows:

$$\widehat{\text{var}}(\theta) = \frac{n-1}{n}W + \frac{1}{n}B.$$

The EPSR is defined as

$$\hat{R}^{1/2} = \left[\widehat{\text{var}}(\theta)/W\right]^{1/2}.$$

As the algorithm converges, $\hat{R}^{1/2}$ should be close to 1.0. In monitoring convergence, all EPSR values for all parameters are computed. Convergence is achieved if all the EPSR values are less than 1.2.


5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000

if we know the corr is low. then set $s=1$.

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; see Albert and Chib (1993). In the numerical illustrations of this book, we use $s = 1$.

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

$$\hat{\boldsymbol{\theta}} = T^{*-1} \sum_{t=1}^{T^*} \boldsymbol{\theta}^{(t)}, \tag{11}$$

$$\widehat{\text{Var}}(\boldsymbol{\theta}|\mathbf{Y}) = (T^* - 1)^{-1} \sum_{t=1}^{T^*} (\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})^T. \tag{12}$$

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.

Let θ_k be the kth element of $\boldsymbol{\theta}$. Based on (12), the positive square root of the kth diagonal element in $\widehat{\text{Var}}(\boldsymbol{\theta}|\mathbf{Y})$ 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 the Bayesian analysis, the issue of hypothesis testing is formulated as a model comparison problem; see Chapter 4.

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:

$$\hat{\boldsymbol{\omega}}_i = T^{*-1} \sum_{t=1}^{T^*} \boldsymbol{\omega}_i^{(t)}, \tag{13}$$

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.

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 SEMs, 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 $\boldsymbol{\omega}_{i0}$. 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).

We illustrate the implementation of MCMC methods through their applications to some SEMs. First, we consider the following linear SEM with fixed covariates. Its measurement equation is given by:

$$\mathbf{y}_i = \mathbf{A}\mathbf{c}_i + \mathbf{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \cdots, n, \tag{14}$$

in which $\mathbf{A}$, $\mathbf{\Lambda}$, $\mathbf{c}_i$, and $\boldsymbol{\omega}_i$ are defined as before, $\boldsymbol{\epsilon}_i$ is a random vector of residual errors with distribution $N[\mathbf{0}, \mathbf{\Psi}_\epsilon]$, where $\mathbf{\Psi}_\epsilon$ is diagonal and $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\omega}_i$. The structural equation is defined as

$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \mathbf{\Pi}\boldsymbol{\eta}_i + \mathbf{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \tag{15}$$

where $\mathbf{B}$, $\mathbf{\Pi}$, $\mathbf{\Gamma}$, $\mathbf{d}_i$, $\boldsymbol{\eta}_i$, and $\boldsymbol{\xi}_i$ are defined as before, $\boldsymbol{\xi}_i$ and $\boldsymbol{\delta}_i$ are independently distributed as $N[\mathbf{0}, \mathbf{\Phi}]$ and $N[\mathbf{0}, \mathbf{\Psi}_\delta]$, respectively, where $\mathbf{\Psi}_\delta$ is a diagonal covariance matrix. Equation (15) can be rewritten as

$$\boldsymbol{\eta}_i = \mathbf{\Lambda}_\omega \mathbf{v}_i + \boldsymbol{\delta}_i, \tag{16}$$

where $\mathbf{\Lambda}_\omega = (\mathbf{B}, \mathbf{\Pi}, \mathbf{\Gamma})$ and $\mathbf{v}_i = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$.

Let

- $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ — data matrix
- $\mathbf{C} = (\mathbf{c}_1, \cdots, \mathbf{c}_n)$ — data matrix
- $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_n)$ — data matrix
- $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n)$ — matrix of latent variables
- $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Lambda}, \mathbf{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Psi}_\delta\} = \{\mathbf{A}, \boldsymbol{\Lambda}, \boldsymbol{\Lambda}_\omega, \boldsymbol{\Phi}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Psi}_\delta\}$.

Our main objective is to use MCMC methods to obtain the Bayesian estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$. A sequence of random observations from the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$ will be generated via the Gibbs sampler. 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.

The 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.

The key idea of the above strategy is data augmentation. We augment $\boldsymbol{\Omega}$ with the observed data $\mathbf{Y}$ and work on the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\Omega}|\mathbf{Y}]$. In Step b of the Gibbs sampler, we need to simulate $\boldsymbol{\theta}$ from $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$. Once $\boldsymbol{\Omega}$ is given rather than random, the SEM becomes the familiar regression model. Consequently, the conditional distribution $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$ can be derived and the implementation of Gibbs sampler is possible.

As an illustration, we present an application of the above strategy to the analysis of nonlinear SEMs with fixed covariates. The model is defined as follows: For $i = 1, \cdots, n$,

$$\mathbf{y}_i = \mathbf{A}\mathbf{c}_i + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i,$$
$$\boldsymbol{\eta}_i = \mathbf{B}\mathbf{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i,$$

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.

Here, $[\boldsymbol{\Omega}|\boldsymbol{\theta}, \mathbf{Y}]$ can be derived based on the distribution of the latent variables and the definition of the model as follows:

$$p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}) p(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \boldsymbol{\theta}) p(\boldsymbol{\xi}_i|\boldsymbol{\theta}). \quad (17)$$

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

$$\exp\left\{ -\frac{1}{2}\boldsymbol{\xi}_i^T \boldsymbol{\Phi}^{-1} \boldsymbol{\xi}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{A}\mathbf{c}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_\epsilon^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{c}_i - \boldsymbol{\Lambda}\boldsymbol{\omega}_i) \right.$$
$$\left. - \frac{1}{2}(\boldsymbol{\eta}_i - \mathbf{B}\mathbf{d}_i - \boldsymbol{\Pi}\boldsymbol{\eta}_i - \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i))^T \boldsymbol{\Psi}_\delta^{-1} (\boldsymbol{\eta}_i - \mathbf{B}\mathbf{d}_i - \boldsymbol{\Pi}\boldsymbol{\eta}_i - \boldsymbol{\Gamma}\mathbf{F}(\boldsymbol{\xi}_i)) \right\}.$$

This distribution is non-standard and complex. The MH algorithm is used to generate observations from the target density $p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta})$ (Appendix 3.2). Once $\boldsymbol{\Omega}$ is given, the nonlinear SEM becomes the familiar regression model. The conditional distribution $[\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}]$ can be easily derived (Appendix 3.4).

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 and the MH algorithm. 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.

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 the following nonlinear SEM with a linear covariate. 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

$$\begin{aligned} \mu^*_{i1} &= \mu_1 + \eta_i, \quad \mu^*_{ij} = \mu_j + \lambda_{j1}\eta_i, \quad j = 2,\ 3, \\ \mu^*_{i4} &= \mu_4 + \xi_{i1}, \ \mu^*_{ij} = \mu_j + \lambda_{j2}\xi_{i1}, \ j = 5,\ 6,\ 7, \\ \mu^*_{i8} &= \mu_8 + \xi_{i2}, \ \mu^*_{ij} = \mu_j + \lambda_{j3}\xi_{i2}, \ j = 9,\ 10, \end{aligned} \tag{18}$$

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

$$\nu_i = b_1 d_i + \gamma_1\xi_{i1} + \gamma_2\xi_{i2} + \gamma_3\xi_{i1}\xi_{i2} + \gamma_4\xi^2_{i1} + \gamma_5\xi^2_{i2}. \tag{19}$$

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:

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

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 (see Equations (6) and (8)):

$$\boldsymbol{\mu}_0 = (0.0, \cdots, 0.0)^T,\ \boldsymbol{\Sigma}_0 = \mathbf{I}_{10},\ \alpha_{0\epsilon k} = \alpha_{0\delta} = 9,\ \beta_{0\epsilon k} = \beta_{0\delta} = 4,$$

elements in $\boldsymbol{\Lambda}_{0k}$ and $\boldsymbol{\Lambda}_{0\omega k}$ are taken to be the true values,

$$\mathbf{H}_{0yk} = \mathbf{I}_{10},\ \mathbf{H}_{0\omega k} = \mathbf{I}_6,\ \rho_0 = 4,\ \mathbf{R}_0 = \boldsymbol{\Phi}_0^{-1},$$

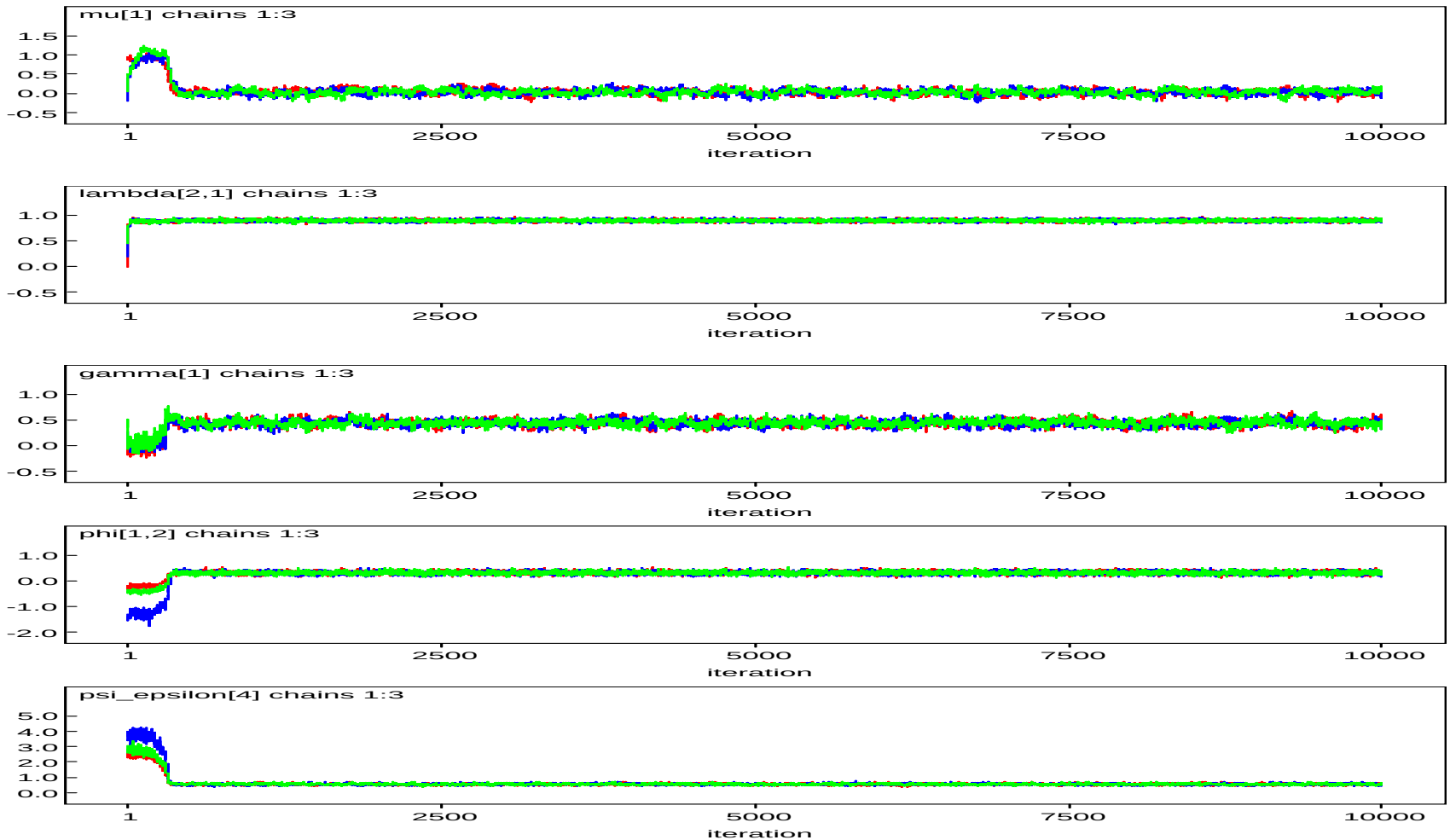
where $\boldsymbol{\Phi}_0$ is the matrix with true values of ϕ_{11}, ϕ_{22}, and ϕ_{12}.

The WinBUGS code is given below (see also Appendix 3.5).

```
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
```

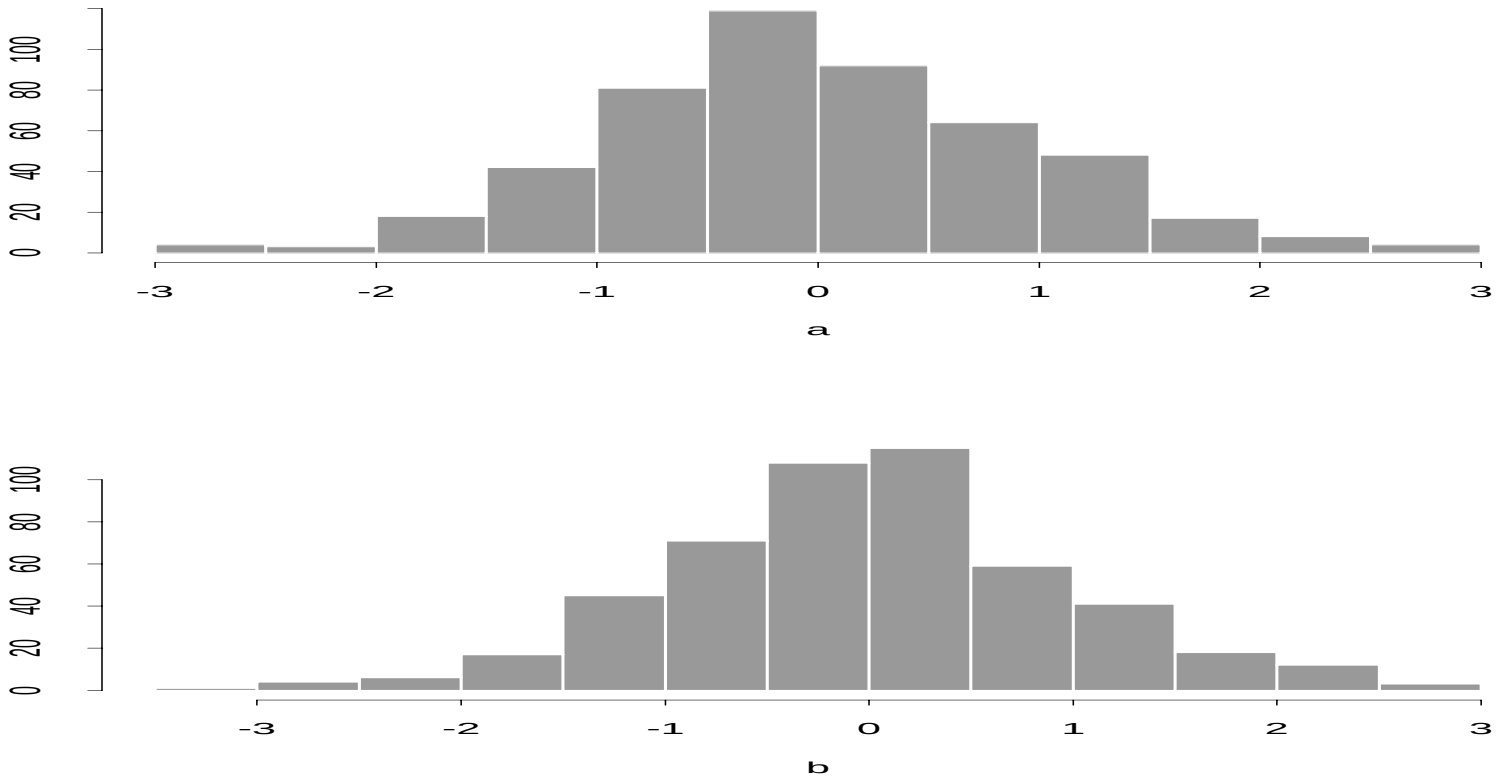
We observed that the WinBUGS program converged in less than 4,000 iterations. Plots of some simulated sequences of observations for monitoring convergence are presented in Figure 3.2.



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.

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

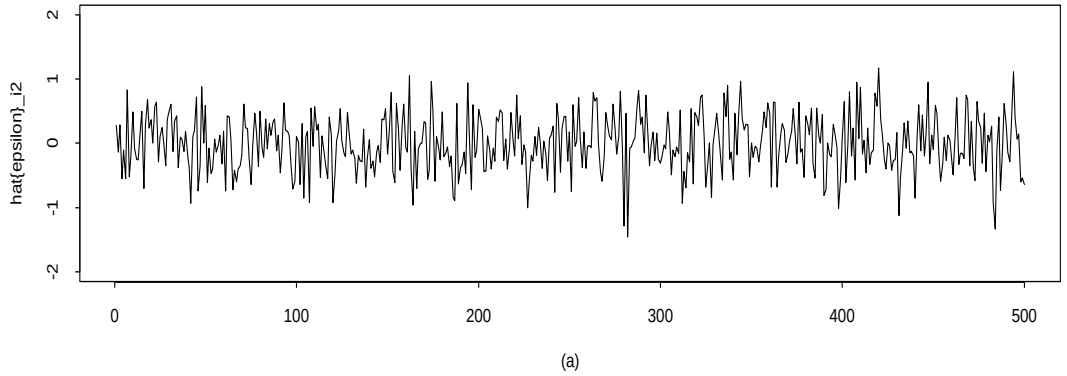
WinBUGS also produces estimates of the latent variables $\{\hat{\boldsymbol{\omega}}_i = (\hat{\eta}_i, \hat{\xi}_{i1}, \hat{\xi}_{i2})^T, 1, \cdots, n\}$. Histograms that correspond to the sets of latent variable estimates $\hat{\xi}_{i1}$ and $\hat{\xi}_{i2}$ are displayed in Figure 3.3. We observe that the empirical distributions are close to the normal. The elements in the sample covariance matrix of $\{\hat{\boldsymbol{\xi}}_i,\ i = 1,\ \cdots, n\}$ are $\{s_{11}, s_{12}, s_{22}\} = \{0.902, 0.311, 0.910\}$, which are close to the true values.



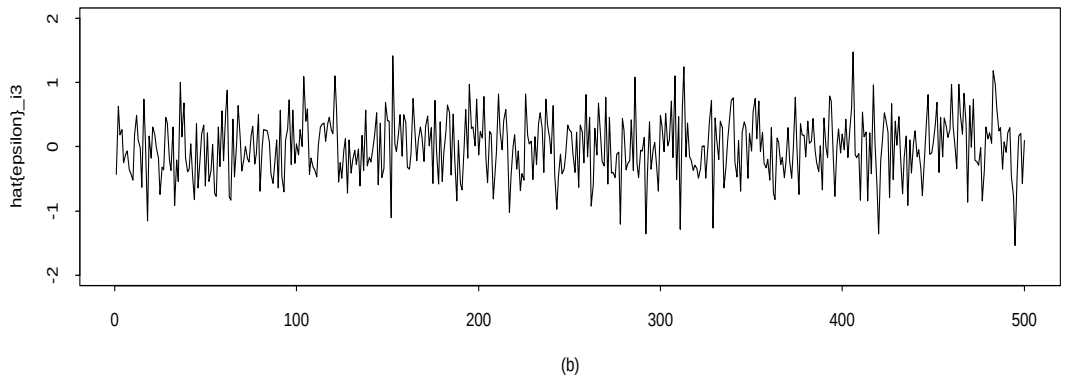
The residuals can be estimated via $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\omega}}_i = (\hat{\eta}_i, \hat{\xi}_{i1}, \hat{\xi}_{i2})^T$ for $i = 1, \cdots, n$ as follows:

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

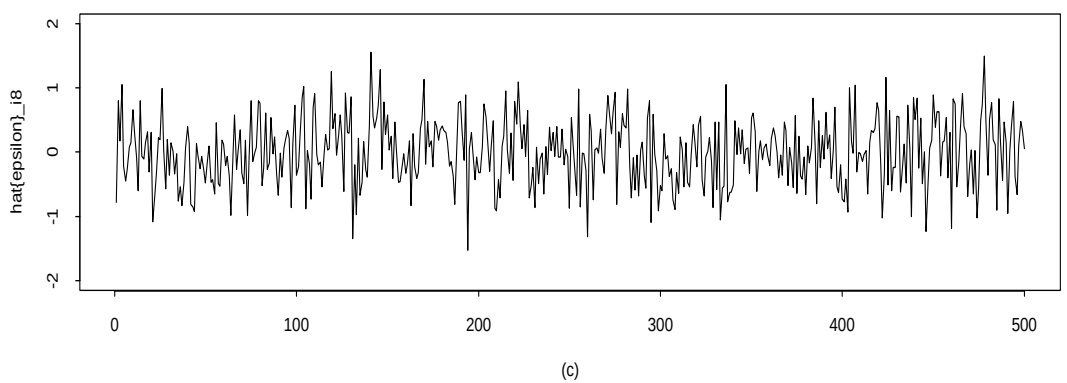
Some estimated residual plots, $\hat{\epsilon}_{i2}$, $\hat{\epsilon}_{i3}$, and $\hat{\epsilon}_{i8}$ against the case number are presented in Figure 3.4. The plots of estimated residuals $\hat{\delta}_i$ versus $\hat{\xi}_{i1}$ and $\hat{\xi}_{i2}$ are presented in Figures 3.5. Other residual plots are similar. 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.


hat{epsilon}_i2
2
1
0
-1
-2
0
100
200
300
400
500

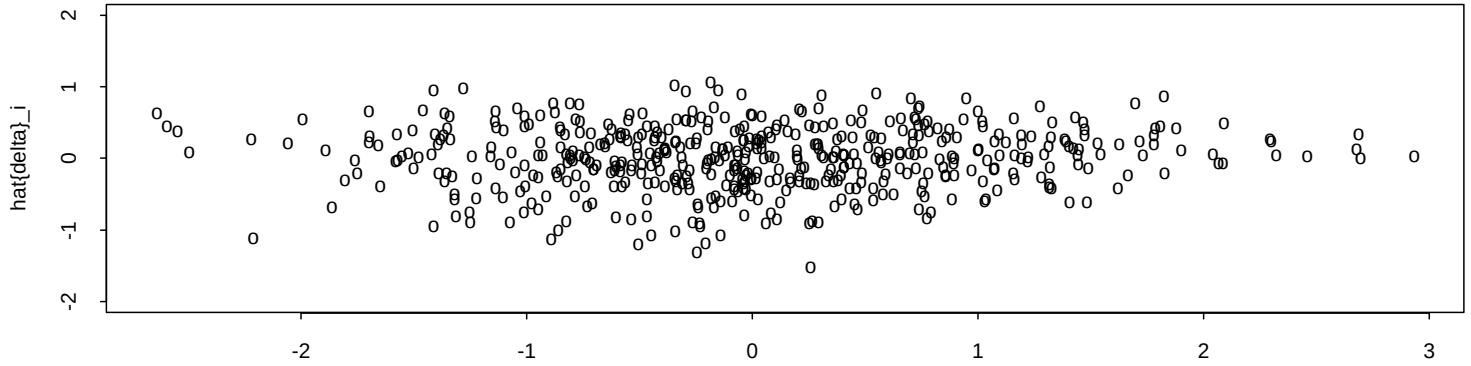
(a)


hat{epsilon}_i3
2
1
0
-1
-2
0
100
200
300
400
500

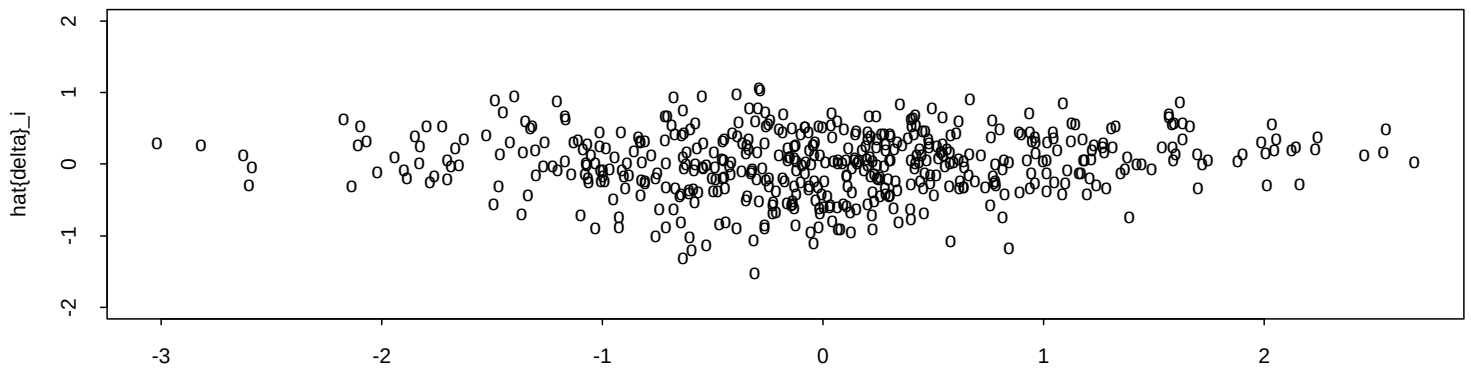
(b)


hat{epsilon}_i8
2
1
0
-1
-2
0
100
200
300
400
500

(c)



(a)



(b)

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 (18), (19), and (5). 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; see Appendices 3.5 and 3.6.