7 Mixture Structural Equation Models

7.1 Introduction

In Chapter 6, we discussed two-level SEMs and multisample SEMs, which are useful tools for analyzing hierarchical data and multiple group data. In this chapter, we focus on another kind of heterogeneous data which involve independent observations that come from one of the K populations with different distributions, and no information is available on which of the K populations an individual observation belongs to. In general, a finite mixture model arises with a population which is a mixture of K components with probability densities $\{f_k,\ k = 1, \cdots, K\}$ and mixing proportions $\{\pi_k,\ k = 1, \cdots, K\}$. This kind of models arise in many fields, including behavioral, medical, and environmental sciences. They have been used in handling outliers (Pettit and Smith, 1985), and density estimation (Roeder and Wasserman, 1997). Statistical analysis of mixture models is not straightforward. For the estimation of mixture models with a fixed number of components K, a variety of methods have been proposed. Examples are the method of moments (Lindsay and Basak, 1993), Bayesian methods with MCMC techniques (Diebolt and Robert, 1994; Robert, 1996), and the ML method (Hathaway, 1985). For mixture models with K treated as random, Richardson and Green (1997) developed a full Bayesian analysis with a reversible jump MCMC method. For the challenging problem of testing the number of components, the classical likelihood-based inference encountered serious difficulties because of some non-regular problems, some standard asymptotic properties associated with the likelihood ratio test are not valid. In contrast, as pointed out by Richardson and Green (1997), the Bayes paradigm is particularly suitable to analyzing mixture models with an unknown K.

In the field of SEM, Jedidi, Jagpal and DeSarbo (1997) considered the estimation of a finite mixtures of SEMs with a fixed number of components, and gave a brief discussion on model selection via the Bayesian Information Criterion (BIC). Yung (1997) investigated finite mixtures of confirmatory factor analysis models, while Dolan and van der Maas (1998) applied a quasi-Newton algorithm to finite mixtures and inferred the estimation by changing the degree of separation and the sample size. Arminger, Stein and Wittenberg (1999) discussed ML analysis for mixtures of conditional mean- and covariance-structure models; and three estimation strategies on the basis of the EM algorithm were established. Zhu and Lee (2001) proposed a Bayesian analysis to finite mixtures in the LISREL model, using the idea of data augmentation and some MCMC methods. Lee and Song (2003) developed a path sampling procedure to compute the observed-data log-likelihood, for evaluating the BIC in selecting the appropriate number of components for a mixture SEM with missing data. A Bayesian approach for analyzing mixtures of SEMs with an unknown number of components has been developed by Lee and Song (2002), based on the Bayes factor computed via a path sampling procedure. Spiegelhalter *et al.* (2003) pointed out that DIC may not be appropriate for model comparison in the context of mixture models. Hence, WinBUGS does not give DIC results for mixture models. In Section 7.3.3, we present a modified DIC for the comparison of mixture SEMs using our tailor-made R code.

For a mixture SEM with K components, the model is invariant with respect to permutation of the labels $k = 1, \cdots, K$. Thus, the model is not identified, and adoption of an unique labeling for identifiability is important. In the literature, a common method is to use some constraints on the components of the mean vector to force a unique labeling. In a Bayesian approach, arbitrary constraints may not be able to solve the problem. We apply the permutation sampler (Frühwirth-Schnatter, 2001) to find the appropriate identifiability constraints.

The objectives of this chapter are to introduce finite mixture SEMs and a modified mixture SEM. The Bayesian methodologies for analyzing heterogeneous data through these models are also discussed. Section 7.2 presents a general finite mixture SEM, in which the probabilities of component memberships, $\pi_1, \cdots, \pi_K$, are unknown and estimated together with other parameters. Section 7.3 extends the general mixture SEM to a modified mixture SEM, in which the probabilities of component memberships are modeled through a multinomial logit model. With this extension, the effects of some important covariates on individuals' component memberships are incorporated to achieve better results. In addition, the modified mixture SEM can accommodate component-specific nonlinear interrelationships among latent variables, as well as nonignorable missing responses and covariates.

7.2 Finite Mixture SEMs

7.2.1 The model

Let $\mathbf{y}_i$ be a $p \times 1$ random vector corresponding to the ith observation, and suppose that its distribution is given by the following probability density function:

$$f(\mathbf{y}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k), \quad i = 1, \cdots, n, \tag{7.1}$$

where K is a given integer, π_k is the unknown mixing proportion such that $\pi_k > 0$ and $\pi_1 + \cdots + \pi_K = 1$, $f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k)$ is the multivariate normal density function with an unknown mean vector $\boldsymbol{\mu}_k$ and a general covariance structure $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k)$ that depends on an unknown parameter vector $\boldsymbol{\theta}_k$. Let $\boldsymbol{\theta}$ be the parameter vector that contains all unknown parameters in π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\theta}_k$, $k = 1, \cdots, K$. For the kth component, the measurement equation of the model is given by:

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{7.2}$$

where $\boldsymbol{\mu}_k$ is the mean vector, $\boldsymbol{\Lambda}_k$ is the $p \times q$ factor loading matrix, $\boldsymbol{\omega}_i$ is a random vector of latent variables, and $\boldsymbol{\epsilon}_i$ is a random vector of residuals which is distributed according to $N[\mathbf{0}, \boldsymbol{\Psi}_k]$, where $\boldsymbol{\Psi}_k$ is a diagonal matrix. It is assumed that $\boldsymbol{\omega}_i$ and $\boldsymbol{\epsilon}_i$ are independent. Let $\boldsymbol{\omega}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$ be a partition of $\boldsymbol{\omega}_i$ into an outcome latent vector $\boldsymbol{\eta}_i$, and an explanatory latent vector $\boldsymbol{\xi}_i$. The structural equation of the model is defined as

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}_k \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_k \boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \tag{7.3}$$

where $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ are $q_1 \times 1$ and $q_2 \times 1$ subvectors of $\boldsymbol{\omega}_i$ respectively, $\boldsymbol{\delta}_i$ is a random vector of residuals that is independent of $\boldsymbol{\xi}_i$, $\boldsymbol{\Pi}_k$ and $\boldsymbol{\Gamma}_k$ are unknown parameter matrices such that $\boldsymbol{\Pi}_{0k}^{-1} = (\mathbf{I} - \boldsymbol{\Pi}_k)^{-1}$ exists, and $|\boldsymbol{\Pi}_{0k}|$ is independent of the elements in $\boldsymbol{\Pi}_k$. Let the distributions of $\boldsymbol{\xi}_i$ and $\boldsymbol{\delta}_i$ in the kth component be $N[\mathbf{0}, \boldsymbol{\Phi}_k]$ and $N[\mathbf{0}, \boldsymbol{\Psi}_{\delta k}]$, respectively, where $\boldsymbol{\Psi}_{\delta k}$ is a diagonal matrix. The parameter vector $\boldsymbol{\theta}_k$ contains the free unknown parameters in $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Pi}_k$, $\boldsymbol{\Gamma}_k$, $\boldsymbol{\Phi}_k$, $\boldsymbol{\Psi}_k$, and $\boldsymbol{\Psi}_{\delta k}$. The covariance structure of $\boldsymbol{\omega}_i$ in the kth component is given by

$$\boldsymbol{\Sigma}_{\omega k} = \begin{bmatrix} \boldsymbol{\Pi}_{0k}^{-1}(\boldsymbol{\Gamma}_k \boldsymbol{\Phi}_k \boldsymbol{\Gamma}_k^T + \boldsymbol{\Psi}_{\delta k})(\boldsymbol{\Pi}_{0k}^{-1})^T & \boldsymbol{\Pi}_{0k}^{-1}\boldsymbol{\Gamma}_k \boldsymbol{\Phi}_k \\ \boldsymbol{\Phi}_k \boldsymbol{\Gamma}_k^T (\boldsymbol{\Pi}_{0k}^{-1})^T & \boldsymbol{\Phi}_k \end{bmatrix},$$

and $\boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) = \boldsymbol{\Lambda}_k \boldsymbol{\Sigma}_{\omega k} \boldsymbol{\Lambda}_k^T + \boldsymbol{\Psi}_k$. Any of these unknown parameter matrices can be invariant across components. However, it is important to assign a different $\boldsymbol{\mu}_k$ in the measurement equation of each component in order to effectively analyze data from the heterogeneous populations that differ by their mean vectors.

As the mixture model defined in (7.1) is invariant with respect to permutation of labels $k = 1, \cdots, K$, adoption of an unique labeling for identifiability is important. Our method is to impose the ordering $\mu_{1,1} < \cdots < \mu_{K,1}$ for solving the label switching problem (jumping between various labeling subspaces), where $\mu_{k,1}$ is the first element of the mean vector $\boldsymbol{\mu}_k$. It works fine if $\mu_{1,1} < \cdots < \mu_{K,1}$ are well separated. However, if $\mu_{1,1} < \cdots < \mu_{K,1}$ are close to each other, it may not be able to eliminate the label switching problem, and may give biased results. Hence, it is important to find an appropriate identifiability constraint. Here, the random permutation sampler that is developed by Frühwirth-Schnatter (2001) will be applied for finding the suitable identifiability constraints. Moreover, for each $k = 1, \cdots K$, structural parameters in the covariance matrix $\mathbf{\Sigma}_k$ corresponding to the model defined by (7.2) and (7.3) are not identified. This problem is solved by the common method in structural equation modeling by fixing appropriate elements in $\mathbf{\Lambda}_k$, $\mathbf{\Pi}_k$, and/or $\mathbf{\Gamma}_k$ at preassigned values that are chosen on a problem-by-problem basis. For clear presentation of the Bayesian method, we assume that all the unknown parameters in the model are identified.

7.2.2 Bayesian Estimation

Let $\boldsymbol{\theta}_{yk}$ be the vector of unknown parameters in $\mathbf{\Lambda}_k$ and $\boldsymbol{\Psi}_k$, and let $\boldsymbol{\theta}_{\omega k}$ be the vector of unknown parameters in $\mathbf{\Pi}_k$, $\mathbf{\Gamma}_k$, $\mathbf{\Phi}_k$, and $\boldsymbol{\Psi}_{\delta k}$. Let $\boldsymbol{\mu}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}_y$, and $\boldsymbol{\theta}_\omega$ be the vectors that contain the unknown parameters in $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$, $\{\pi_1, \cdots, \pi_K\}$, $\{\boldsymbol{\theta}_{y1}, \cdots, \boldsymbol{\theta}_{yK}\}$, and $\{\boldsymbol{\theta}_{\omega 1}, \cdots, \boldsymbol{\theta}_{\omega K}\}$, respectively; and let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega)$ be the overall parameter vector. Inspired by other works in finite mixture models, we introduce a group label w_i for the ith observation $\mathbf{y}_i$ as a latent allocation variable, and assume that it is independently drawn from the following distribution:

$$p(w_i = k) = \pi_k, \qquad \text{for} \quad k = 1, \cdots, K. \tag{7.4}$$

Moreover, let $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ be the observed data matrix, $\mathbf{\Omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n)$ be the matrix of latent vectors, and $\mathbf{W} = (w_1, \cdots, w_n)$ be the vector of allocation variables.

In a standard Bayesian analysis, we need to evaluate the posterior distribution of the unknown parameters, $p[\boldsymbol{\theta}|\mathbf{Y}]$. Due to the nature of the mixture model, this posterior distribution is complicated. However, if $\mathbf{W}$ is observed, the component of every $\mathbf{y}_i$ can be identified and the mixture model becomes the more familiar multiple group model. In addition, if $\mathbf{\Omega}$ is observed, the underlying SEM will become the linear simultaneous equation model which is comparatively easy to handle. Hence, the observed data $\mathbf{Y}$ are augmented with the latent quantities $\mathbf{\Omega}$ and $\mathbf{W}$ in the posterior analysis, and the posterior analysis is based on $p(\boldsymbol{\theta}, \mathbf{\Omega}, \mathbf{W}|\mathbf{Y})$.

The label switching problem has to be solved in the posterior analysis. For general mixture models with K-components, the unconstrained parameter space contains $K!$ subspaces, each one corresponding to a different way to label the states. In the current mixture of SEM, the likelihood is invariant to relabeling the states. If the prior distributions of $\boldsymbol{\pi}$ and other parameters in $\boldsymbol{\theta}$ are also invariant, the unconstrained posterior is invariant to relabeling the states and identical on all labeling subspaces. This induces a multimodal posterior and a serious problem in Bayesian estimation.

We will use the MCMC approach proposed by Frühwirth-Schnatter (2001) to deal with the above label switching problem. In this approach, an unidentified model is first estimated by sampling from the unconstrained posterior using the random permutation sampler, where each sweep is concluded by a random permutation of the current labeling of the components. The random permutation sampler delivers a sample that explores the whole unconstrained parameter space and jumps between various labeling subspaces in a balanced fashion. As pointed out by Frühwirth-Schnatter (2001), although the model is unidentified, the output of the random permutation sampler can be used to estimate unknown parameters that are invariant to relabeling the states, and can be explored to find a suitable identifiability constraints. Then, the model is reestimated by sampling from the posterior distribution under the imposed identifiability constraints, again using the permutation sampler. The implementation of the permutation sampler in relation to the mixture of SEMs, and the method of selecting the identifiability constraint are briefly described in Appendices 7.1 and 7.2, respectively.

The main task in the Bayesian estimation is to simulate a sufficiently large sample of observations from $[\boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{W}|\mathbf{Y}]$. Similar to many Bayesian analyses of SEMs, this is done by the Gibbs sampler (Geman and Geman, 1984) as follows: At the rth iteration with current values $\boldsymbol{\theta}^{(r)}$, $\boldsymbol{\Omega}^{(r)}$, and $\mathbf{W}^{(r)}$:

Step (a): Generate $(\mathbf{W}^{(r+1)}, \boldsymbol{\Omega}^{(r+1)})$ from $p(\boldsymbol{\Omega}, \mathbf{W}|\mathbf{Y}, \boldsymbol{\theta}^{(r)})$;

Step (b): Generate $\boldsymbol{\theta}^{(r+1)}$ from $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}^{(r+1)}, \mathbf{W}^{(r+1)})$;

Step (c): Reorder the label through the permutation sampler to achieve the identifiability.

As $p(\boldsymbol{\Omega}, \mathbf{W}|\mathbf{Y}, \boldsymbol{\theta}) = p(\mathbf{W}|\mathbf{Y}, \boldsymbol{\theta})p(\boldsymbol{\Omega}|\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta})$, Step (a) can be further decomposed into the following two steps:

Step (a1): Generate $\mathbf{W}^{(r+1)}$ from $p(\mathbf{W}|\mathbf{Y}, \boldsymbol{\theta}^{(r)})$;

Step (a2): Generate $\boldsymbol{\Omega}^{(r+1)}$ from $p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}^{(r)}, \mathbf{W}^{(r+1)})$.

Simulating observations $(\mathbf{W}, \boldsymbol{\Omega})$ through Steps (a1) and (a2) is more efficient than using Step (a). Conditional distributions required for implementing the Gibbs sampler are discussed below.

We first consider the conditional distribution associated with Step (a1). As w_i are independent,

$$p(\mathbf{W}|\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(w_i|\mathbf{y}_i, \boldsymbol{\theta}). \tag{7.5}$$

Moreover,

$$p(w_i = k|\mathbf{y}_i, \boldsymbol{\theta}) = \frac{p(w_i = k, \mathbf{y}_i|\boldsymbol{\theta})}{p(\mathbf{y}_i|\boldsymbol{\theta})} = \frac{p(w_i = k|\boldsymbol{\pi})p(\mathbf{y}_i|w_i = k, \boldsymbol{\theta})}{p(\mathbf{y}_i|\boldsymbol{\theta})} = \frac{\pi_k f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k)}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \tag{7.6}$$

where $f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k)$ is the probability density function of $N[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k)]$. Hence, the conditional distribution of $\mathbf{W}$ given $\mathbf{Y}$ and $\boldsymbol{\theta}$ can be derived from (7.5) and (7.6).

Consider the conditional distribution involved in Step (a2). Because $\boldsymbol{\omega}_i$ are mutually independent, we have

$$\begin{aligned}\prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\mathbf{y}_i, w_i, \boldsymbol{\theta}) = p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{W}) &\propto p(\mathbf{Y}|\boldsymbol{\Omega}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\theta}_y)p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega) \\ &= \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\omega}_i, w_i, \boldsymbol{\mu}, \boldsymbol{\theta}_y)p(\boldsymbol{\omega}_i|w_i, \boldsymbol{\theta}_\omega).\end{aligned} \tag{7.7}$$

Let $\mathbf{C}_k = \boldsymbol{\Sigma}_{\omega k}^{-1} + \boldsymbol{\Lambda}_k^T\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k$, where $\boldsymbol{\Sigma}_{\omega k}$ is the covariance matrix of $\boldsymbol{\omega}_i$ in the kth component. Moreover, as the conditional distribution of $\boldsymbol{\omega}_i$ given $\boldsymbol{\theta}_\omega$ and '$w_i = k$' is $N[\mathbf{0}, \boldsymbol{\Sigma}_{wk}]$, and the conditional distribution of $\mathbf{y}_i$ given $\boldsymbol{\omega}_i$, $\boldsymbol{\mu}$, $\boldsymbol{\theta}_y$, and '$w_i = k$' is $N[\boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i, \boldsymbol{\Psi}_k]$, it can be shown (see Lindley and Smith, 1972) that

$$[\boldsymbol{\omega}_i|\mathbf{y}_i, w_i = k, \boldsymbol{\theta}] \overset{D}{=} N[\mathbf{C}_k^{-1}\boldsymbol{\Lambda}_k^T\boldsymbol{\Psi}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k), \mathbf{C}_k^{-1}]. \tag{7.8}$$

The conditional distribution of $p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{W})$ can be obtained from (7.7) and (7.8). Drawing observations from this familiar normal distribution is fast.

We now consider the conditional distribution $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}, \mathbf{W})$ in Step (b) of the Gibbs sampler. This conditional distribution is quite complicated. However, the difficulty can be reduced by assuming the following mild conditions on the prior distribution of $\boldsymbol{\theta}$. We assume that the prior distribution of the mixing proportion $\boldsymbol{\pi}$ is independent of the prior distributions of $\boldsymbol{\mu}$, $\boldsymbol{\theta}_y$, and $\boldsymbol{\theta}_\omega$. Like many Bayesian analyses in SEMs, the prior distribution of the mean vector $\boldsymbol{\mu}$ can be taken to be independent of the prior distributions of the parameters $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\omega$ in the covariance structures. Moreover, when $\boldsymbol{\Omega}$ is given, the parameters in $\boldsymbol{\theta}_{yk} = \{\boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k\}$ are the parameters involved in the linear regression model with the observed variables in $\mathbf{y}$, see (7.2); and the parameters in $\boldsymbol{\theta}_{\omega k} = \{\boldsymbol{\Pi}_k, \boldsymbol{\Gamma}_k, \boldsymbol{\Phi}_k, \boldsymbol{\Psi}_{\delta k}\}$ are the parameters involved in the other simultaneous equation model with the latent variables, see (7.3). Hence, we assume that the prior distributions of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\omega$ are independent. As a result, $p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)$. Moreover, from the definition of the model and the properties of $\mathbf{W}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\theta}$, we have

$p(\mathbf{W}|\boldsymbol{\theta}) = p(\mathbf{W}|\boldsymbol{\pi})$, and $p(\boldsymbol{\Omega}, \mathbf{Y}|\mathbf{W}, \boldsymbol{\theta}) = p(\mathbf{Y}|\boldsymbol{\Omega}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\theta}_y)p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega)$. Hence, the joint conditional distribution of $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega)$ can be expressed as

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{W}, \boldsymbol{\Omega}, \mathbf{Y}) &= p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega|\mathbf{W}, \boldsymbol{\Omega}, \mathbf{Y}) \\ &\propto p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)p(\mathbf{W}, \boldsymbol{\Omega}, \mathbf{Y}|\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)p(\mathbf{W}|\boldsymbol{\theta})p(\boldsymbol{\Omega}, \mathbf{Y}|\boldsymbol{\theta}, \mathbf{W}) \qquad (7.9) \\ &\propto p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)p(\mathbf{W}|\boldsymbol{\pi})p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega)p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\theta}_y) \\ &= [p(\boldsymbol{\pi})p(\mathbf{W}|\boldsymbol{\pi})][p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\theta}_y)][p(\boldsymbol{\theta}_\omega)p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega)] \end{aligned}$$

Using this result, the marginal densities $p(\boldsymbol{\pi}|\cdot)$, $p(\boldsymbol{\mu}, \boldsymbol{\theta}_y|\cdot)$, and $p(\boldsymbol{\theta}_\omega|\cdot)$ can be treated separately.

The prior distribution of $\boldsymbol{\pi}$ is taken as the symmetric Dirichlet distribution, that is, $\boldsymbol{\pi} \overset{D}{=} D(\alpha, \cdots, \alpha)$ with probability density function given by

$$p(\boldsymbol{\pi}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \ \pi_1^\alpha \cdots \pi_K^\alpha,$$

where $\Gamma(\cdot)$ is the Gamma function. Since $p(\mathbf{W}|\boldsymbol{\pi}) \propto \prod_{k=1}^K \pi_k^{n_k}$, it follows from (7.9) that the full conditional distribution of $\boldsymbol{\pi}$ remains Dirichlet in the following form:

$$p(\boldsymbol{\pi}|\cdot) \propto p(\boldsymbol{\pi})p(\mathbf{W}|\boldsymbol{\pi}) \propto \prod_{k=1}^K \pi_k^{n_k+\alpha}, \qquad (7.10)$$

where n_k is the total number of i such that $w_i = k$. Thus, $p(\boldsymbol{\pi}|\cdot)$ is distributed as $D(\alpha + n_1, \cdots, \alpha + n_K)$.

Let $\mathbf{Y}_k$ and $\boldsymbol{\Omega}_k$ be the respective submatrices of $\mathbf{Y}$ and $\boldsymbol{\Omega}$, such that all the ith columns with $w_i \neq k$ are deleted. It is natural to assume that for $k \neq h$, $(\boldsymbol{\mu}_k, \boldsymbol{\theta}_{yk}, \boldsymbol{\theta}_{\omega k})$ and $(\boldsymbol{\mu}_h, \boldsymbol{\theta}_{yh}, \boldsymbol{\theta}_{\omega h})$ are independent. Hence, given $\mathbf{W}$, we have

$$p(\boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega|\mathbf{Y}, \boldsymbol{\Omega}, \mathbf{W}) \propto \prod_{k=1}^K p(\boldsymbol{\mu}_k)p(\boldsymbol{\theta}_{yk})p(\boldsymbol{\theta}_{\omega k})p(\mathbf{Y}_k|\boldsymbol{\Omega}_k, \boldsymbol{\mu}_k, \boldsymbol{\theta}_{yk})p(\boldsymbol{\Omega}_k|\boldsymbol{\theta}_{\omega k}), \qquad (7.11)$$

and we can treat the product in (7.11) separately with each k. When $\mathbf{W}$ is given, the original complicated problem of finite mixtures reduces to a much simpler multisample problem. Here, for brevity, we assume that there are no cross-group constraints, and the analysis can be carried out separately with each individual sample. Situations with cross-group constraints can be handled by the similar manner as a multiple groups analysis.

For mixture models, Roeder and Wasserman (1997) pointed out that using fully non-informative prior distributions may lead to improper posterior distributions. Thus, most existing Bayesian analyses on normal mixtures used conjugate type prior distributions (see, Roeder and Wasserman, 1997). This type of prior distributions for various components of $\boldsymbol{\theta}$ are adopted here. Let $\mathbf{\Lambda}_{\omega k} = (\mathbf{\Pi}_k, \mathbf{\Gamma}_k)$, for $m = 1, \cdots, p$, and $l = 1, \cdots, q_1$, we take

$$\begin{aligned}
&[\mathbf{\Lambda}_{km}|\psi_{km}] \stackrel{D}{=} N[\mathbf{\Lambda}_{0km}, \psi_{km}\mathbf{H}_{0ykm}],\ \psi_{km}^{-1} \stackrel{D}{=} Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}],\\
&[\mathbf{\Lambda}_{\omega kl}|\psi_{\delta kl}] \stackrel{D}{=} N[\mathbf{\Lambda}_{0\omega kl}, \psi_{\delta kl}\mathbf{H}_{0\omega kl}],\ \psi_{\delta kl}^{-1} \stackrel{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}], \qquad (7.12)\\
&\boldsymbol{\mu}_k \stackrel{D}{=} N[\boldsymbol{\mu}_{0k}, \mathbf{\Sigma}_{0k}],\ \mathbf{\Phi}_k^{-1} \stackrel{D}{=} W_{q_2}[\mathbf{R}_{0k}, \rho_{0k}],
\end{aligned}$$

where ψ_{km} and $\psi_{\delta kl}$ are the mth diagonal element of $\mathbf{\Psi}_k$ and the lth diagonal element of $\mathbf{\Psi}_{\delta k}$, respectively; $\mathbf{\Lambda}_{km}^T$ and $\mathbf{\Lambda}_{\omega kl}^T$ are vectors that contain unknown parameters in the mth row of $\mathbf{\Lambda}_k$ and the lth row of $\mathbf{\Lambda}_{\omega k}$, respectively; $\alpha_{0\epsilon k},\ \beta_{0\epsilon k},\ \mathbf{\Lambda}_{0km},\ \alpha_{0\delta k}, \beta_{0\delta k},\ \mathbf{\Lambda}_{0\omega kl},\ \boldsymbol{\mu}_{0k}, \rho_{0k}$, and positive definite matrices $\mathbf{H}_{0ykm},\ \mathbf{H}_{0\omega kl},\ \mathbf{\Sigma}_{0k}$, and $\mathbf{R}_{0k}$ are hyperparameters whose values are assumed given. Moreover, we also assume the mild assumptions that $(\psi_{km}, \mathbf{\Lambda}_{km})$ is independent of $(\psi_{kh}, \mathbf{\Lambda}_{kh})$ for $m \neq h$, and $(\psi_{\delta kl}, \mathbf{\Lambda}_{\omega kl})$ is independent of $(\psi_{\delta kh}, \mathbf{\Lambda}_{\omega kh})$ for $l \neq h$.

Let $\mathbf{\Omega}_1 = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)$, $\mathbf{\Omega}_2 = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)$, and let $\mathbf{\Omega}_{1k}$ and $\mathbf{\Omega}_{2k}$ be the submatrices of $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ respectively such that all the ith columns with $w_i \neq k$ are deleted. It can be shown by similar derivations as in previous chapters that the conditional distributions of components of $\boldsymbol{\theta}_k$ are the following familiar normal, Gamma, and inverted Wishart distributions:

$$[\boldsymbol{\mu}_k|\mathbf{Y}_k, \mathbf{\Omega}_k, \mathbf{\Lambda}_k, \mathbf{\Psi}_k] \stackrel{D}{=} N[(\mathbf{\Sigma}_0^{-1} + n_k\mathbf{\Psi}_k^{-1})^{-1}(n_k\mathbf{\Psi}_k^{-1}\bar{\mathbf{Y}}_k + \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0), (\mathbf{\Sigma}_0^{-1} + n_k\mathbf{\Psi}_k^{-1})^{-1}],$$

$$[\mathbf{\Lambda}_{km}|\mathbf{Y}_k,\mathbf{\Omega}_k,\psi_{km}^{-1}] \stackrel{D}{=} N[\mathbf{a}_{ykm},\psi_{km}\mathbf{A}_{ykm}],$$

$$[\psi_{km}^{-1}|\mathbf{Y}_k,\mathbf{\Omega}_k,\mu_{km}] \stackrel{D}{=} Gamma[n_k/2+\alpha_{0\epsilon k},\beta_{\epsilon km}],$$

$$[\mathbf{\Lambda}_{\omega kl}|\mathbf{Y}_k,\mathbf{\Omega}_k,\psi_{\delta kl}^{-1}] \stackrel{D}{=} N[\mathbf{a}_{\delta kl},\psi_{\delta kl}\mathbf{A}_{\omega kl}],$$

$$[\psi_{\delta kl}^{-1}|\mathbf{Y}_k,\mathbf{\Omega}_k] \stackrel{D}{=} Gamma[n_k/2+\alpha_{0\delta k},\beta_{\delta kl}],$$

$$[\mathbf{\Phi}_k|\mathbf{\Omega}_{2k}] \stackrel{D}{=} IW_{q_2}[(\mathbf{\Omega}_{2k}\mathbf{\Omega}_{2k}^T+\mathbf{R}_0^{-1}),n_k+\rho_0],$$

where $\bar{\mathbf{Y}}_k = \sum_{i:w_i=k}(\mathbf{y}_i - \mathbf{\Lambda}_k\boldsymbol{\omega}_i)/n_k$, with $\sum_{i:w_i=k}$ denotes the summation with respect to those i such that $w_i = k$, and

$$\begin{aligned}
\mathbf{a}_{ykm} =& \mathbf{A}_{ykm}(\mathbf{H}_{0ykm}^{-1}\mathbf{\Lambda}_{0km}+\mathbf{\Omega}_k\tilde{\mathbf{Y}}_{km}),\ \mathbf{A}_{ykm} = (\mathbf{H}_{0ykm}^{-1}+\mathbf{\Omega}_k\mathbf{\Omega}_k^T)^{-1},\\
\beta_{\epsilon km} =& \beta_{0\epsilon k} + [\tilde{\mathbf{Y}}_{km}^T\tilde{\mathbf{Y}}_{km} - \mathbf{a}_{ykm}^T\mathbf{A}_{ykm}^{-1}\mathbf{a}_{ykm} + \mathbf{\Lambda}_{0km}^T\mathbf{H}_{0ykm}^{-1}\mathbf{\Lambda}_{0km}]/2,\\
\mathbf{a}_{\delta kl} =& \mathbf{A}_{\omega kl}(\mathbf{H}_{0\omega kl}^{-1}\mathbf{\Lambda}_{0\omega kl}+\mathbf{\Omega}_k\mathbf{\Omega}_{1kl}),\ \mathbf{A}_{\omega kl} = (\mathbf{H}_{0\omega kl}^{-1}+\mathbf{\Omega}_k\mathbf{\Omega}_k^T)^{-1},\\
\beta_{\delta kl} =& \beta_{0\delta k} + [\mathbf{\Omega}_{1kl}^T\mathbf{\Omega}_{1kl} - \mathbf{a}_{\delta kl}^T\mathbf{A}_{\omega kl}^{-1}\mathbf{a}_{\delta kl} + \mathbf{\Lambda}_{0\omega kl}^T\mathbf{H}_{0\omega kl}^{-1}\mathbf{\Lambda}_{0\omega kl}]/2,
\end{aligned}$$

in which $\tilde{\mathbf{Y}}_{km}^T$ is the mth row of $\tilde{\mathbf{Y}}_k$, and $\tilde{\mathbf{Y}}_k$ is a matrix whose columns are equal to the columns of $\mathbf{Y}_k$ minus $\boldsymbol{\mu}_k$, and $\mathbf{\Omega}_{1kl}^T$ is the lth row of $\mathbf{\Omega}_{1k}$. The computational burden in simulating observations from these conditional distributions is light. The situation with fixed parameters in $\mathbf{\Lambda}_k$ or $\mathbf{\Lambda}_{\omega k}$ can be similarly handled as in Zhu and Lee (2001); see also Chapter 3, Appendix 3.3. The conditional distributions given above are familiar and simple distributions. The computational burden required in simulating observations from these distributions is not heavy.

In addition to their role in facilitating estimation, the allocation variables in $\mathbf{W}$ also form a coherent basis for Bayesian classification of the observations. Classification can be addressed either on a within-sample basis or on a predictive basis. See Lee (2007a, Chapter 12) for more details.

7.2.3 Analysis of an Artificial Example

An important issue in the analysis of mixture SEMs is the separation of the components. Yung (1997), and Dolan and van der Maas (1998) pointed out that some of their statistical results cannot be trusted when the separation is poor. Yung (1997) considered the following measure of separation: $d_{kh} = \max_{l \in \{k,h\}} \{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_l^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_h)\}^{1/2}$ and suggested that d_{kh} should be about 3.8 or over. In view of this, an objective of this artificial example is to investigate the performance of the Bayesian approach in analyzing a mixture of SEMs with two not well-separated components. Another objective is to demonstrate the random permutation sampler for finding suitable identifiability constraints. Random observations are simulated from a mixture SEM with two components defined by (7.1), (7.2), and (7.3). The model for each $k = 1,\ 2$ involves nine observed variables that are indicators of three latent variables η, ξ_1, and ξ_2. The structure of the factor loading matrix in each component is

$$\boldsymbol{\Lambda}_k^T = \begin{bmatrix} 1 & \lambda_{k,21} & \lambda_{k,31} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \lambda_{k,52} & \lambda_{k,62} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{k,83} & \lambda_{k,93} \end{bmatrix},$$

where 1's and 0's are fixed parameters for achieving an identified model. In the kth component, the structural equation is given by: $\eta = \gamma_{k,1}\xi_1 + \gamma_{k,2}\xi_2 + \delta$. The true population values of the unknown parameters are given by: $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0)^T$, $\boldsymbol{\mu}_2 = (0.0, 0.0, 0.0, 0.5, 1.5, 0.0,\ 1.0, 1.0, 1.0)^T$, $\lambda_{1,21} = \lambda_{1,31} = \lambda_{1,83} = \lambda_{1,93} = 0.4$, $\lambda_{1,52} = \lambda_{1,62} = 0.8$, $\lambda_{2,21} = \lambda_{2,31} = \lambda_{2,83} = \lambda_{2,93} = 0.8$, $\lambda_{2,52} = \lambda_{2,62} = 0.4$, $\gamma_{1,1} = 0.2$, $\gamma_{1,2} = 0.7$, $\gamma_{2,1} = 0.7$, $\gamma_{2,2} = 0.2$, $\phi_{1,11} = \phi_{1,22} = \phi_{2,11} = \phi_{2,22} = 1.0$, $\phi_{1,12} = \phi_{2,12} = 0.3$, $\psi_{11} = \cdots = \psi_{19} = \psi_{21} = \cdots = \psi_{29} = \psi_{\delta 11} = \psi_{\delta 21} = 0.5$. In this 2-component mixture SEM, the total number of unknown parameters is 62. The separation d_{12} is equal to 2.56, which is less than the suggested value in Yung (1997).

Based on the above settings, we simulate 400 observations from each component, and the total sample size is 800. We focus on $\boldsymbol{\mu}_1$ (or $\boldsymbol{\mu}_2$) in finding a suitable identifiability constraint. The first step is to apply the random permutation sampler to produce an MCMC sample from the unconstrained posterior with size 5,000 after a burn-in phase of 500 simulations. This random permutation sampler delivers a sample that explores a whole unconstrained parameter space and jumps between the various labeling subspaces in a balanced fashion. For a mixture of SEMs with two components, we have only 2! labeling subspaces. In the random permutation sampler, after each sweep the first state (1s) and the second state (2s) are permuted randomly; that is, with probability 0.5, the 1s stay as 1s, and with probability 0.5 they become 2s. The output can be explored to find a suitable identifiability constraint. Based on the reasoning given in Appendix 7.2, it suffices to consider the parameters in $\boldsymbol{\mu}_1$. To search for an appropriate identifiability constraint, we look at the scatterplots of $\mu_{1,1}$ versus $\mu_{1,l}$, $l = 2, \cdots, 9$, for getting information on aspects of the states that are most different. These scatterplots are presented in Figure 7.1, which clearly indicates that the most two significant differences between the two components are sampled values corresponding to $\mu_{1,5}$ and $\mu_{1,6}$. If permutation sampling is based on the constraint $\mu_{1,5} < \mu_{2,5}$ or $\mu_{1,6} > \mu_{2,6}$, the label switching will not appear.

Bayesian estimates are obtained using the permutation sampler with the identifiability constraint $\mu_{1,5} < \mu_{2,5}$. Values of hyperparameters in the conjugate prior distributions (see (7.12)) are taken as follows. For $m = 1, \cdots, 9$, $\mu_{0,m}$ equals to the sample mean $\bar{y}_m$, $\mathbf{\Sigma}_0 = 10^2\mathbf{I}$, elements in $\mathbf{\Lambda}_{0km}$, and $\mathbf{\Lambda}_{0\omega kl}$ (which only involves the γ's) are taken to be true parameter values, $\mathbf{H}_{0ykm}$ and $\mathbf{H}_{0\omega kl}$ are the identity matrices, $\alpha_{0\epsilon k} = \alpha_{0\delta k} = 10$, $\beta_{0\epsilon k} = \beta_{0\delta k} = 8$, $\rho_0 = 6$, and $\mathbf{R}_0^{-1} = 5\mathbf{I}$. The α in the Dirichlet distribution of $\boldsymbol{\pi}$ is taken as 1. A few test runs show that the algorithm converges in less than 500 iterations. Hence, Bayesian estimates are obtained using a burn-in phase of 500 iterations and a total of

2,000 observations collected after the burn-in phase. Results are reported in Table 7.1. We observe that the Bayesian estimates are pretty close to their true parameter values.

Figure 7.1 and Table 7.1 here

This artificial data set has also been analyzed by using WinBUGS (Spiegelhalter, *et al.*, 2003). First, an initial Bayesian estimation was conducted without the identifiability constraints to determine the appropriate identifiability constraint $\mu_{1,5} < \mu_{2,5}$ from the output as before. The model is then reestimated with this identifiability constraint, and three starting values of the parameters that are obtained from the sample mean, the 5th, and 95th percentile of the corresponding simulated samples. Bayesian estimates are obtained using the permutation samples with the identifiability constraint and the hyperparameter values given above. Results are presented in Table 7.2. The WinBUGS code under the identifiability constraint $\mu_{1,5} < \mu_{2,5}$, and the data are respectively given in the following websites:

http://www.sta.cuhk.edu.hk/song-lee/book-chapter7(section7.2.3)/WinBUGS-code.
http://www.sta.cuhk.edu.hk/song-lee/book-chapter7(section7.2.3)/WinBUGS-data.
(PLEASE CHANGE TO WEB-SITES HOUSED IN JOHN-WILEY).

Table 7.2 here

7.2.4 An Example on 'Job' and 'Home life'

A small portion of the ICPSR data set collected in the project WORLD VALUES SURVEY 1981-1984 AND 1990-1993 (World Values Study Group, ICPSR Version) is analyzed. The whole data set was collected in 45 societies around the world on broad topics such as work, the meaning and purpose of life, family life, and contemporary social issues. As an illustration, only the data obtained from the United Kingdom with a sample size 1,484 are used. Eight variables in the original data set (variables 116, 117, 180, 132, 96, 255, 254, and 252) that are related to respondents' job and home life are taken as observed variables in $\mathbf{y}$. A description of these variables in the questionnaire is given in Appendix 1.1; see also Lee, 2007a, Section 12.4.3. For simplicity, we delete the cases with missing data, and obtain a sample of size 824. The data set is analyzed with a mixture SEM with two components. In each component, three latent variables, which can be roughly interpreted as 'job satisfaction, η', 'home life, ξ_1', and 'job attitude, ξ_2', are considered. For $k = 1,\ 2$, the specification of the parameter matrices in the model formulation are given by: $\mathbf{\Pi}_k = \mathbf{0}$, $\mathbf{\Psi}_{\delta k} = \psi_{\delta k}$, $\mathbf{\Gamma}_k = (\gamma_{k,1}, \gamma_{k,2})$,

$$\mathbf{\Lambda}_k^T = \begin{bmatrix} 1 & \lambda_{k,21} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{k,42} & \lambda_{k,52} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{k,73} & \lambda_{k,83} \end{bmatrix}, \quad \mathbf{\Phi}_k = \begin{bmatrix} \phi_{k,11} & \phi_{k,12} \\ \phi_{k,21} & \phi_{k,22} \end{bmatrix}, \qquad (7.13)$$

and $\mathbf{\Psi}_k = \text{diag}(\psi_{k1}, \cdots, \psi_{k8})$. To identify the model, elements 1's and 0's in $\mathbf{\Lambda}_k$ are fixed. The total number of unknown parameters is 56.

Bayesian estimates of the structural parameters and the factor scores are obtained via the Gibbs sampler. The following hyperparameters are used: $\alpha = 1$, $\boldsymbol{\mu}_0 = \bar{\mathbf{y}}$, $\mathbf{\Sigma}_0 = \mathbf{S}_y^2/2.0$, $\rho_0 = 5$, and $\mathbf{R}_0^{-1} = 5\mathbf{I}_2$, where $\bar{\mathbf{y}}$ and $\mathbf{S}_y^2$ are the sample mean and sample covariance matrix calculated using the observed data; $\alpha_{0\epsilon k} = \alpha_{0\delta k} = \beta_{0\epsilon k} = \beta_{0\delta k} = 6$ for all k; $\mathbf{H}_{0ykm} = \mathbf{I}$, $\mathbf{H}_{0\omega kl} = \mathbf{I}$; $\mathbf{\Lambda}_{0km} = \tilde{\mathbf{\Lambda}}_{0km}$, and $\mathbf{\Lambda}_{0\omega kl} = \tilde{\mathbf{\Lambda}}_{0\omega kl}$ for all k, m, and l, where $\tilde{\mathbf{\Lambda}}_{0km}$ and $\tilde{\mathbf{\Lambda}}_{0\omega kl}$ are the initial estimates of $\mathbf{\Lambda}_{0km}$ and $\mathbf{\Lambda}_{0\omega kl}$ obtained using noninformative prior distributions. We first use MCMC samples simulated by the random permutation sampler to find an identifiability constraint, and observe that $\mu_{1,1} < \mu_{2,1}$ is a suitable one. Based on different starting values of the parameters, three parallel sequences of observations are generated and the EPSR values are calculated. We observe that the Gibbs sampler algorithm converges within 1,000 iterations. After the convergence of the Gibbs sampler, a total of 1,000 observations with a spacing of 10 are collected for analysis. The Bayesian estimates of the structural parameters and their standard error estimates are reported in Table 7.3. From this table, it can be seen that there are at least two components which have different sets of Bayesian parameter estimates.

Table 7.3 here

7.2.5 Bayesian Model Comparison of Mixture SEMs

The objective of this subsection is to consider the Bayesian model selection problem in selecting one of the two mixtures of SEMs with different number of components. An approach based on the Bayes factor (Kass and Raftery, 1995), together with the path sampling procedure, will be introduced. The underlying finite mixture of SEMs is defined again by equations (7.1), (7.2), and (7.3), except that K is not fixed and π_k are non-negative component probabilities that sum to 1.0. Note that some π_k may be equal to zero. When using the likelihood ratio test, some unknown parameters may be on the boundary of the parameter space, and this causes serious difficulty in developing the test statistics for the hypothesis testing of the number of components. On the contrary, the Bayesian approach for model comparison with the Bayes factor does not have this problem.

Let M_1 be a mixture SEM with K components, and M_0 be a mixture SEM with c components, where $c < K$. The Bayes factor for selection between M_0 and M_1 is defined by

$$B_{10} = \frac{P(\mathbf{Y}|M_1)}{P(\mathbf{Y}|M_0)}. \tag{7.14}$$

In computing the Bayes factor through path sampling, the observed data set $\mathbf{Y}$ is augmented with the matrix of latent variables $\mathbf{\Omega}$. Based on similar reasoning and derivation as given in previous chapters, $\log B_{10}$ can be estimated as follows. Let

$$U(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}, t) = \frac{d}{dt} \log\{p(\mathbf{Y}, \mathbf{\Omega}|\boldsymbol{\theta}, t)\},$$

where $p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}, t)$ is the complete-data likelihood function. Then,

$$\widehat{\log B}_{10} = \frac{1}{2}\sum_{s=0}^{S}(t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}), \tag{7.15}$$

where $t_{(s)}$ are fixed grids in [0, 1] such that $0 = t_{(0)} < t_{(1)} < t_{(2)} < \cdots < t_{(S)} < t_{(S+1)} = 1$, and

$$\bar{U}_{(s)} = J^{-1}\sum_{j=1}^{J} U(\mathbf{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}), \tag{7.16}$$

in which $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ are simulated observations drawn from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$.

The implementation of path sampling is straightforward. Drawing observations $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}) :\ j = 1, \cdots, J\}$ from the posterior distribution $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$ is the major task in the proposed procedure. Similar to estimation, we further utilize the idea of data augmentation to augment the observed data $\mathbf{Y}$ with the latent matrix $\mathbf{W}$ of the allocation variables. As a consequence, the Gibbs sampler described in Section 7.2.2 can be applied to simulate observations from $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t_{(s)})$. In computing Bayes factors, the inclusion of an identifiability constraint in simulating $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ is not necessary. As the likelihood is invariant to relabeling the states, the inclusion of such a constraint will not change the values of $U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)$. As a result, the logarithm Bayes factor estimated through (7.15) and (7.16) is not changed.

Finding a good path to link competing models M_1 and M_0 is important in applying path sampling. An illustrative example is given as follows. Consider the following competing models:

$$M_1 :\quad \mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\pi} \overset{D}{=} \sum_{k=1}^{K} \pi_k f_k(\mathbf{Y}|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k), \tag{7.17}$$

corresponding to a model of K components with positive component probabilities π_k, and

$$M_0 :\quad \mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\pi}^* \overset{D}{=} \sum_{k=1}^{c} \pi_k^* f_k(\mathbf{Y}|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k), \tag{7.18}$$

corresponding to a model of c components with positive component probabilities π_k^*, where $1 \le c < K$. Clearly, these competing models are linked up by a path $M_t : t \in [0, 1]$

as follows:

$$
\begin{aligned}
M_t: \quad [\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\pi}, t] \overset{D}{=} & [\pi_1 + (1-t)a_1(\pi_{c+1} + \cdots + \pi_K)] f_1(\mathbf{y}|\boldsymbol{\mu}_1, \boldsymbol{\theta}_1) + \cdots \\
& + [\pi_c + (1-t)a_c(\pi_{c+1} + \cdots + \pi_K)] f_c(\mathbf{y}|\boldsymbol{\mu}_c, \boldsymbol{\theta}_c) \\
& + t\pi_{c+1} f_{c+1}(\mathbf{y}|\boldsymbol{\mu}_{c+1}, \boldsymbol{\theta}_{c+1}) + \cdots + t\pi_K f_K(\mathbf{y}|\boldsymbol{\mu}_K, \boldsymbol{\theta}_K),
\end{aligned} \tag{7.19}
$$

where $a_1, \cdots, a_c$ are given positive weights such that $a_1 + \cdots + a_c = 1$. When $t = 1$, M_t reduces to M_1; and when $t = 0$, M_t reduces to M_0 with $\pi_k^* = \pi_k + a_k(\pi_{c+1} + \cdots + \pi_K)$, $k = 1, \cdots, c$. The weights $a_1, \cdots, a_c$ represent the increases of the corresponding component probabilities from a K-component SEM to a c-component SEM. A natural and simple suggestion for practical applications is to take $a_k = c^{-1}$.

The complete-data log-likelihood function can be written as

$$
\begin{aligned}
\log p(\mathbf{Y}, \boldsymbol{\Omega}|\boldsymbol{\theta}, t) = & \sum_{i=1}^n \log \Bigg\{ \sum_{k=1}^c [\pi_k + (1-t)a_k \sum_{h=c+1}^K \pi_h] \\
& \times f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k) + \sum_{k=c+1}^K t\pi_k f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k) \Bigg\}.
\end{aligned} \tag{7.20}
$$

By differentiation with respect to t, we have

$$
\begin{aligned}
& U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t) = \\
& \sum_{i=1}^n \frac{-\sum\limits_{h=c+1}^K \pi_h \sum\limits_{k=1}^c a_k f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k) + \sum\limits_{k=c+1}^K \pi_k f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k)}{\sum\limits_{k=1}^c \left[\pi_k + (1-t)a_k \sum\limits_{h=c+1}^K \pi_h\right] f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k) + \sum\limits_{k=c+1}^K t\pi_k f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k)},
\end{aligned} \tag{7.21}
$$

where

$$
\begin{aligned}
f_k(\mathbf{y}_i, \boldsymbol{\omega}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k) = & (2\pi)^{-p/2} |\boldsymbol{\Psi}_k|^{-1/2} \\
& \times \exp\left[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)\right] \\
& \times (2\pi)^{-q_1/2} |\mathbf{I}_{q_1} - \boldsymbol{\Pi}_k| |\boldsymbol{\Psi}_{\delta k}|^{-1/2} \\
& \times \exp\left[-\frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_{\delta k}^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\boldsymbol{\omega}_i)\right] \\
& \times (2\pi)^{-q_2/2} |\boldsymbol{\Phi}_k|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\xi}_i^T \boldsymbol{\Phi}_k^{-1} \boldsymbol{\xi}_i\right),
\end{aligned}
$$

and $\mathbf{\Lambda}_{\omega k} = (\mathbf{\Pi}_k, \mathbf{\Gamma}_k)$. Thus, the Bayes factor can be estimated with

$$\bar{U}_{(s)} = J^{-1} \sum_{j=1}^{J} U(\mathbf{Y}, \mathbf{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}),$$

where $\{(\mathbf{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}) : \ j = 1, \cdots, J\}$ are observations drawn from $p(\mathbf{\Omega}, \boldsymbol{\theta} | \mathbf{Y}, t_{(s)})$.

7.2.6 An Illustrative Example

The same portion of the ICPSR data set as described in Section 7.2.4 is reanalyzed to illustrate the path sampling procedure. We wish to find out whether there are some mixture models that are better than the mixture SEM with two components presented in Section 7.2.4. For each component, the specification of the model is the same as before. However, the number of components, K, is not fixed.

The hyperparameter values are assigned as follows. First $\alpha = 1$; $\boldsymbol{\mu}_0 = \bar{\mathbf{y}}$, $\mathbf{\Sigma}_0 = \mathbf{S}_y^2/2$, where $\bar{\mathbf{y}}$ and $\mathbf{S}_y^2$ are the sample mean and sample covariance matrix calculated using the observed data; $\rho_0 = 6$ and $\mathbf{R}_0^{-1} = 5\mathbf{I}$; $\mathbf{H}_{0ykm} = \mathbf{I}$ and $\mathbf{H}_{0\omega kl} = \mathbf{I}$ are selected for each k, $m = 1, \cdots, p$, and $l = 1, \cdots, q_1$. Moreover, $\{\alpha_{0\epsilon k}, \beta_{0\epsilon k}\}$, and $\{\alpha_{0\delta k}, \beta_{0\delta k}\}$ are selected such that the means and standard deviations of the prior distributions associated with ψ_{km} and $\psi_{\delta kl}$ are equal to 5. Finally, we take $\mathbf{\Lambda}_{0km} = \tilde{\mathbf{\Lambda}}_{0km}$, $\mathbf{\Lambda}_{0\omega kl} = \tilde{\mathbf{\Lambda}}_{0\omega kl}$ for all k, m, and l, where $\tilde{\mathbf{\Lambda}}_{0km}$ and $\tilde{\mathbf{\Lambda}}_{0\omega kl}$ are the corresponding Bayesian estimates obtained through a single component model with noninformative prior distributions. Again, for each $t_{(s)}$, the convergence of the Gibbs sampler algorithm is monitored by parallel sequences of the generated observations from very different starting values. We observe that the algorithm converges quickly within 200 iterations. A total of $J = 1,000$ additional observations are collected after a burn-in phase of 200 iterations for computing $\bar{U}_{(s)}$ in (7.16), and then the logarithms of Bayes factors are estimated via (7.15), using 20 fixed grids in [0,1]. Let M_k denotes the mixture model with k components, the estimated logarithms of Bayes factors are equal to: $\widehat{\log B}_{21} = 75.055$, $\widehat{\log B}_{32} = 4.381$, $\widehat{\log B}_{43} = -0.824$, and $\widehat{\log B}_{53} = -1.395$. Based on the criterion of the logarithm Bayes factor, the one-component model is significantly worse than the two-component model which is significantly worse than the three-component model; while the three-component model is almost as good as the more complicated four-component and five-component models. Hence, a mixture model with three components should be chosen. Although the two-component model suggested in Section 7.2.4 is a plausible model, it does not give as strong support of evidence as the three-component model.

In estimation, based on the MCMC samples simulated by the random permutation sampler, we find $\mu_{1,1} < \mu_{2,1} < \mu_{3,1}$ is a suitable identifiability constraint. Bayesian estimates of the selected three-component mixture model obtained under the constraint $\mu_{1,1} < \mu_{2,1} < \mu_{3,1}$ are presented in Table 7.4, together with the corresponding standard error estimates. For parameters directly associated with observed variables y_1 to y_5, we observe from this table that their Bayesian estimates under component 2 are close to those under component 3, but these estimates are quite different from those under component 1. In contrast, for parameters directly associated with observed variables y_6 to y_8, estimates under component 1 are close to those under component 3, but are quite different from those under component 2. Hence, it is reasonable to select a three-component model for this data set. We estimate the separations of these components and find that they are equal to $d_{12} = 2.257$, $d_{13} = 2.590$, and $d_{23} = 2.473$. These results indicate that the introduced procedure is able to select the appropriate three-component model whose components are not well separated.

Table 7.4 here

7.3 A Modified Mixture SEM

In the analysis of mixture SEMs, it is of interest to examine the effects of covariates on the probability of component membership, which plays an extremely important role in a mixture model. This kind of modeling has been discussed in latent class models and latent growth mixture models (Muthén and Shedden, 1999; Elliott *et al.*, 2005; Guo, Wall and Amemiya, 2006). Another important issue in the analysis of mixture SEMs is related to missing data. Lee and Song (2003) pointed out that the case deletion method which deleted observations with missing entries would produce less accurate estimation results. Specifically, the bias and the root mean square values of parameter estimates and their true population values are larger compared with those obtained by the methods that include observations with missing entries. Lee (2007b) further demonstrated that the selection of the number of components in mixture SEMs would result in misleading conclusions if observations with missing entries are ignored. Moreover, the assumption of missing at random (MAR) may not be realistic for heterogeneous data because the probability of missingness for an individual may highly depend on its associated component with some special characteristics. Here, we present statistical methods to handle missing responses and covariates with a nonignorable missing mechanism.

In this section, we introduce a modified mixture SEM which extends the previous mixture SEMs in three aspects. First, a multinomial logit model with covariates is incorporated to predict the unknown component membership. Second, a nonlinear structural equation is introduced in each component to capture the component-specific nonlinear effects of explanatory latent variables and covariates on outcome latent variables. Third, nonignorable missing data are considered for both responses and covariates. Again, Bayesian methods are used to conduct the analysis. The methodologies are illustrated through a longitudinal study of polydrug use conducted in five California counties in 2004.

7.3.1 Model Description

The modified mixture SEM for a $p \times 1$ random vector $\mathbf{y}_i,\ i = 1, \cdots, n$ is defined as follows (Cai, Song and Hser, 2010):

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_{ik} f_k(\mathbf{y}_i|\boldsymbol{\theta}_k), \quad i = 1, \cdots, n, \tag{7.22}$$

where K is the number of mixture components, π_{ik} is the subject probability of component membership for $\mathbf{y}_i$ such that $\sum_{k=1}^{K} \pi_{ik} = 1$, and $f_k(\cdot)$ is the density function of $\mathbf{y}_i$ with a parameter vector $\boldsymbol{\theta}_k$. The component probabilities of $\mathbf{y}_i$'s, π_{ik}, are further related to an $m_1 \times 1$ vector of covariates $\mathbf{x}_i$ via the following multinomial logit model: For $k = 1, \cdots, K$,

$$\pi_{ik} = p(z_i = k|\mathbf{x}_i) = \frac{\exp\{\boldsymbol{\tau}_k^T \mathbf{x}_i\}}{\sum_{j=1}^{K} \exp\{\boldsymbol{\tau}_j^T \mathbf{x}_i\}}, \tag{7.23}$$

where z_i is a latent allocation variable of $\mathbf{y}_i$, $\boldsymbol{\tau}_k(m_1 \times 1)$ is an unknown vector of coefficients, and the elements in $\boldsymbol{\tau}_K$ are fixed at zeros for identification purpose. The elements of $\boldsymbol{\tau}_k$ carry the information about the probability of component membership present in $\mathbf{x}_i$, which includes covariates that may come from continuous or discrete distributions with a parameter vector $\boldsymbol{\tau}_x$. In contrast to conventional mixture models, the subject probability of unknown component membership in the current mixture SEM is modeled by incorporating explanatory covariates $\mathbf{x}_i$ through a multinomial logit model. Alternative terms in $\mathbf{x}_i$ can be concomitant variables, grouping variables, external variables, and exogenous variables. In substantive research, $\mathbf{x}_i$ and $\mathbf{y}_i$ may or may not have common variables. For example, $\mathbf{x}_i$ can be considered as the same set or a subset of $\mathbf{y}_i$ to help explain differences in component-specific parameters (Muthén and Shedden, 1999; Guo, Wall and Amemiya, 2006). In contrast, Elliott *et al.* (2005) considered $\mathbf{x}_i$ to include an intercept and an indicator of baseline depression, which were excluded in $\mathbf{y}_i$. The main purpose of Equation (7.23) is to provide an improved model with some covariates for predicting unknown component probabilities.

In order to group observed variables in $\mathbf{y}_i$ into latent factors, the measurement equation is defined as follows. Conditional on the kth component,

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{7.24}$$

where $\boldsymbol{\mu}_k$ is a $p \times 1$ intercept vector, $\boldsymbol{\Lambda}_k$ is a $p \times q$ factor loading matrix, $\boldsymbol{\omega}_i$ is a $q \times 1$ random vector of latent variables, $\boldsymbol{\epsilon}_i$ is a $p \times 1$ random vector of error measurements with distribution $N[\mathbf{0}, \boldsymbol{\Psi}_k]$, $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\omega}_i$, and $\boldsymbol{\Psi}_k$ is a diagonal matrix. Furthermore, we consider a partition of $\boldsymbol{\omega}_i$ into a $q_1 \times 1$ outcome latent vector $\boldsymbol{\eta}_i$ and a $q_2 \times 1$ explanatory latent vector $\boldsymbol{\xi}_i$. To assess the interrelationships among $\boldsymbol{\eta}_i$, $\boldsymbol{\xi}_i$, and some fixed covariates, a general structural equation is defined as follows:

$$\boldsymbol{\eta}_i = \mathbf{B}_k \mathbf{d}_i + \boldsymbol{\Pi}_k \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_k \mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{7.25}$$

where $\mathbf{d}_i$ is an $m_2 \times 1$ vector of fixed covariates, conditional on the kth component which may come from continuous or discrete distributions with a parameter vector $\boldsymbol{\tau}_{kd}$; $\mathbf{F}(\boldsymbol{\xi}_i) = (f_1(\boldsymbol{\xi}_i), \cdots, f_r(\boldsymbol{\xi}_i))^T$ $(r \geq q_2)$ is a vector of general differentiable functions $f_1, \cdots, f_r$ that are linearly independent; $\mathbf{B}_k(q_1 \times m_2)$, $\boldsymbol{\Pi}_k(q_1 \times q_1)$, and $\boldsymbol{\Gamma}_k(q_1 \times r)$ are unknown parameter matrices; and $\boldsymbol{\xi}_i$ and $\boldsymbol{\delta}_i$ are independently distributed as $N[\mathbf{0}, \boldsymbol{\Phi}_k]$ and $N[\mathbf{0}, \boldsymbol{\Psi}_{\delta k}]$ with a diagonal matrix $\boldsymbol{\Psi}_{\delta k}$, respectively. Similar to many SEMs, we assume that $|\mathbf{I} - \boldsymbol{\Pi}_k|$ is nonzero and independent of the elements in $\boldsymbol{\Pi}_k$. Let $\boldsymbol{\Lambda}_{\omega k} = (\mathbf{B}_k, \boldsymbol{\Pi}_k, \boldsymbol{\Gamma}_k)$, and $\mathbf{G}(\boldsymbol{\omega}_i) = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \mathbf{F}(\boldsymbol{\xi}_i)^T)^T$, (7.25) can be rewritten as

$$\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_{\omega k} \mathbf{G}(\boldsymbol{\omega}_i) + \boldsymbol{\delta}_i. \tag{7.26}$$

For each $k = 1, \cdots, K$, the measurement and structural equations defined by Equations (7.24) and (7.26) are not identified. A common method in structural equation modeling for model identification is fixing the appropriate elements in $\boldsymbol{\Lambda}_k$ and/or $\boldsymbol{\Lambda}_{\omega k}$ at preassigned values. In the subsequent analysis, let $\boldsymbol{\theta}_k$ be the parameter vector that includes all unknown parameters in $\boldsymbol{\mu}_k$, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Lambda}_{\omega k}$, $\boldsymbol{\Phi}_k$, $\boldsymbol{\Psi}_k$, and $\boldsymbol{\Psi}_{\delta k}$, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$ be the vector including all unknown structural parameters that define an identified model.

To deal with the missing data in mixture SEMs, we define the missing indicator vector $\mathbf{r}_i^y = (r_{i1}^y, \cdots, r_{ip}^y)^T$ of $\mathbf{y}_i$ such that $r_{ij}^y = 1$ if y_{ij} is missing, and $r_{ij}^y = 0$ otherwise. Similarly, the missing indicator vector of $\mathbf{d}_i$ is defined by $\mathbf{r}_i^d = (r_{i1}^d, \cdots, r_{im_2}^d)^T$. To cope with the nonignorable missing data in both responses and covariates, we need to define appropriate mechanisms to model the conditional distributions of $\mathbf{r}_i^y$ given $\mathbf{y}_i$ and $\boldsymbol{\omega}_i$, as well as $\mathbf{r}_i^d$ given $\boldsymbol{\omega}_i$ and $\mathbf{d}_i$. We assume that the conditional distributions of r_{ij}^y ($j = 1, \cdots, p$) given $\mathbf{y}_i$ and $\boldsymbol{\omega}_i$ are independent (Song and Lee, 2007), and the conditional distributions of r_{ij}^d ($j = 1, \cdots, m_2$) given $\mathbf{d}_i$ and $\boldsymbol{\omega}_i$ are independent. Thus, conditional on the kth component,

$$p(\mathbf{r}_i^y|\mathbf{y}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{ky}) = \prod_{j=1}^{p}\{p(r_{ij}^y = 1|\mathbf{y}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{ky})\}^{r_{ij}^y}\{1 - p(r_{ij}^y = 1|\mathbf{y}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{ky})\}^{1-r_{ij}^y},$$
$$p(\mathbf{r}_i^d|\mathbf{d}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{kd}) = \prod_{j=1}^{m_2}\{p(r_{ij}^d = 1|\mathbf{d}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{kd})\}^{r_{ij}^d}\{1 - p(r_{ij}^d = 1|\mathbf{d}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{kd})\}^{1-r_{ij}^d},$$

where $\boldsymbol{\varphi}_{ky}$ and $\boldsymbol{\varphi}_{kd}$ are the component-specific parameter vectors associated with $\mathbf{r}_i^y$ and $\mathbf{r}_i^d$, respectively. For simplicity, we further assume that $p(r_{ij}^y|\mathbf{y}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{ky}) = p(r_{ij}^y|\mathbf{y}_i, \boldsymbol{\varphi}_{ky})$ and $p(r_{ij}^d|\mathbf{d}_i, \boldsymbol{\omega}_i, \boldsymbol{\varphi}_{kd}) = p(r_{ij}^d|\mathbf{d}_i, \boldsymbol{\varphi}_{kd})$, and propose the following logistic regression models:

$$\text{logit}\{p(r_{ij}^y = 1|\mathbf{y}_i, \boldsymbol{\varphi}_{ky})\} = \varphi_{k0}^y + \varphi_{k1}^y y_{i1} + \cdots + \varphi_{kp}^y y_{ip} = \boldsymbol{\varphi}_{ky}^T \mathbf{u}_i^y, \tag{7.27}$$

$$\text{logit}\{p(r_{ij}^d = 1|\mathbf{d}_i, \boldsymbol{\varphi}_{kd})\} = \varphi_{k0}^d + \varphi_{k1}^d d_{i1} + \cdots + \varphi_{km_2}^d d_{im_2} = \boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d, \tag{7.28}$$

where $\mathbf{u}_i^y = (1, \mathbf{y}_i^T)^T$, $\boldsymbol{\varphi}_{ky} = (\varphi_{k0}^y, \varphi_{k1}^y, \cdots, \varphi_{kp}^y)^T$, $\mathbf{u}_i^d = (1, \mathbf{d}_i^T)^T$, and $\boldsymbol{\varphi}_{kd} = (\varphi_{k0}^d, \varphi_{k1}^d, \cdots, \varphi_{km_2}^d)^T$. To deal with the missing covariates in the multinomial logit model (7.23), we use $\mathbf{r}_i^x = (r_{i1}^x, \cdots, r_{im_1}^x)^T$ to represent the missing indicator vectors of $\mathbf{x}_i$, where r_{ij}^x is similarly defined as r_{ij}^y. The following mechanism is used to model the conditional distribution of

$\mathbf{r}_i^x$ given $\mathbf{x}_i$ and $\boldsymbol{\varphi}_x$:

$$p(\mathbf{r}_i^x|\mathbf{x}_i, \boldsymbol{\varphi}_x) = \prod_{j=1}^{m_1} \{p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\}^{r_{ij}^x}\{1 - p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\}^{1-r_{ij}^x},$$
$$\text{logit}\{p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\} = \varphi_0^x + \varphi_1^x x_{i1} + \cdots + \varphi_{m_1}^x x_{im_1} = \boldsymbol{\varphi}_x^T \mathbf{u}_i^x, \tag{7.29}$$

where $\mathbf{u}_i^x = (1, \mathbf{x}_i^T)^T$, and $\boldsymbol{\varphi}_x = (\varphi_0^x, \cdots, \varphi_{m_1}^x)^T$ is a parameter vector.

The above binomial modeling with logistic regressions is not the only choice for modeling the nonignorable missing mechanisms. For instance, one can also use probit regression models to model $p(r_{ij}^y|\cdot), p(r_{ij}^d|\cdot)$, and/or $p(r_{ij}^x|\cdot)$. We use the current modeling strategy because (i) conditional distributions associated with the missing components can easily be derived, (ii) not too many nuisance parameters are involved, and (iii) the logistic regression model is a natural way to model the probability of missingness (Ibrahim, Chen and Lipsitz, 2001). As the true missing mechanism is unknown, a comparison of modeling strategies can be viewed as a sensitivity analysis for model misspecification of the missing data mechanisms. Note that unknown parameters $\boldsymbol{\varphi}_{ky}$ and $\boldsymbol{\varphi}_{kd}$ in the missing mechanisms (7.27)-(7.28) are different across distinct components. Hence, the possible heterogeneity in relation to the missing mechanisms can be addressed.

7.3.2 Bayesian Estimation

Let $\mathbf{Y} = \{\mathbf{y}_i, i = 1, \cdots, n\}$, $\mathbf{D} = \{\mathbf{d}_i, i = 1, \cdots, n\}$, $\mathbf{X} = \{\mathbf{x}_i, i = 1, \cdots, n\}$, $\mathbf{y}_i = \{\mathbf{y}_{oi}, \mathbf{y}_{mi}\}$, $\mathbf{d}_i = \{\mathbf{d}_{oi}, \mathbf{d}_{mi}\}$, $\mathbf{x}_i = \{\mathbf{x}_{oi}, \mathbf{x}_{mi}\}$, where $\{\mathbf{y}_{oi}, \mathbf{d}_{oi}, \mathbf{x}_{oi}\}$ and $\{\mathbf{y}_{mi}, \mathbf{d}_{mi}, \mathbf{x}_{mi}\}$ denote the observed elements and missing elements in $\mathbf{y}_i, \mathbf{d}_i$, and $\mathbf{x}_i$, respectively. Let $\mathbf{Y}_o = \{\mathbf{y}_{oi}, i = 1, \cdots, n\}$, $\mathbf{R}^y = \{\mathbf{r}_i^y, i = 1, \cdots, n\}$, $\mathbf{D}_o = \{\mathbf{d}_{oi}, i = 1, \cdots, n\}$, $\mathbf{R}^d = \{\mathbf{r}_i^d, i = 1, \cdots, n\}$, $\mathbf{X}_o = \{\mathbf{x}_{oi}, i = 1, \cdots, n\}$, $\mathbf{R}^x = \{\mathbf{r}_i^x, i = 1, \cdots, n\}$, and $\mathbf{F}_o = \{\mathbf{Y}_o, \mathbf{D}_o, \mathbf{X}_o, \mathbf{R}^y, \mathbf{R}^d, \mathbf{R}^x\}$. Let $\boldsymbol{\pi}_k = \{\pi_{ik}, i = 1, \cdots, n\}$, $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_{K-1}\}$, $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{K-1}\}$, $\boldsymbol{\tau}_d = \{\boldsymbol{\tau}_{1d}, \cdots, \boldsymbol{\tau}_{Kd}\}$, $\boldsymbol{\varphi}_y = \{\boldsymbol{\varphi}_{1y}, \cdots, \boldsymbol{\varphi}_{Ky}\}$, $\boldsymbol{\varphi}_d = \{\boldsymbol{\varphi}_{1d}, \cdots, \boldsymbol{\varphi}_{Kd}\}$, $\boldsymbol{\vartheta}_s = \{\boldsymbol{\varphi}_y, \boldsymbol{\varphi}_d, \boldsymbol{\varphi}_x, \boldsymbol{\tau}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_d\}$, and $\boldsymbol{\theta}_* = \{\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\vartheta}_s\}$, where $\boldsymbol{\vartheta}_s$ contains the parameters in logistic regression models (7.27)-(7.29) and those involved in the probability distributions of $\mathbf{x}_i$ and $\mathbf{d}_i$ in models (7.23) and (7.25). As all the components in $\boldsymbol{\vartheta}_s$ are associated with the missing mechanisms rather than the modified mixture SEM, they are considered as a nuisance parameter vector in Bayesian analysis. To obtain the Bayesian estimate of $\boldsymbol{\theta}_*$, the main task is to draw observations from the posterior distribution $p(\boldsymbol{\theta}_*|\mathbf{F}_o)$. Due to the complexity of the model and the existence of nonignorable missing data, this posterior distribution involves high-dimensional integrals and does not have a closed form. We utilize the idea of data augmentation with the help of a latent allocation variable z_i for each $\mathbf{y}_i$. Here, we assume that z_i follows a multinomial distribution Multi$(\pi_{i1}, \cdots, \pi_{iK})$ with

$$p(z_i = k|\mathbf{x}_i) = \pi_{ik}, \quad k = 1, \cdots, K. \tag{7.30}$$

Let $\mathbf{Z} = \{z_1, \cdots, z_n\}$. If $\mathbf{Z}$ is given, the allocation of each $\mathbf{y}_i$ is identified, and the mixture model becomes a familiar multiple group model. Furthermore, let $\mathbf{Y}_m = \{\mathbf{y}_{mi}, i = 1, \cdots, n\}$, $\mathbf{D}_m = \{\mathbf{d}_{mi}, i = 1, \cdots, n\}$, $\mathbf{X}_m = \{\mathbf{x}_{mi}, i = 1, \cdots, n\}$, and $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_i, i = 1, \cdots, n\}$. The observed data $\mathbf{F}_o$ will be augmented with the latent quantities $\mathbf{F}_m = \{\mathbf{Y}_m, \mathbf{D}_m, \mathbf{X}_m, \boldsymbol{\Omega}, \mathbf{Z}\}$ in the posterior analysis. Hence, the Bayesian estimate of $\boldsymbol{\theta}_*$ can be obtained by drawing samples from the joint posterior distribution $p(\boldsymbol{\theta}_*, \mathbf{F}_m|\mathbf{F}_o)$ through Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler (Geman and Geman, 1984) and the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970). The random permutation sampler proposed by Frühwirth-Schnatter (2001) is used to deal with the label switching problem in the MCMC algorithm (see details in Section 7.2.2).

The full conditional distributions in implementing the MCMC algorithm involve the prior distributions of unknown parameters in $\boldsymbol{\theta}_*$. According to the suggestions given in the literature (Lee, 2007a; Lee and Song, 2003), the following conjugate prior distributions are used. Let $\boldsymbol{\Lambda}_{kj}^T$ and $\boldsymbol{\Lambda}_{\omega kl}^T$ be the jth and lth rows of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Lambda}_{\omega k}$, and ψ_{kj} and $\psi_{\delta kl}$ be the jth and the lth diagonal elements of $\boldsymbol{\Psi}_k$ and $\boldsymbol{\Psi}_{\delta k}$, respectively. The prior distributions of the unknown parameters in $\boldsymbol{\theta}_k,\ k = 1, \cdots, K$ are given as follows: For $j = 1, \cdots, p,\ l = 1, \cdots, q_1$,

$$\begin{aligned}
&[\boldsymbol{\Lambda}_{kj}|\psi_{kj}] \overset{D}{=} N[\boldsymbol{\Lambda}_{0kj}, \psi_{kj}\mathbf{H}_{0kj}], && \psi_{kj}^{-1} \overset{D}{=} Gamma[\alpha_{0kj}, \beta_{0kj}],\\
&[\boldsymbol{\Lambda}_{\omega kl}|\psi_{\delta kl}] \overset{D}{=} N[\tilde{\boldsymbol{\Lambda}}_{0kl}, \psi_{\delta kl}\tilde{\mathbf{H}}_{0kl}], && \psi_{\delta kl}^{-1} \overset{D}{=} Gamma[\tilde{\alpha}_{0kl}, \tilde{\beta}_{0kl}],\\
&\boldsymbol{\mu}_k \overset{D}{=} N[\boldsymbol{\mu}_{0k}, \boldsymbol{\Sigma}_{0k}], && \boldsymbol{\Phi}_k^{-1} \overset{D}{=} W_{q_2}[\mathbf{V}_{0k}, \rho_{0k}].
\end{aligned} \tag{7.31}$$

The prior distributions of $\boldsymbol{\tau}_k, \boldsymbol{\varphi}_x, \boldsymbol{\varphi}_{ky}$, and $\boldsymbol{\varphi}_{kd}$ are given as follows:

$$\begin{aligned}
&\boldsymbol{\tau}_k \overset{D}{=} N[\boldsymbol{\mu}_{0k\tau}, \boldsymbol{\Sigma}_{0k\tau}], && \boldsymbol{\varphi}_x \overset{D}{=} N[\boldsymbol{\varphi}_{0x}, \boldsymbol{\Sigma}_{0x}],\\
&\boldsymbol{\varphi}_{ky} \overset{D}{=} N[\boldsymbol{\varphi}_{0ky}, \boldsymbol{\Sigma}_{0ky}], && \boldsymbol{\varphi}_{kd} \overset{D}{=} N[\boldsymbol{\varphi}_{0kd}, \boldsymbol{\Sigma}_{0kd}].
\end{aligned} \tag{7.32}$$

Moreover, the conjugate prior distributions of $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_d$ can be related on the basis of the distributions of $\mathbf{x}_i$ and $\mathbf{d}_i$. The hyperparameters in (7.31) and (7.32) include $\boldsymbol{\Lambda}_{0kj}$, $\tilde{\boldsymbol{\Lambda}}_{0kl}$, α_{0kj}, β_{0kj}, $\tilde{\alpha}_{0kl}$, $\tilde{\beta}_{0kl}$, $\boldsymbol{\mu}_{0k}$, ρ_{0k}, $\boldsymbol{\mu}_{0k\tau}$, $\boldsymbol{\varphi}_{0x}$, $\boldsymbol{\varphi}_{0ky}$, $\boldsymbol{\varphi}_{0kd}$, and positive definite matrices $\mathbf{H}_{0kj}$, $\tilde{\mathbf{H}}_{0kl}$, $\boldsymbol{\Sigma}_{0k}$, $\boldsymbol{\Sigma}_{0k\tau}$, $\boldsymbol{\Sigma}_{0x}$, $\boldsymbol{\Sigma}_{0ky}$, $\boldsymbol{\Sigma}_{0kd}$, and $\mathbf{V}_{0k}$, whose values are assumed to be given according to prior information. With the above prior distributions, the full conditional distributions of the unknown parameters in $\boldsymbol{\theta}_*$ are derived in Appendix 7.3.

7.3.3 Bayesian Model Selection Using a Modified DIC

For the modified mixture SEM with missing data, competing models are compared with respect to (i) the different numbers of components involved in the mixture model; and (ii) the different missing mechanisms for the missing data. In this section, a modified Deviance Information Criterion (DIC; Spiegelhalter *et al.*, 2002) is considered for model selection. It is well known that directly applying DIC to the model selection of mixture models with incomplete data is problematic (Spiegelhalter *et al.*, 2003). Recently, Celeux *et al.* (2006) explored a wide range of options for constructing an appropriate DIC for mixture models. Here, we use one of these adaptations, namely, a modified DIC, as follows:

$$\text{DIC} = -4E_{\boldsymbol{\theta}_*,\mathbf{F}_m}\{\log p(\mathbf{F}_o, \mathbf{F}_m|\boldsymbol{\theta}_*)|\mathbf{F}_o\} + 2E_{\mathbf{F}_m}\{\log p(\mathbf{F}_o, \mathbf{F}_m|E_{\boldsymbol{\theta}_*}[\boldsymbol{\theta}_*|\mathbf{F}_o, \mathbf{F}_m])|\mathbf{F}_o\}, \tag{7.33}$$

where $\log p(\mathbf{F}_o, \mathbf{F}_m|\boldsymbol{\theta}_*)$ is the complete-data log-likelihood function, which can be written as follows:

$$\log p(\mathbf{F}_o, \mathbf{F}_m|\boldsymbol{\theta}_*) = \sum_{i=1}^{n} \log p(\mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i, z_i, \mathbf{x}_i, \mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{r}_i^x|\boldsymbol{\theta}_*),$$

where

$$\begin{aligned}
&\log p(\mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i, z_i, \mathbf{x}_i, \mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{r}_i^x|\boldsymbol{\theta}_*) \\
=& \log(\mathbf{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k, z_i = k) + \log p(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \mathbf{d}_i, \boldsymbol{\Lambda}_{\omega k}, \boldsymbol{\Psi}_{\delta k}, z_i = k) + \log p(\boldsymbol{\xi}_i|\boldsymbol{\Phi}_k, z_i = k) \\
&+ \log p(\mathbf{d}_i|\boldsymbol{\tau}_{kd}, z_i = k) + \log p(z_i = k|\boldsymbol{\tau}, \mathbf{x}_i) + \log p(\mathbf{x}_i|\boldsymbol{\tau}_x) \\
&+ \log p(\mathbf{r}_i^y|\mathbf{y}_i, \boldsymbol{\varphi}_{ky}, z_i = k) + \log p(\mathbf{r}_i^d|\mathbf{d}_i, \boldsymbol{\varphi}_{kd}, z_i = k) + \log p(\mathbf{r}_i^x|\boldsymbol{\varphi}_x, \mathbf{x}_i) \\
=& -\frac{1}{2}\big\{p\log(2\pi) + \log|\boldsymbol{\Psi}_k| + (\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)^T\boldsymbol{\Psi}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)\big\} \\
&- \frac{1}{2}\big\{q_1\log(2\pi) + \log|\boldsymbol{\Psi}_{\delta k}| + (\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i))^T\boldsymbol{\Psi}_{\delta k}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i))\big\} \\
&- \frac{1}{2}\big\{q_2\log(2\pi) + \log|\boldsymbol{\Phi}_k| + \boldsymbol{\xi}_i^T\boldsymbol{\Phi}_k^{-1}\boldsymbol{\xi}_i\big\} \\
&+ \log p(\mathbf{d}_i|\boldsymbol{\tau}_{kd}, z_i = k) + \boldsymbol{\tau}_k^T\mathbf{x}_i - \log\{\sum_{j=1}^{K}\exp(\boldsymbol{\tau}_j^T\mathbf{x}_i)\} + \log p(\mathbf{x}_i|\boldsymbol{\tau}_x) \\
&+ (\sum_{j=1}^{p} r_{ij}^y)\left(\boldsymbol{\varphi}_{ky}^T\mathbf{u}_i^y\right) - p\log\{1 + \exp(\boldsymbol{\varphi}_{ky}^T\mathbf{u}_i^y)\} + (\sum_{j=1}^{m_2} r_{ij}^d)\left(\boldsymbol{\varphi}_{kd}^T\mathbf{u}_i^d\right) - m_2\log\{1 + \exp(\boldsymbol{\varphi}_{kd}^T\mathbf{u}_i^d)\} \\
&+ (\sum_{j=1}^{m_1} r_{ij}^x)\left(\boldsymbol{\varphi}_x^T\mathbf{u}_i^x\right) - m_1\log\{1 + \exp(\boldsymbol{\varphi}_x^T\mathbf{u}_i^x)\}.
\end{aligned}$$

Hence, the first expectation in (7.33) can be obtained using the following approximation with the MCMC output $\{(\mathbf{F}_m^{(j)}, \boldsymbol{\theta}_*{}^{(j)}),\ j = 1, \cdots, J\}$:

$$E_{\boldsymbol{\theta}_*,\mathbf{F}_m}\{\log p(\mathbf{F}_o, \mathbf{F}_m|\boldsymbol{\theta}_*)|\mathbf{F}_o\} \approx \frac{1}{J}\sum_{j=1}^{J}\log p(\mathbf{F}_o, \mathbf{F}_m^{(j)}|\boldsymbol{\theta}_*{}^{(j)}).$$

Furthermore, let ${\boldsymbol{\theta}_*}^{(j,l)}$, $l = 1, \cdots, L$ be the observations generated from $p(\boldsymbol{\theta}_*|\mathbf{F}_o, \mathbf{F}_m^{(j)})$ via the MCMC method described above, we have

$$E_{\boldsymbol{\theta}_*}[\boldsymbol{\theta}_*|\mathbf{F}_o, \mathbf{F}_m^{(j)}] \approx \bar{\boldsymbol{\theta}}_*^{(j)} = \frac{1}{L}\sum_{l=1}^{L} {\boldsymbol{\theta}_*}^{(j,l)}.$$

Thus, the second expectation in (7.33) can be approximated by

$$E_{\mathbf{F}_m}\{\log p(\mathbf{F}_o, \mathbf{F}_m|E_{\boldsymbol{\theta}_*}[\boldsymbol{\theta}_*|\mathbf{F}_o, \mathbf{F}_m])|\mathbf{F}_o\} \approx \frac{1}{J}\sum_{j=1}^{J} \log p(\mathbf{F}_o, \mathbf{F}_m^{(j)}|\bar{\boldsymbol{\theta}}_*^{(j)}).$$

Finally, we can obtain the approximation of the modified DIC as follows:

$$\text{DIC} = -\frac{4}{J}\sum_{j=1}^{J} \log p(\mathbf{F}_o, \mathbf{F}_m^{(j)}|{\boldsymbol{\theta}_*}^{(j)}) + \frac{2}{J}\sum_{j=1}^{J} \log p(\mathbf{F}_o, \mathbf{F}_m^{(j)}|\bar{\boldsymbol{\theta}}_*^{(j)}).$$

7.3.4 An Illustrative Example

The methodology is illustrated through a longitudinal study of polydrug use conducted in five California counties in 2004 (Cai, Song and Hser, 2010). Data were collected from self-reported and administrative questionnaires about the retention of drug treatment, drug use history, drug-related crime history, motivation of drug treatment, and received service and test for 1,588 participants at intake, 3-month, and 12-month follow-up interviews. The modified mixture SEM is applied to examine the possible explanatory effects on treatment retention, and to explore possible heterogeneity in the data. Our primary interest is 'retention (Retent), y_1', which was collected at 12-month follow-up interview and which indicated the days of stay in treatment. Other observed variables include 'Drug use in past 30 days at intake (drgday30), y_2', 'Drug problems in past 30 days at intake (Drgplm30), y_3', 'The number of arrests in lifetime at intake (ArrN), y_4', 'The number of incarcerations in lifetime at intake (Incar), y_5', and 'The age of first arrest (Agefirstarrest), y_6'; see Appendix 1.1. These variables are treated as continuous. As $\{y_2, y_3\}$ are associated with the severity of drug use, they were grouped into a latent variable, 'drug severity, ξ_1', and as $\{y_4, y_5, y_6\}$ are associated with drug-related crime history, they were grouped into a latent variable, 'crime, ξ_2'. Therefore, an SEM is adopted, in which a measurement equation is used to identify two latent variables, 'drug severity' and 'crime', and a structural equation is used to study the influence of these latent variables on the outcome variable 'retention, η'. In addition, variables about treatment motivation (Mtsum01, Mtsum02, and Mtsum03) were collected at intake. They were treated as fixed covariates in the structural equation to incorporate their possible effects on retention. Moreover, some variables were collected at 3-month follow-up interview, including 'Services received in past 3 months at TSI 3 month interview (Servicem)', 'The number of drug tests by TX in past 3 months at TSI 3 month interview (DrugtestTX)', and 'The number of drug tests by criminal justice in past 3 months at TSI 3 month interview (DrugtestCJ)'. As these variables are related to the service and test received, they are likely to affect the pattern of influence of 'drug severity', 'crime', and treatment motivation on 'retention'. Thus, they were used to predict the component probability (see Equation (7.23)). In this study, there are 1,588 random observations, many of which have missing entries. Because most of variables in polydrug use data were non-normal, the logarithm and square root transformations were applied to the non-normal variables. Furthermore, the continuous measurements were standardized to unify the scale.

According to the aforementioned description, the model was formulated as follows. The six observed variables $y_1, \cdots, y_6$ were grouped into three latent variables η, ξ_1, and ξ_2, which were interpreted as 'retention', 'drug severity', and 'crime', respectively. In order to achieve clear interpretation of each latent variable, the following non-overlapping loading matrix $\mathbf{\Lambda}_k$ was used in each component:

$$\mathbf{\Lambda}_k^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \lambda_{k,32} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \lambda_{k,53} & \lambda_{k,63} \end{bmatrix},$$

where the ones and zeros were fixed for model identification. Furthermore, because there was only one indicator corresponding to the latent variable 'retention, η' ($\eta = y_1$), we fixed $\psi_{k1} = 0.0$ to identify the model. In each component, a structural equation was used to assess the effects of 'drug severity' and 'crime', together with covariates of treatment motivations (d_1, d_2, d_3), on 'retention' as follows:

$$\eta_i = b_{k1} d_{i1} + b_{k2} d_{i2} + b_{k3} d_{i3} + \gamma_{k1} \xi_{i1} + \gamma_{k2} \xi_{i2} + \delta_i.$$

The component probabilities π_{ik}'s were determined by the following multinomial logit model:

$$\pi_{ik} = \frac{\exp\{\tau_{k0} + \tau_{k1} x_{i1} + \tau_{k2} x_{i2} + \tau_{k3} x_{i3}\}}{\sum_{j=1}^{K} \exp\{\tau_{j0} + \tau_{j1} x_{i1} + \tau_{j2} x_{i2} + \tau_{j3} x_{i3}\}}, \qquad k = 1, \cdots, K.$$

Based on the nature of the questionnaires, we assumed that d_j, $j = 1, 2, 3$ came from multinomial distributions Multi$(\pi_{j1}, \cdots, \pi_{j5})$, x_1 came from a normal distribution, and x_2 and x_3 came from Poisson distributions.

The first step was to check the heterogeneity of the data using the modified DIC. For $k = 1, 2, 3$, let M_k be the k-component mixture SEM with nonignorable missing mechanisms defined by (7.27)-(7.29). In the following analysis, a vague prior was taken as follows: $\boldsymbol{\mu}_{0k} = \bar{\mathbf{y}}$, $\boldsymbol{\Sigma}_{0k} = 4\mathbf{S}_y^2$, where $\bar{\mathbf{y}}$ and $\mathbf{S}_y^2$ are the sample mean and sample covariance matrix calculated from the fully observed data; the elements in $\boldsymbol{\Lambda}_{0kj}$, $\tilde{\boldsymbol{\Lambda}}_{0kl}$, $\boldsymbol{\mu}_{0k\tau}$, $\boldsymbol{\varphi}_{0x}$, $\boldsymbol{\varphi}_{0ky}$, and $\boldsymbol{\varphi}_{0kd}$ are ones; and $\mathbf{H}_{0kj}$, $\tilde{\mathbf{H}}_{0kl}$, $\boldsymbol{\Sigma}_{0k\tau}$, $\boldsymbol{\Sigma}_{0x}$, $\boldsymbol{\Sigma}_{0ky}$, $\boldsymbol{\Sigma}_{0kd}$, and $\mathbf{V}_{0k}^{-1}$ are 10 times the identity matrices of appropriate order, $\alpha_{0kj} = \tilde{\alpha}_{0kl} = 5$, $\beta_{0kj} = \tilde{\beta}_{0kl} = 6$, and $\rho_{0k} = 13$. Based on some pilot runs, we found that the MCMC algorithm converged within 8,000 iterations. After discarding 8,000 burn-in iterations, additional 10,000 observations were used to compute the modified DIC values. The modified DIC values corresponding to M_k were $\text{DIC}_{M_1} = 74,512$, $\text{DIC}_{M_2} = 73,035$, and $\text{DIC}_{M_3} = 78,160$, respectively. Therefore, the two-component model M_2 was selected. To select appropriate mechanisms for missing responses and covariates, we compared M_2 with the following two-component models:

M_4: the missing data in $\mathbf{y}$ are treated as MAR, and in $\mathbf{d}$ and $\mathbf{x}$ are treated as nonignorable;

M_5: the missing data in $\mathbf{d}$ are treated as MAR, and in $\mathbf{y}$ and $\mathbf{x}$ are treated as nonignorable;

M_6: the missing data in $\mathbf{x}$ are treated as MAR, and in $\mathbf{y}$ and $\mathbf{d}$ are treated as nonignorable;

M_7: the missing data in $\mathbf{y}$, $\mathbf{d}$, and $\mathbf{x}$ are all treated as MAR.

M_8: instead of using logistic regression models (7.27)-(7.29), the probit regression models are used to model the missing mechanisms.

M_9: the explanatory covariates in models (7.27)-(7.29) are specified as:

$$\text{logit}\{p(r_{ij}^y = 1|\mathbf{y}_i, \boldsymbol{\varphi}_{ky})\} = \varphi_0^y + \varphi_j^y y_{ij}, \quad \text{logit}\{p(r_{ij}^d = 1|\mathbf{d}_i, \boldsymbol{\varphi}_{kd})\} = \varphi_0^d + \varphi_j^d d_{ij},$$

$$\text{logit}\{p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\} = \varphi_0^x + \varphi_j^x x_{ij}.$$

In the above model settings, M_4 to M_9 have the same number of components as the true model but have different missing mechanisms for $\mathbf{y}$, $\mathbf{d}$, and $\mathbf{x}$, respectively. The modified DIC values for M_4 to M_9 were equal to $\text{DIC}_{M_4} = 73,458$, $\text{DIC}_{M_5} = 73,162$, $\text{DIC}_{M_6} = 73,131$, $\text{DIC}_{M_7} = 74,001$, $\text{DIC}_{M_8} = 73,088$, and $\text{DIC}_{M_9} = 73,862$. Again, M_2 with the smallest $\text{DIC}_{M_2} = 73,035$ was selected.

On the basis of the selected model M_2, 10,000 observations collected after convergence were used to obtain the Bayesian estimates. The path diagrams of components 1 and 2 are presented in Figures 7.2 and 7.3, respectively, together with the Bayesian estimates of some interesting component-specific parameters. The Bayesian estimates of other parameters and their corresponding standard error estimates (SE) are presented in Table 7.5. We draw the following conclusions: (i) For components 1 and 2, the influences of latent variables ‘drug severity’ and ‘crime’ on retention have different patterns. This indicates that the mixture SEM with two components is necessary. For component 1, the effect of ‘drug severity’ is significant ($\hat{\gamma}_{11} = -0.236$, SE $= 0.081$), while the effect of ‘crime’ is insignificant ($\hat{\gamma}_{12} = -0.046$, SE $= 0.051$). Hence, it seems that in this component, participants with less drug use tend to stay in the treatment longer. In contrast, for component 2, the effect of 'drug severity' is insignificant ($\hat{\gamma}_{21} = -0.076$, SE $= 0.134$), while the effect of 'crime' is significant ($\hat{\gamma}_{22} = -0.151$, SE $= 0.074$). Thus, the participants with fewer drug-related crime records tend to stay in the treatment longer. Moreover, as the difference in sizes of $\hat{\gamma}_{11}$ and $\hat{\gamma}_{12}$, and $\hat{\gamma}_{21}$ and $\hat{\gamma}_{22}$ are substantial, more attention should be paid to the significant effects. (ii) For components 1 and 2, the effects of treatment motivation on retention are different. In particular, for component 1, the effect of 'Mtm03' on 'retention' is insignificant ($\hat{b}_{13} = 0.025$, SE $= 0.036$), while for component 2, the corresponding effect is significant ($\hat{b}_{23} = -0.208$, SE$=0.071$). (iii) There are differences in other parameter estimates in the two components. For example, estimates of $\lambda_{1,32}$ and $\lambda_{2,32}$, μ_{1j} and μ_{2j} for $j = 1, \cdots, 6$, and some nuisance parameters in the logistic regression models are quite different. (iv) τ_{11} and τ_{13} are significant, indicating that both service and drug tests received by the participants are useful to predict their component probabilities. (v) Many parameter estimates in $\boldsymbol{\varphi}_y$, $\boldsymbol{\varphi}_d$, and $\boldsymbol{\varphi}_x$ are substantially different from zero, indicating the necessity of the nonignorable mechanisms in the analysis of missing data. (vi) The different estimates of $\boldsymbol{\varphi}_y$ and $\boldsymbol{\varphi}_d$ in the two components imply the existence of component-specific patterns in the missing mechanisms for both responses and covariates. These conclusions reconfirm the above model comparison result in selecting a mixture SEM with two components.

Table 7.5, Figures 7.2 and 7.3 here

To investigate the sensitivity of Bayesian estimation and model selection to the prior input given in (7.31) and (7.32), the above analyses were repeated with some perturbations of the current prior inputs. Bayesian estimates obtained are close to those given in Table 7.5. The DIC values consistently selected the model (M_2) under the different prior inputs. The computer program for conducting this analysis is written in R compiled C code, which is given in the following website:

http://www.sta.cuhk.edu.hk/song-lee/book-chapter7(section7.3.4)/R-complied-C-code.

(PLEASE CHANGE TO WEB-SITES HOUSED IN JOHN-WILEY).

Appendix 7.1: The Permutation Sampler

Let $\boldsymbol{\psi} = (\boldsymbol{\Omega}, \mathbf{W}, \boldsymbol{\theta})$, the permutation sampler for generating $\boldsymbol{\psi}$ from the posterior $p(\boldsymbol{\psi}|\mathbf{Y})$ is implemented as follows:

1. First generate $\tilde{\boldsymbol{\psi}}$ from the unconstrained posterior $p(\boldsymbol{\psi}|\mathbf{Y})$ using standard Gibbs sampling steps;

2. Select some permutation $\rho(1), \cdots, \rho(K)$ of the current labeling of the states and define $\boldsymbol{\psi} = \rho(\tilde{\boldsymbol{\psi}})$ from $\tilde{\boldsymbol{\psi}}$ by reordering the labeling through this permutation, $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) := (\boldsymbol{\theta}_{\rho(1)}, \cdots, \boldsymbol{\theta}_{\rho(K)})$, and $\mathbf{W} = (w_1, \cdots, w_n) := (\rho(w_1), \cdots, \rho(w_n))$.

One application of permutation sampling is the random permutation sampler, where each sweep of the MCMC chain is concluded by relabeling the states through a random permutation of $\{1, \cdots, K\}$. This method delivers a sample that explores the whole unconstrained parameter space and jumps between the various labeling subspaces in a balanced fashion. Another application of the permutation sampler is the permutation sampling under identifiability constraints. A common way to include an identifiability constraint is to use a permutation sampler, where the permutation is selected in such a way that the identifiability constraint is fulfilled.

Appendix 7.2: Searching for Identifiability Constraints

For $k = 1, \cdots, K$, let $\boldsymbol{\theta}_k$ denote parameter vector corresponding to the kth component. According to the suggestion by Frühwirth-Schnatter (2001), the MCMC output of the random permutation sampler can be explored to find a suitable identifiability constraint. It is sufficient to consider only the parameters in $\boldsymbol{\theta}_1$, because a balanced sample from the unconstrained posterior will contain the same information for all parameters in $\boldsymbol{\theta}_k$ with $k > 1$. As the random permutation sampler jumps between the various labeling subspaces, part of the values sampled for $\boldsymbol{\theta}_1$ will belong to the first state, part will belong to the second state, and so on. To differ for various states, it is most useful to consider bivariate scatter plots of θ_{1i} versus θ_{1l} for possible combinations of i and l, where θ_{1i} and θ_{1l} indicate the ith and the lth element of $\boldsymbol{\theta}_1$, respectively. Jumping between the labeling subspaces produces groups in these scatter plots that correspond to different states. By describing the difference between the various groups geometrically, identification of a unique labeling subspace through conditions on the state-specific parameters is attempted. If the values sampled for a certain component of $\boldsymbol{\theta}$ differ markedly between the groups when jumping between the labeling subspaces, then an order condition on this component could be used to separate the labeling subspaces, while if the values sampled for a certain component of $\boldsymbol{\theta}$ hardly differ between the states when jumping between the labeling subspaces, then this component will be a poor candidate for separating the labeling subspaces.

Appendix 7.3: Conditional Distributions: Modified Mixture SEMs

(a) The full conditional distribution of $\mathbf{Y}_m$:

$$p(\mathbf{Y}_m|\mathbf{R}^y, \mathbf{Y}_o, \boldsymbol{\Omega}, \mathbf{Z}, \boldsymbol{\theta}_*) = \prod_{i=1}^{n} p(\mathbf{y}_{mi}|\mathbf{r}_i^y, \mathbf{y}_{oi}, \boldsymbol{\omega}_i, z_i = k, \boldsymbol{\theta}_*), \qquad \text{and}$$

$$\begin{aligned} p(\mathbf{y}_{mi}|\mathbf{r}_i^y, \mathbf{y}_{oi}, \boldsymbol{\omega}_i, z_i = k, \boldsymbol{\theta}_*) &\propto p(\mathbf{r}_i^y|\mathbf{y}_i, z_i = k, \boldsymbol{\varphi}_{ky}) p(\mathbf{y}_{mi}|\boldsymbol{\omega}_i, z_i = k, \boldsymbol{\theta}_k) \\ \propto \exp\Big[-\frac{1}{2}(\mathbf{y}_{mi} - \boldsymbol{\mu}_{k,\text{mis}} - \boldsymbol{\Lambda}_{k,\text{mis}}\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_{k,\text{mis}}^{-1} &(\mathbf{y}_{mi} - \boldsymbol{\mu}_{k,\text{mis}} - \boldsymbol{\Lambda}_{k,\text{mis}}\boldsymbol{\omega}_i) \\ + (\sum_{j=1}^{p} r_{ij}^y)(\boldsymbol{\varphi}_{ky}^T \mathbf{u}_i^y) - p \log\left\{1 + \exp(\boldsymbol{\varphi}_{ky}^T \mathbf{u}_i^y)\right\}\Big], \end{aligned} \tag{7.A1}$$

where $\boldsymbol{\mu}_{k,\text{mis}}$ is the subvector of $\boldsymbol{\mu}_k$, $\boldsymbol{\Lambda}_{k,\text{mis}}$ and $\boldsymbol{\Psi}_{k,\text{mis}}$ are the submatrices of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$ corresponding to $\mathbf{y}_{mi}$, respectively.

(b) The full conditional distribution of $\mathbf{D}_m$:

$$p(\mathbf{D}_m|\mathbf{R}^d, \mathbf{D}_o, \boldsymbol{\Omega}, \mathbf{Z}, \boldsymbol{\theta}_*) = \prod_{i=1}^{n} p(\mathbf{d}_{mi}|\mathbf{r}_i^d, \mathbf{d}_{oi}, \boldsymbol{\omega}_i, z_i = k, \boldsymbol{\theta}_*), \qquad \text{and}$$

$$
\begin{aligned}
&p(\mathbf{d}_{mi}|\mathbf{r}_i^d, \mathbf{d}_{oi}, \boldsymbol{\omega}_i, z_i = k, \boldsymbol{\theta}_*) \\
&\propto p(\mathbf{r}_i^d|\mathbf{d}_i, z_i = k, \boldsymbol{\varphi}_{kd}) p(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \mathbf{d}_i, z_i = k, \boldsymbol{\theta}_k) p(\mathbf{d}_{mi}|z_i = k, \boldsymbol{\tau}_{kd}) \\
&\propto \exp\Big[-\frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i))^T \boldsymbol{\Psi}_{\delta k}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i)) \\
&\qquad + (\sum_{j=1}^{m_2} r_{ij}^d)(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d) - m_2 \log\big\{1 + \exp(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d)\big\}\Big] \cdot p(\mathbf{d}_{mi}|\boldsymbol{\tau}_{kd}, z_i = k).
\end{aligned}
\tag{7.A2}
$$

(c) The full conditional distribution of $\mathbf{X}_m$:

$$
p(\mathbf{X}_m|\mathbf{R}^x, \mathbf{X}_o, \mathbf{Z}, \boldsymbol{\theta}_*) = \prod_{i=1}^n p(\mathbf{x}_{mi}|\mathbf{r}_i^x, \mathbf{x}_{oi}, z_i, \boldsymbol{\theta}_*), \qquad \text{and}
$$

$$
\begin{aligned}
p(\mathbf{x}_{mi}|\mathbf{r}_i^x, \mathbf{x}_{oi}, z_i, \boldsymbol{\theta}_*) &\propto p(z_i|\boldsymbol{\tau}, \mathbf{x}_i) p(\mathbf{r}_i^x|\mathbf{x}_i, \boldsymbol{\varphi}_x) p(\mathbf{x}_i|\boldsymbol{\tau}_x) \\
&\propto \exp\Big[\boldsymbol{\tau}_{z_i}^T \mathbf{x}_i - \log\{\sum_{j=1}^K \exp(\boldsymbol{\tau}_j^T \mathbf{x}_i)\} + (\sum_{j=1}^{m_1} r_{ij}^x)(\boldsymbol{\varphi}_x^T \mathbf{u}_i^x) \\
&\qquad - m_1 \log\big\{1 + \exp(\boldsymbol{\varphi}_x^T \mathbf{u}_i^x)\big\}\Big] \cdot p(\mathbf{x}_i|\boldsymbol{\tau}_x).
\end{aligned}
\tag{7.A3}
$$

(d) The full conditional distribution of $\boldsymbol{\Omega}$:

$$
p(\boldsymbol{\Omega}|\mathbf{Y}, \mathbf{D}, \mathbf{Z}, \boldsymbol{\theta}_*) = \prod_{i=1}^n p(\boldsymbol{\omega}_i|\mathbf{y}_i, \mathbf{d}_i, z_i = k, \boldsymbol{\theta}_*), \quad \text{and}
$$

$$
\begin{aligned}
&p(\boldsymbol{\omega}_i|\mathbf{y}_i, \mathbf{d}_i, z_i = k, \boldsymbol{\theta}_*) \\
&\propto p(\mathbf{y}_i|\boldsymbol{\omega}_i, z_i = k, \boldsymbol{\theta}_k) p(\boldsymbol{\eta}_i|\boldsymbol{\xi}_i, \mathbf{d}_i, z_i = k, \boldsymbol{\theta}_k) p(\boldsymbol{\xi}_i|z_i = k, \boldsymbol{\theta}_k) \\
&\propto \exp\Big[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i) \\
&\qquad - \frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i))^T \boldsymbol{\Psi}_{\delta k}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i)) - \frac{1}{2}\boldsymbol{\xi}_i^T \boldsymbol{\Phi}_k^{-1} \boldsymbol{\xi}_i\Big].
\end{aligned}
\tag{7.A4}
$$

(e) The full distribution of $\mathbf{Z}$:

$$
p(\mathbf{Z}|\mathbf{R}^y, \mathbf{R}^d, \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{D}, \mathbf{X}, \boldsymbol{\theta}_*) = \prod_{i=1}^n p(z_i|\mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i, \mathbf{x}_i, \boldsymbol{\theta}_*), \qquad \text{and}
$$

$$
p(z_i = k|\mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i, \mathbf{x}_i, \boldsymbol{\theta}_*) = \frac{\pi_{ik} p_k(\mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i|\boldsymbol{\theta}_k, \boldsymbol{\varphi}_{ky}, \boldsymbol{\varphi}_{kd}, \boldsymbol{\tau}_{kd})}{\sum_{j=1}^K \pi_{ij} p_j(\mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i|\boldsymbol{\theta}_k, \boldsymbol{\varphi}_{ky}, \boldsymbol{\varphi}_{kd}, \boldsymbol{\tau}_{kd})}. \tag{7.A5}
$$

where π_{ik} is defined in (7.23) and $p_k(\mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i|\boldsymbol{\theta}_k, \boldsymbol{\varphi}_{ky}, \boldsymbol{\varphi}_{kd}, \boldsymbol{\tau}_{kd})$ is given as follows:

$$
\begin{aligned}
&\log p_k(\mathbf{r}_i^y, \mathbf{r}_i^d, \mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{d}_i|\boldsymbol{\theta}_k, \boldsymbol{\varphi}_{ky}, \boldsymbol{\varphi}_{kd}, \boldsymbol{\tau}_{kd}) \\
&= (\sum_{j=1}^{p} r_{ij}^y)\left(\boldsymbol{\varphi}_{ky}^T \mathbf{u}_i^y\right) - p\log\left\{1 + \exp(\boldsymbol{\varphi}_{ky}^T \mathbf{u}_i^y)\right\} - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)^T \boldsymbol{\Psi}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i) \\
&- \frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i))^T \boldsymbol{\Psi}_{\delta k}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\Lambda}_{\omega k}\mathbf{G}(\boldsymbol{\omega}_i)) - \frac{1}{2}\boldsymbol{\xi}_i^T \boldsymbol{\Phi}_k^{-1}\boldsymbol{\xi}_i \\
&+ (\sum_{j=1}^{m_2} r_{ij}^d)\left(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d\right) - m_2\log\left\{1 + \exp(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d)\right\} + \log p(\mathbf{d}_i|\boldsymbol{\tau}_{kd}, z_i = k) \\
&- \frac{1}{2}\left\{\log|\boldsymbol{\Psi}_k| + \log|\boldsymbol{\Psi}_{\delta k}| + \log|\boldsymbol{\Phi}_k| + (p+q)\log(2\pi)\right\}.
\end{aligned}
$$

(f) The full conditional distribution of $\boldsymbol{\theta}$ are given as follows: Let $\boldsymbol{\Omega}_1 = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n)$, $\boldsymbol{\Omega}_2 = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n)$, and $\mathbf{G} = (\mathbf{G}(\boldsymbol{\omega}_1), \cdots, \mathbf{G}(\boldsymbol{\omega}_n))$. Further, let $\mathbf{Y}_k$, $\boldsymbol{\Omega}_k$, $\boldsymbol{\Omega}_{1k}$, $\boldsymbol{\Omega}_{2k}$, and $\mathbf{G}_k$ be the submatrices of $\mathbf{Y}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_1$, $\boldsymbol{\Omega}_2$, and $\mathbf{G}$ respectively such that all the i-th columns with $z_i \neq k$ are deleted. Under the conjugate prior distributions in (7.31), it can be shown that the full conditional distributions of components of $\boldsymbol{\theta}_k$ are:

$$
\begin{aligned}
&[\boldsymbol{\mu}_k|\cdot] \overset{D}{=} N\left[(\boldsymbol{\Sigma}_{0k}^{-1} + n_k\boldsymbol{\Psi}_k^{-1})^{-1}(n_k\boldsymbol{\Psi}_k^{-1}\bar{\mathbf{Y}}_k + \boldsymbol{\Sigma}_{0k}^{-1}\boldsymbol{\mu}_{0k}),\ (\boldsymbol{\Sigma}_{0k}^{-1} + n_k\boldsymbol{\Psi}_k^{-1})^{-1}\right], \\
&[\boldsymbol{\Lambda}_{kj}|\cdot] \overset{D}{=} N[\tilde{\boldsymbol{\Lambda}}_{kj}^*, \psi_{kj}\mathbf{H}_{kj}^*], \qquad [\psi_{kj}^{-1}|\cdot] \overset{D}{=} \text{Gamma}[n_k/2 + \alpha_{0kj}, \beta_{kj}^*], \\
&[\boldsymbol{\Lambda}_{\omega kl}|\cdot] \overset{D}{=} N[\tilde{\boldsymbol{\Lambda}}_{\omega kl}^*, \psi_{\delta kl}\tilde{\mathbf{H}}_{kl}^*], \qquad [\psi_{\delta kl}^{-1}|\cdot] \overset{D}{=} Gamma[n_k/2 + \tilde{\alpha}_{0kl}, \tilde{\beta}_{kl}^*], \\
&[\boldsymbol{\Phi}_k|\cdot] \overset{D}{=} IW_{q_2}[(\boldsymbol{\Omega}_{2k}\boldsymbol{\Omega}_{2k}^T + \mathbf{V}_{0k}^{-1}),\ n_k + \rho_{0k}],
\end{aligned}
$$

where $n_k = \sum_{i=1}^n I(z_i = k)$, where $I(\cdot)$ is an indicator function, $\bar{\mathbf{Y}}_k = \sum_{i,z_i=k}(\mathbf{y}_i - \boldsymbol{\Lambda}_k\boldsymbol{\omega}_i)/n_k$, and

$$
\begin{aligned}
&\tilde{\boldsymbol{\Lambda}}_{kj}^* = \mathbf{H}_{kj}^*[\mathbf{H}_{0kj}^{-1}\boldsymbol{\Lambda}_{0kj} + \boldsymbol{\Omega}_k\mathbf{Y}_{kj}^*], \qquad \mathbf{H}_{kj}^* = (\mathbf{H}_{0kj}^{-1} + \boldsymbol{\Omega}_k\boldsymbol{\Omega}_k^T)^{-1}, \\
&\beta_{kj}^* = \beta_{0kj} + [\mathbf{Y}_{kj}^{*^T}\mathbf{Y}_{kj}^* - \tilde{\boldsymbol{\Lambda}}_{kj}^{*^T}\mathbf{H}_{kj}^{*-1}\tilde{\boldsymbol{\Lambda}}_{kj}^* + \boldsymbol{\Lambda}_{0kj}^T\mathbf{H}_{0kj}^{-1}\boldsymbol{\Lambda}_{0kj}]/2, \\
&\tilde{\boldsymbol{\Lambda}}_{\omega kl}^* = \tilde{\mathbf{H}}_{kl}^*[\tilde{\mathbf{H}}_{0kl}^{-1}\tilde{\boldsymbol{\Lambda}}_{0kl} + \mathbf{G}_k\boldsymbol{\Omega}_{1kl}], \qquad \tilde{\mathbf{H}}_{kl}^* = (\tilde{\mathbf{H}}_{0kl}^{-1} + \mathbf{G}_k\mathbf{G}_k^T)^{-1}, \\
&\tilde{\beta}_{kl}^* = \tilde{\beta}_{0kl} + [\boldsymbol{\Omega}_{1kl}^T\boldsymbol{\Omega}_{1kl} - \tilde{\boldsymbol{\Lambda}}_{\omega kl}^{*^T}\tilde{\mathbf{H}}_{kl}^{*-1}\tilde{\boldsymbol{\Lambda}}_{\omega kl}^* + \tilde{\boldsymbol{\Lambda}}_{0kl}^T\tilde{\mathbf{H}}_{0kl}^{-1}\tilde{\boldsymbol{\Lambda}}_{0kl}]/2,
\end{aligned}
$$

where $\mathbf{Y}_{kj}^{*^T}$ is the jth row of $\mathbf{Y}_k^*$ which is a matrix whose columns are equal to the columns of $\mathbf{Y}_k$ minus μ_k, and $\boldsymbol{\Omega}_{1kl}^T$ is the lth row of $\boldsymbol{\Omega}_{1k}$.

(g) The full distribution of $\boldsymbol{\varphi}_y$:

$$p(\boldsymbol{\varphi}_{ky}|\mathbf{R}^y, \mathbf{Y}, \mathbf{Z}) \propto \Big\{ \prod_{i,z_i=k} p(\mathbf{r}_i^y|\mathbf{y}_i, \boldsymbol{\varphi}_{ky})\Big\} p(\boldsymbol{\varphi}_{ky}) \propto \tag{7.A6}$$

$$\exp\Big[\sum_{i,z_i=k} \{(\sum_{j=1}^{p} r_{ij}^y)(\boldsymbol{\varphi}_{ky}^T \mathbf{u}_i^y) - p\log(1+\exp(\boldsymbol{\varphi}_{ky}^T\mathbf{u}_i^y))\} - \frac{1}{2}(\boldsymbol{\varphi}_{ky}-\boldsymbol{\varphi}_{0ky})^T\boldsymbol{\Sigma}_{0ky}^{-1}(\boldsymbol{\varphi}_{ky}-\boldsymbol{\varphi}_{0ky})\Big].$$

(h) The full distribution of $\boldsymbol{\varphi}_d$:

$$p(\boldsymbol{\varphi}_{kd}|\mathbf{R}^d, \mathbf{D}, \mathbf{Z}) \propto \Big\{ \prod_{i,z_i=k} p(\mathbf{r}_i^d|\mathbf{d}_i, \boldsymbol{\varphi}_{kd})\Big\} p(\boldsymbol{\varphi}_{kd}) \propto \tag{7.A7}$$

$$\exp\Big[\sum_{i,z_i=k} \{(\sum_{j=1}^{m_2} r_{ij}^d)(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d) - m_2\log(1+\exp(\boldsymbol{\varphi}_{kd}^T\mathbf{u}_i^d))\} - \frac{1}{2}(\boldsymbol{\varphi}_{kd}-\boldsymbol{\varphi}_{0kd})^T\boldsymbol{\Sigma}_{0kd}^{-1}(\boldsymbol{\varphi}_{kd}-\boldsymbol{\varphi}_{0kd})\Big].$$

(i) The full distribution of $\boldsymbol{\varphi}_x$:

$$p(\boldsymbol{\varphi}_{x}|\mathbf{R}^x, \mathbf{X}) \propto \Big\{ \prod_{i=1}^{n} p(\mathbf{r}_i^x|\mathbf{x}_i, \boldsymbol{\varphi}_{x})\Big\} p(\boldsymbol{\varphi}_{x}) \propto \tag{7.A8}$$

$$\exp\Big[\sum_{i=1}^{n} \{(\sum_{j=1}^{m_1} r_{ij}^x)(\boldsymbol{\varphi}_{x}^T \mathbf{u}_i^x) - m_1\log(1+\exp(\boldsymbol{\varphi}_{x}^T\mathbf{u}_i^x))\} - \frac{1}{2}(\boldsymbol{\varphi}_{x}-\boldsymbol{\varphi}_{0x})^T\boldsymbol{\Sigma}_{0x}^{-1}(\boldsymbol{\varphi}_{x}-\boldsymbol{\varphi}_{0x})\Big].$$

(j) The full conditional distribution of $\boldsymbol{\tau}$:

$$\begin{aligned} p(\boldsymbol{\tau}_k|\mathbf{Z}, \mathbf{X}, \boldsymbol{\tau}_{-k}) &\propto p(\mathbf{Z}|\boldsymbol{\tau}, \mathbf{X})p(\boldsymbol{\tau}_k) = \Big\{\prod_{i=1}^{n} p(z_i|\boldsymbol{\tau}, \mathbf{x}_i)\Big\} p(\boldsymbol{\tau}_k) \\ &\propto \Big\{\prod_{i=1}^{n} \frac{\exp(\boldsymbol{\tau}_{z_i}^T\mathbf{x}_i)}{\sum_{j=1}^{K}\exp(\boldsymbol{\tau}_j^T\mathbf{x}_i)}\Big\} \exp\Big\{ -\frac{1}{2}(\boldsymbol{\tau}_k - \boldsymbol{\mu}_{0k\tau})^T\boldsymbol{\Sigma}_{0k\tau}^{-1}(\boldsymbol{\tau}_k - \boldsymbol{\mu}_{0k\tau})\Big\}, \end{aligned} \tag{7.A9}$$

where $\boldsymbol{\tau}_{-k}$ is $\boldsymbol{\tau}$ with $\boldsymbol{\tau}_k$ deleted.

The full conditional distributions in (7.A1)-(7.A9) are non-standard, we will use the MH algorithm to simulate observations from them. The full conditional distributions of $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_d$ depend on their prior distributions and the distributions of $\mathbf{x}_i$ and $\mathbf{d}_i$, respectively; they can be easily derived on a problem-by-problem basis.

References

Arminger, G., Stein, P. and Wittenberg, J. (1999) Mixtures of conditional mean- and covariance-structure models. *Psychometrika*, **64**, 475-494.

Cai, J. H., Song, X. Y. and Hser, Y. I. (2010) A Bayesian analysis of mixture structural equation models with nonignorable missing responses and covariates. *Statistics in Medicine* **29**, 1861-1874.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651-674.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363-375.

Dolan, C. V. and van der Maas, H. L. J. (1998) Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, **63**, 227-253.

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R. and Katz, I. R. (2005) Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction, *Biostatistics*, **6**, 119-143.

Frühwirth-Schnatter, S. (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194-209.

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.

Guo, J., Wall, M. and Amemiya, Y. (2006) Latent class regression on latent factors. *Biostatistics*, **7**, 145-163.

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

Hathaway, R. J. (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**, 795-800.

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2001) Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**, 551-564.

Jedidi, K., Jagpal, H. S. and DeSarbo, W. S. (1997) STEMM: A general finite mixture structural equation model. *Journal of Classification*, **14**, 23-50.

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

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

Lee, S. Y. (2007b) Bayesian analysis of mixtures structural equation models with missing data. In: Lee, S. Y. (ed.), *Handbook of Latent Variable and Related Models.* Amsterdam: Elsevier.

Lee, S. Y. and Song, X. Y. (2002) Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, **29**, 23-39.

Lee, S. Y. and Song, X. Y. (2003) Bayesian model selection for mixtures of structural equation models with an unknown number of components. *British Journal of Mathematical and Statistical Psychology*, **56**, 145-165.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 1-41.

Lindsay, B. G. and Basak, P. (1993) Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association*, **88**, 468-476.

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.

Muthén, B and Shedden, K. (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463-469.

Pettit, L. I. and Smith, A. F. M. (1985) Outliers and influential observations in linear models. In J. M. Bernardo *et al.* (eds), *Bayesian Statistics*, **2**, pp. 473-494. Amsterdam: Elsevier.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731-792.

Robert, C. P. (1996) Mixtures of distributions: Inference and estimation. In W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, pp. 441-464. London: Chapman and Hall.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894-902.

Song, X. Y. and Lee, S. Y. (2007) Bayesian analysis of latent variable models with non-ignorable missing outcomes from exponential family. *Statistics in Medicine*, **26**, 681-693.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583-639.

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

Yung, Y. F. (1997) Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, **62**, 297-330.

Zhu, H. T. and Lee, S. Y. (2001) A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, **66**, 133-152.

Table 7.1: Bayesian estimates in the artificial example.

Component 1			Component 2		
Par	EST	SE	Par	EST	SE
$\pi_1 = 0.5$	0.504	0.208	$\pi_2 = 0.5$	0.496	0.028
$\mu_{1,1} = 0.0$	0.070	0.082	$\mu_{2,1} = 0.0$	0.030	0.084
$\mu_{1,2} = 0.0$	0.056	0.046	$\mu_{2,2} = 0.0$	-0.056	0.061
$\mu_{1,3} = 0.0$	0.036	0.045	$\mu_{2,3} = 0.0$	-0.014	0.061
$\mu_{1,4} = 0.0$	0.108	0.075	$\mu_{2,4} = 0.5$	0.430	0.071
$\mu_{1,5} = 0.0$	0.209	0.061	$\mu_{2,5} = 1.5$	1.576	0.057
$\mu_{1,6} = 1.0$	1.101	0.062	$\mu_{2,6} = 0.0$	-0.084	0.052
$\mu_{1,7} = 1.0$	1.147	0.112	$\mu_{2,7} = 1.0$	0.953	0.110
$\mu_{1,8} = 1.0$	1.033	0.046	$\mu_{2,8} = 1.0$	1.041	0.056
$\mu_{1,9} = 1.0$	0.974	0.046	$\mu_{2,9} = 1.0$	0.941	0.057
$\lambda_{1,21} = 0.4$	0.322	0.052	$\lambda_{2,21} = 0.8$	0.851	0.060
$\lambda_{1,31} = 0.4$	0.411	0.052	$\lambda_{2,31} = 0.8$	0.810	0.060
$\lambda_{1,52} = 0.8$	0.712	0.060	$\lambda_{2,52} = 0.4$	0.498	0.067
$\lambda_{1,62} = 0.8$	0.695	0.066	$\lambda_{2,62} = 0.4$	0.480	0.064
$\lambda_{1,83} = 0.4$	0.386	0.075	$\lambda_{2,83} = 0.8$	0.738	0.077
$\lambda_{1,93} = 0.4$	0.428	0.079	$\lambda_{2,93} = 0.8$	0.826	0.083
$\gamma_{1,1} = 0.2$	0.236	0.104	$\gamma_{2,1} = 0.7$	0.817	0.104
$\gamma_{1,2} = 0.7$	0.740	0.074	$\gamma_{2,2} = 0.2$	0.210	0.074
$\phi_{1,11} = 1.0$	1.017	0.121	$\phi_{2,11} = 1.0$	0.820	0.118
$\phi_{1,12} = 0.3$	0.249	0.090	$\phi_{2,12} = 0.3$	0.283	0.074
$\phi_{1,22} = 1.0$	0.900	0.185	$\phi_{2,22} = 1.0$	0.982	0.163
$\psi_{11} = 0.5$	0.535	0.092	$\psi_{21} = 0.5$	0.588	0.080
$\psi_{12} = 0.5$	0.558	0.046	$\psi_{22} = 0.5$	0.489	0.053
$\psi_{13} = 0.5$	0.510	0.045	$\psi_{23} = 0.5$	0.565	0.057
$\psi_{14} = 0.5$	0.483	0.067	$\psi_{24} = 0.5$	0.620	0.085
$\psi_{15} = 0.5$	0.492	0.056	$\psi_{25} = 0.5$	0.556	0.056
$\psi_{16} = 0.5$	0.554	0.063	$\psi_{26} = 0.5$	0.507	0.050
$\psi_{17} = 0.5$	0.696	0.126	$\psi_{27} = 0.5$	0.569	0.108
$\psi_{18} = 0.5$	0.563	0.052	$\psi_{28} = 0.5$	0.578	0.062
$\psi_{19} = 0.5$	0.566	0.053	$\psi_{29} = 0.5$	0.508	0.065
$\psi_{\delta 11} = 0.5$	0.549	0.094	$\psi_{\delta 21} = 0.5$	0.549	0.082

Table 7.2: Bayesian estimates in the artificial example via WinBUGS.

Component 1			Component 2		
Par	EST	SE	Par	EST	SE
$\pi_1 = 0.5$	0.503	0.027	$\pi_2 = 0.5$	0.498	0.027
$\mu_{1,1} = 0.0$	0.035	0.068	$\mu_{2,1} = 0.0$	0.028	0.070
$\mu_{1,2} = 0.0$	0.050	0.046	$\mu_{2,2} = 0.0$	-0.043	0.062
$\mu_{1,3} = 0.0$	0.034	0.046	$\mu_{2,3} = 0.0$	-0.008	0.062
$\mu_{1,4} = 0.0$	0.124	0.067	$\mu_{2,4} = 0.5$	0.435	0.066
$\mu_{1,5} = 0.0$	0.192	0.060	$\mu_{2,5} = 1.5$	1.590	0.056
$\mu_{1,6} = 1.0$	1.089	0.061	$\mu_{2,6} = 0.0$	-0.065	0.053
$\mu_{1,7} = 1.0$	1.068	0.067	$\mu_{2,7} = 1.0$	0.961	0.066
$\mu_{1,8} = 1.0$	1.024	0.046	$\mu_{2,8} = 1.0$	1.053	0.056
$\mu_{1,9} = 1.0$	0.974	0.047	$\mu_{2,9} = 1.0$	0.916	0.057
$\lambda_{1,21} = 0.4$	0.355	0.047	$\lambda_{2,21} = 0.8$	0.850	0.053
$\lambda_{1,31} = 0.4$	0.428	0.047	$\lambda_{2,31} = 0.8$	0.817	0.053
$\lambda_{1,52} = 0.8$	0.723	0.062	$\lambda_{2,52} = 0.4$	0.507	0.065
$\lambda_{1,62} = 0.8$	0.741	0.068	$\lambda_{2,62} = 0.4$	0.495	0.064
$\lambda_{1,83} = 0.4$	0.403	0.058	$\lambda_{2,83} = 0.8$	0.741	0.061
$\lambda_{1,93} = 0.4$	0.385	0.061	$\lambda_{2,93} = 0.8$	0.837	0.064
$\gamma_{1,1} = 0.2$	0.208	0.071	$\gamma_{2,1} = 0.7$	0.873	0.103
$\gamma_{1,2} = 0.7$	0.740	0.094	$\gamma_{2,2} = 0.2$	0.152	0.068
$\phi_{1,11} = 1.0$	0.953	0.115	$\phi_{2,11} = 1.0$	0.785	0.110
$\phi_{1,12} = 0.3$	0.278	0.071	$\phi_{2,12} = 0.3$	0.305	0.067
$\phi_{1,22} = 1.0$	0.932	0.133	$\phi_{2,22} = 1.0$	0.988	0.117
$\psi_{11} = 0.5$	0.496	0.072	$\psi_{21} = 0.5$	0.519	0.059
$\psi_{12} = 0.5$	0.536	0.043	$\psi_{22} = 0.5$	0.512	0.052
$\psi_{13} = 0.5$	0.504	0.042	$\psi_{23} = 0.5$	0.566	0.054
$\psi_{14} = 0.5$	0.517	0.067	$\psi_{24} = 0.5$	0.620	0.076
$\psi_{15} = 0.5$	0.492	0.056	$\psi_{25} = 0.5$	0.536	0.052
$\psi_{16} = 0.5$	0.542	0.063	$\psi_{26} = 0.5$	0.518	0.051
$\psi_{17} = 0.5$	0.636	0.093	$\psi_{27} = 0.5$	0.533	0.066
$\psi_{18} = 0.5$	0.536	0.045	$\psi_{28} = 0.5$	0.574	0.054
$\psi_{19} = 0.5$	0.586	0.048	$\psi_{29} = 0.5$	0.487	0.054
$\psi_{\delta 11} = 0.5$	0.479	0.077	$\psi_{\delta 21} = 0.5$	0.510	0.072

Table 7.3: Bayesian estimates and standard error estimates of the ICPSR example.

	Component 1		Component 2	
Par	EST	SE	EST	SE
π_k	0.56	0.03	0.44	0.03
$\mu_{k,1}$	6.91	0.11	8.09	0.09
$\mu_{k,2}$	6.30	0.14	7.90	0.14
$\mu_{k,3}$	5.87	0.14	7.83	0.11
$\mu_{k,4}$	7.83	0.10	8.70	0.07
$\mu_{k,5}$	7.10	0.11	8.07	0.09
$\mu_{k,6}$	5.41	0.14	4.01	0.15
$\mu_{k,7}$	4.06	0.13	3.61	0.14
$\mu_{k,8}$	5.59	0.14	4.61	0.14
$\lambda_{k,11}$	1^*	–	1^*	–
$\lambda_{k,21}$	0.49	0.11	0.86	0.13
$\lambda_{k,32}$	1^*	–	1^*	–
$\lambda_{k,42}$	1.30	0.17	0.94	0.10
$\lambda_{k,52}$	1.58	0.20	1.02	0.11
$\lambda_{k,63}$	1^*	–	1^*	–
$\lambda_{k,73}$	2.05	0.44	0.98	0.07
$\lambda_{k,83}$	1.08	0.27	0.74	0.08
$\gamma_{k,1}$	0.68	0.14	0.77	0.11
$\gamma_{k,2}$	-0.02	0.15	-0.09	0.04
$\phi_{k,11}$	1.18	0.26	0.90	0.18
$\phi_{k,21}$	-0.12	0.09	-0.28	0.15
$\phi_{k,22}$	0.92	0.30	4.30	0.52
ψ_{k1}	1.56	0.65	0.56	0.11
ψ_{k2}	6.92	0.50	2.80	0.34
ψ_{k3}	4.86	0.37	1.35	0.18
ψ_{k4}	2.51	0.27	0.45	0.07
ψ_{k5}	1.29	0.27	0.55	0.08
ψ_{k6}	6.31	0.50	1.25	0.35
ψ_{k7}	2.43	0.76	1.07	0.23
ψ_{k8}	6.39	0.57	3.15	0.40
$\psi_{\delta k}$	3.38	0.72	0.70	0.12

This table is extracted from Zhu and Lee (2001).

Table 7.4: Bayesian estimates of parameters and standard errors for the selected model with 3 components in analyzing the ICPSR data set.

	Component 1		Component 2		Component 3	
Par	EST	SE	EST	SE	EST	SE
π_k	0.51	0.03	0.23	0.03	0.26	0.03
$\mu_{k,1}$	6.75	0.13	8.05	0.15	8.25	0.15
$\mu_{k,2}$	5.95	0.12	7.53	0.21	8.63	0.17
$\mu_{k,3}$	5.76	0.18	7.67	0.18	7.87	0.14
$\mu_{k,4}$	7.74	0.13	8.65	0.12	8.77	0.12
$\mu_{k,5}$	7.05	0.12	8.06	0.12	8.04	0.11
$\mu_{k,6}$	5.50	0.25	2.70	0.18	5.20	0.15
$\mu_{k,7}$	4.12	0.23	2.60	0.16	4.35	0.14
$\mu_{k,8}$	5.66	0.24	3.08	0.23	5.94	0.15
$\lambda_{k,21}$	0.31	0.12	1.10	0.21	0.66	0.08
$\lambda_{k,42}$	1.38	0.18	0.84	0.13	0.87	0.16
$\lambda_{k,52}$	1.67	0.19	0.92	0.15	1.10	0.18
$\lambda_{k,73}$	2.15	0.31	0.98	0.09	1.94	0.22
$\lambda_{k,83}$	0.88	0.22	0.97	0.11	0.64	0.20
$\gamma_{k,1}$	0.62	0.16	0.52	0.15	0.69	0.14
$\gamma_{k,2}$	0.01	0.11	-0.37	0.12	-0.12	0.14
$\phi_{k,11}$	1.07	0.21	1.30	0.33	0.81	0.20
$\phi_{k,21}$	-0.13	0.13	-0.59	0.20	0.07	0.07
$\phi_{k,22}$	1.10	0.49	1.45	0.38	1.57	0.21
ψ_{k1}	1.08	0.12	0.87	0.19	0.56	0.35
ψ_{k2}	6.79	0.17	2.02	0.45	0.83	0.46
ψ_{k3}	4.75	0.38	1.71	0.40	1.17	0.37
ψ_{k4}	2.54	0.13	0.45	0.08	0.69	0.27
ψ_{k5}	1.11	0.16	0.55	0.09	0.71	0.23
ψ_{k6}	5.75	0.55	0.55	0.13	4.71	0.46
ψ_{k7}	1.36	0.35	0.60	0.11	1.13	0.47
ψ_{k8}	6.10	0.52	1.05	0.50	4.52	0.47
$\psi_{\delta k}$	3.80	0.13	0.63	0.15	0.68	0.51

Table 7.5: Bayesian estimates of the parameters in the illustrative example.

	Component 1			Component 2	
Par.	Est	SE		Est	SE
μ_1	-0.008	0.069		0.460	0.124
μ_2	-0.297	0.055		0.615	0.077
μ_3	-0.620	0.079		1.209	0.141
μ_4	-0.122	0.041		0.248	0.052
μ_5	-0.052	0.033		0.114	0.051
μ_6	0.019	0.038		-0.003	0.049
ψ_2	0.364	0.051		0.545	0.079
ψ_3	0.371	0.053		0.163	0.090
ψ_4	0.364	0.032		0.441	0.051
ψ_5	0.304	0.036		0.383	0.049
ψ_6	0.956	0.059		0.762	0.061
ϕ_{11}	0.435	0.069		0.403	0.068
ϕ_{21}	0.037	0.025		0.041	0.038
ϕ_{22}	0.537	0.047		0.655	0.074
ψ_δ	0.903	0.051		0.833	0.066
φ_0^y	-4.393	0.598		-1.892	0.516
φ_1^y	-0.309	0.083		-0.236	0.153
φ_2^y	0.228	0.086		0.107	0.110
φ_3^y	-1.787	0.477		-1.075	0.360
φ_4^y	-0.162	0.125		-0.023	0.108
φ_5^y	0.050	0.105		0.044	0.124
φ_6^y	0.040	0.073		0.185	0.110
φ_0^d	-4.001	0.841		-1.326	0.586
φ_1^d	5.715	0.344		3.117	0.402
φ_2^d	-5.589	0.184		-4.965	0.595
φ_3^d	-2.139	0.463		-0.407	0.279
τ_{10}	0.827	0.246			
τ_{11}	-0.257	0.074			
τ_{12}	0.028	0.016			
τ_{13}	0.058	0.018			
non-component-specific parameters:					
φ_0^x	-3.930	0.301	φ_1^x	-3.069	0.217
φ_1^x	1.055	0.142	φ_3^x	2.451	0.169

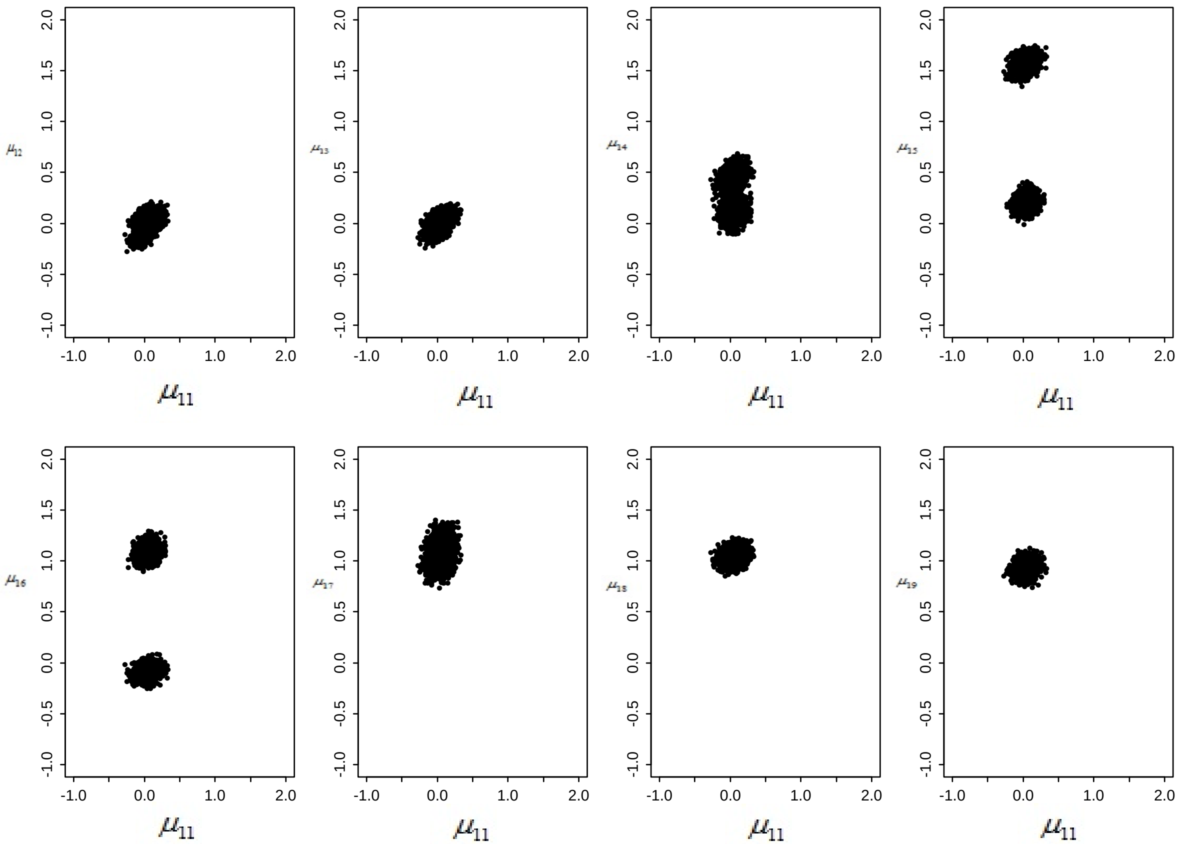

Figure 7.1: Scatter plots of MCMC output for components of $\boldsymbol{\mu}_1$.

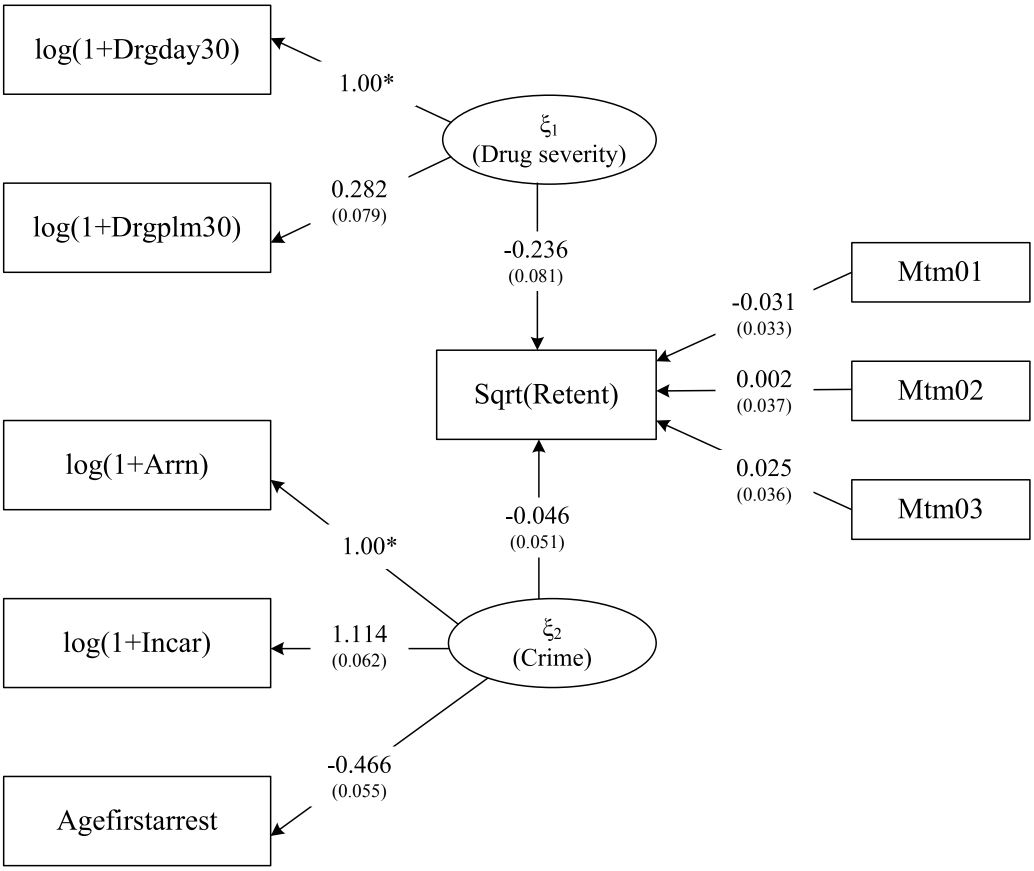

Figure 7.2: Path diagram of component 1 in the polydrug use example.

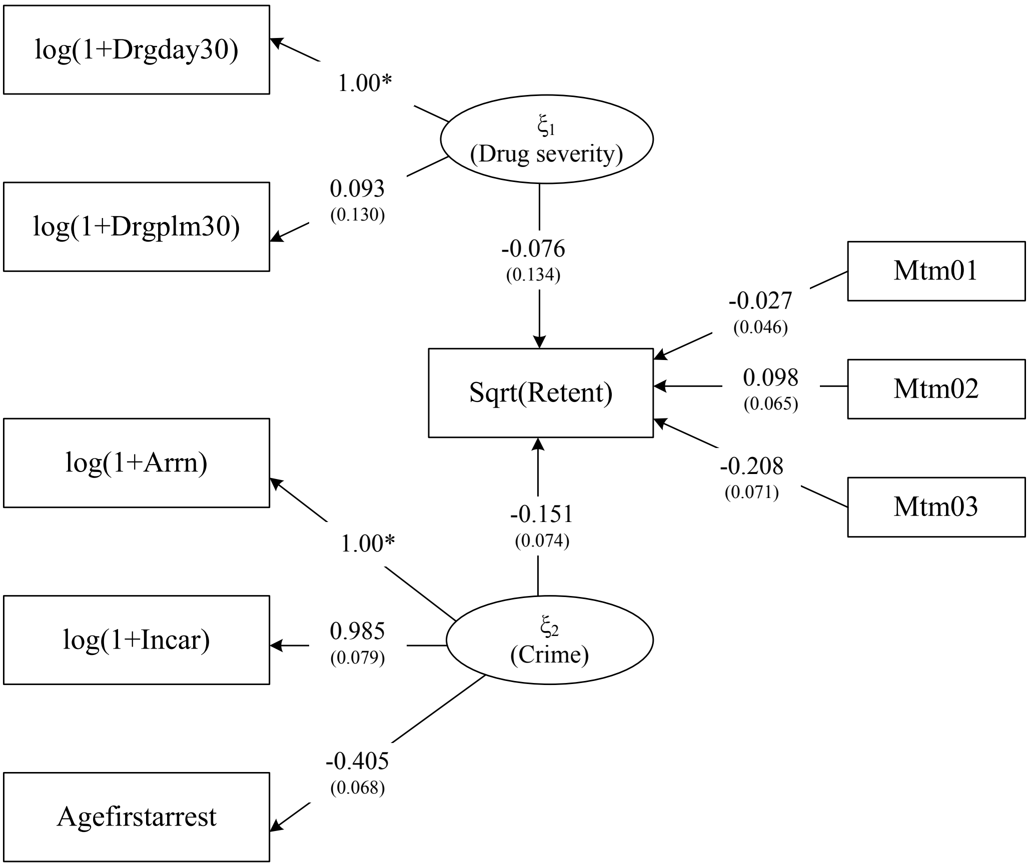

Figure 7.3: Path diagram of component 2 in the polydrug use example.