Chapter 7 Mixture Structural Equation Models

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 $\{\pi_k,\ k = 1, \cdots, K\}$.

The objectives of this chapter:

- Introduce a general finite mixture SEMs, in which $\pi_1, \cdots, \pi_K$ are unknown and estimated together with other parameters.
- Extend the general mixture SEM to a modified mixture SEM, in which the probabilities of component memberships are modeled through a multinomial logit model. This extension can accommodate
  - the effects of some important covariates on individuals' component memberships,
  - component-specific nonlinear interrelationships among latent variables,
  - nonigonralbe missing responses and covariates.
- Bayesian estimation and model comparison for mixture SEMs and modified mixture SEMs.

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:

every y_i is composed by K components, which is mixtured.

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

where

- K is a given integer,
- π_k is the unknown mixing proportion such that

intercepts must be different to distinguish the mixture component.

$$\pi_k > 0, \quad \text{and} \quad \pi_1 + \cdots + \pi_K = 1,$$

- $f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\theta}_k)$ 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:

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

and the structural equation of the model is defined as

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}_k \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_k \boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \tag{3}$$

where

- $\boldsymbol{\mu}_k$, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Pi}_k$, and $\boldsymbol{\Gamma}_k$ are defined as before,
- $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ are $q_1 \times 1$ and $q_2 \times 1$ subvectors of $\boldsymbol{\omega}_i$,
- $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\omega}_i$, and $\boldsymbol{\delta}_i$ is independent of $\boldsymbol{\xi}_i$,
- $\boldsymbol{\xi}_i \sim N[\mathbf{0}, \boldsymbol{\Phi}_k]$, $\boldsymbol{\epsilon}_i \sim N[\mathbf{0}, \boldsymbol{\Psi}_k]$, $\boldsymbol{\delta}_i \sim N[\mathbf{0}, \boldsymbol{\Psi}_{\delta k}]$, where $\boldsymbol{\Psi}_k$ and $\boldsymbol{\Psi}_{\delta k}$ are diagonal.

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 $\boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) = \boldsymbol{\Lambda}_k \boldsymbol{\Sigma}_{\omega k} \boldsymbol{\Lambda}_k^T + \boldsymbol{\Psi}_k$.

$\boldsymbol{\Phi}_k$: ξ's covariance.

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 (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, searching for an appropriate identifiability constraint is important. In this chapter, the random permutation sampler developed by Frühwirth-Schnatter (2001) will be utilized to achieve the purpose. Moreover, for each $k = 1, \cdots K$, the SEM is identified by fixing appropriate elements in $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Pi}_k$, and/or $\boldsymbol{\Gamma}_k$ at preassigned values.

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:

$$p(w_i = k) = \pi_k, \quad \text{for} \quad k = 1, \cdots, K. \tag{4}$$

Let

- $\boldsymbol{\theta}_{yk}$ — the vector of unknown parameters in $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$,
- $\boldsymbol{\theta}_{\omega k}$ — the vector of unknown parameters in $\boldsymbol{\Pi}_k$, $\boldsymbol{\Gamma}_k$, $\boldsymbol{\Phi}_k$, and $\boldsymbol{\Psi}_{\delta k}$,
- $\boldsymbol{\mu}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}_y$, and $\boldsymbol{\theta}_\omega$ — 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,
- $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega)$ — the overall parameter vector,
- $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$ — the matrix of observed variables,
- $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n)$ — the matrix of latent variables,
- $\mathbf{W} = (w_1, \cdots, w_n)$ — the vector of allocation variables.

In the posterior analysis, we should note that:

- Due to the nature of the mixture model, the posterior distribution $p[\boldsymbol{\theta}|\mathbf{Y}]$ is complicated. However, if $\mathbf{W}$ is observed, the component of every $\mathbf{y}_i$ can be identified and mixture models becomes familiar multiple group models. Hence, the posterior analysis is based on $p(\boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{W}|\mathbf{Y})$.

- The label switching problem. 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 $\boldsymbol{\theta}$ 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 use the random permutation sampler proposed by Frühwirth-Schnatter (2001) to deal with the label switching problem.

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

1. Generate $\tilde{\psi}$ from the unconstrained posterior $p(\psi|\mathbf{Y})$ using standard Gibbs sampling steps;
2. Select some permutation $\rho(1), \cdots, \rho(K)$ and define $\psi = \rho(\tilde{\psi})$ from $\tilde{\psi}$ by reordering the labeling through this permutation: $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) := (\boldsymbol{\theta}_{\rho(1)}, \cdots, \boldsymbol{\theta}_{\rho(K)})$, and $\mathbf{W} = (w_1, \cdots, w_n) := (\rho(w_1), \cdots, \rho(w_n))$.

Applications of permutation sampling:

- Random permutation sampler, in which 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.
- Permutation sampling with identifiability constraints, in which the permutation is selected such that the identifiability constraint is fulfilled.

Searching for identifiability constraints. According to the suggestion given by Frühwirth-Schnatter (2001),

- The MCMC output of the random permutation sampler can be explored to find a suitable identifiability constraint.
- Considering only the parameters in $\boldsymbol{\theta}_1$ is sufficient 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_{1j} for possible combinations of i and j.
- If certain θ_{1j} 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.

The Gibbs sampler for simulating observations from $[\boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{W}|\mathbf{Y}]$ is as follows: at the rth iteration with current values $\boldsymbol{\theta}^{(r)}$, $\boldsymbol{\Omega}^{(r)}$, and $\mathbf{W}^{(r)}$:

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

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

(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,

$$p(\mathbf{W}|\mathbf{Y},\boldsymbol{\theta}) = \prod_{i=1}^{n} p(w_i|\mathbf{y}_i,\boldsymbol{\theta}). \tag{5}$$

Moreover,

$$\begin{aligned} p(w_i = k|\mathbf{y}_i,\boldsymbol{\theta}) &= \frac{p(w_i = k,\mathbf{y}_i|\boldsymbol{\theta})}{p(\mathbf{y}_i|\boldsymbol{\theta})} \\ &= \frac{p(w_i = k|\boldsymbol{\pi})p(\mathbf{y}_i|w_i = k,\boldsymbol{\theta})}{p(\mathbf{y}_i|\boldsymbol{\theta})} = \frac{\pi_k f_k(\mathbf{y}_i|\boldsymbol{\mu}_k,\boldsymbol{\theta}_k)}{f(\mathbf{y}_i|\boldsymbol{\theta})}, \end{aligned} \tag{6}$$

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 $\boldsymbol{\theta}$ is independent multinomial distribution with the group probabilities given in (6).

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

$$\begin{aligned} p(\boldsymbol{\Omega}|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{W}) &= \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\mathbf{y}_i, w_i, \boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\omega}_i, w_i, \boldsymbol{\mu}, \boldsymbol{\theta}_y) p(\boldsymbol{\omega}_i|w_i, \boldsymbol{\theta}_\omega). \end{aligned} \tag{7}$$

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 that

$$[\boldsymbol{\omega}_i|\mathbf{y}_i, w_i = k, \boldsymbol{\theta}] \overset{D}{=} N[\mathbf{C}_k^{-1} \boldsymbol{\Lambda}_k^T \boldsymbol{\Psi}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k), \mathbf{C}_k^{-1}]. \tag{8}$$

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). This conditional distribution is quite complicated. The difficulty can be reduced by assuming the following mild conditions on the prior distribution of $\boldsymbol{\theta}$:

- The prior distribution of the mixing proportion $\boldsymbol{\pi}$ is independent of the prior distributions of $\boldsymbol{\mu}$, $\boldsymbol{\theta}_y$, and $\boldsymbol{\theta}_\omega$.
- The prior distribution of the mean vector $\boldsymbol{\mu}$ is independent of those of $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_\omega$ in the covariance structures.
- 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,

$$p(\mathbf{W}|\boldsymbol{\theta}) = p(\mathbf{W}|\boldsymbol{\pi}), \quad p(\boldsymbol{\Omega}, \mathbf{Y}|\mathbf{W}, \boldsymbol{\theta}) = p(\mathbf{Y}|\boldsymbol{\Omega}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\theta}_y)p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega).$$

Hence, the joint conditional distribution of $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega)$ can be expressed as

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{W}, \boldsymbol{\Omega}, \mathbf{Y}) &= p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega|\mathbf{W}, \boldsymbol{\Omega}, \mathbf{Y}) \\
&\propto p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)p(\mathbf{W}, \boldsymbol{\Omega}, \mathbf{Y}|\boldsymbol{\theta}) \\
&\propto p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)p(\mathbf{W}|\boldsymbol{\theta})p(\boldsymbol{\Omega}, \mathbf{Y}|\boldsymbol{\theta}, \mathbf{W}) \\
&\propto p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_\omega)p(\mathbf{W}|\boldsymbol{\pi})p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega)p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\theta}_y) \\
&= [p(\boldsymbol{\pi})p(\mathbf{W}|\boldsymbol{\pi})][p(\boldsymbol{\mu})p(\boldsymbol{\theta}_y)p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\theta}_y)][p(\boldsymbol{\theta}_\omega)p(\boldsymbol{\Omega}|\mathbf{W}, \boldsymbol{\theta}_\omega)]
\end{aligned}
\tag{9}
$$

mix E ME SE

Using this result, the conditional distirbutions of $p(\boldsymbol{\pi}|\cdot)$, $p(\boldsymbol{\mu}, \boldsymbol{\theta}_y|\cdot)$, 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 $p(\mathbf{W}|\boldsymbol{\pi}) \ \propto \ \prod_{k=1}^{K} \ \pi_k^{n_k}$, thus the full conditional distribution of $\boldsymbol{\pi}$ remains Dirichlet in the following form:

$$p(\boldsymbol{\pi}|\cdot) \propto p(\boldsymbol{\pi}) p(\mathbf{W}|\boldsymbol{\pi}) \propto \prod_{k=1}^{K} \ \pi_k^{n_k+\alpha}, \tag{10}$$

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 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

$$\begin{aligned} & p(\boldsymbol{\mu}, \boldsymbol{\theta}_y, \boldsymbol{\theta}_\omega | \mathbf{Y}, \boldsymbol{\Omega}, \mathbf{W}) \\ & \propto \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) p(\boldsymbol{\theta}_{yk}) p(\boldsymbol{\theta}_{\omega k}) p(\mathbf{Y}_k | \boldsymbol{\Omega}_k, \boldsymbol{\mu}_k, \boldsymbol{\theta}_{yk}) p(\boldsymbol{\Omega}_k | \boldsymbol{\theta}_{\omega k}), \end{aligned} \tag{11}$$

and we can treat the product in (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. Situations with cross-group constraints can be handled by the similar manner as a multiple groups analysis.

Let $\mathbf{\Lambda}_{\omega k} = (\mathbf{\Pi}_k, \mathbf{\Gamma}_k)$, for $m = 1, \cdots, p$, and $l = 1, \cdots, q_1$, we take

$$
\begin{aligned}
&[\mathbf{\Lambda}_{km}|\psi_{km}] \stackrel{D}{=} N[\mathbf{\Lambda}_{0km}, \psi_{km}\mathbf{H}_{0ykm}],\ \psi_{km}^{-1} \stackrel{D}{=} Gamma[\alpha_{0\epsilon k}, \beta_{0\epsilon k}],\\
&[\mathbf{\Lambda}_{\omega kl}|\psi_{\delta kl}] \stackrel{D}{=} N[\mathbf{\Lambda}_{0\omega kl}, \psi_{\delta kl}\mathbf{H}_{0\omega kl}],\ \psi_{\delta kl}^{-1} \stackrel{D}{=} Gamma[\alpha_{0\delta k}, \beta_{0\delta k}],\\
&\boldsymbol{\mu}_k \stackrel{D}{=} N[\boldsymbol{\mu}_0, \mathbf{\Sigma}_0],\ \mathbf{\Phi}_k^{-1} \stackrel{D}{=} W_{q_2}[\mathbf{R}_0, \rho_0],
\end{aligned}
\tag{12}
$$

where ψ_{km} and $\psi_{\delta kl}$ are the mth diagonal element of $\mathbf{\Psi}_k$ and the lth diagonal element of $\mathbf{\Psi}_{\delta k}$, $\mathbf{\Lambda}_{km}^T$ 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 $\mathbf{\Lambda}_{\omega k}$. Moreover, we also assume 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 $l \neq h$.

The conditional distributions of components of $\boldsymbol{\theta}_k$ are the familiar normal, Gamma, and inverted Wishart distributions; see details in Song and Lee's book. The computational burden required in simulating observations from these distributions is not heavy.

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_{hk} should be about 3.8 or over.

Objectives of this artificial example:

1. Investigate the performance of the Bayesian approach in analyzing mixture SEMs with two not well-separated components.
2. Demonstrate the random permutation sampler in finding suitable identifiability constraints.

Random observations are simulated from a mixture SEM with two components defined by

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

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i = \boldsymbol{\Pi}_k \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_k \boldsymbol{\xi}_i + \boldsymbol{\delta}_i.$$

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:

$$\eta = \gamma_{k,1}\xi_1 + \gamma_{k,2}\xi_2 + \delta.$$

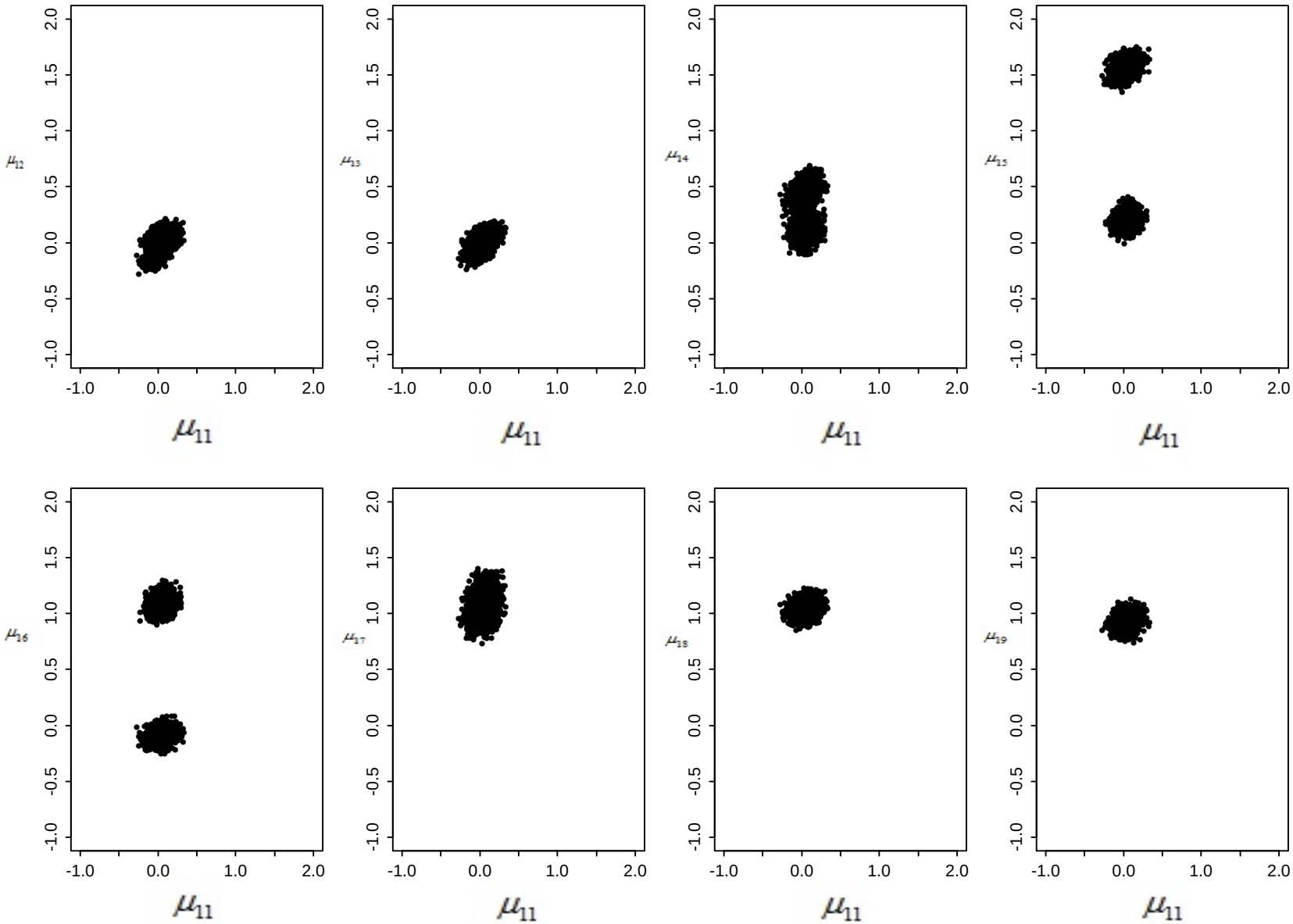
The true population values of the unknown parameters are given by:

$$\pi_1 = \pi_2 = 0.5,$$
$$\boldsymbol{\mu}_1 = (0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0)^T,$$
$$\boldsymbol{\mu}_2 = (0.0, 0.0, 0.0, 0.5, 1.5, 0.0, 1.0, 1.0, 1.0)^T,$$
$$\lambda_{1,21} = \lambda_{1,31} = \lambda_{1,83} = \lambda_{1,93} = 0.4,\ \lambda_{1,52} = \lambda_{1,62} = 0.8,$$
$$\lambda_{2,21} = \lambda_{2,31} = \lambda_{2,83} = \lambda_{2,93} = 0.8,\ \lambda_{2,52} = \lambda_{2,62} = 0.4,$$
$$\gamma_{1,1} = 0.2,\ \gamma_{1,2} = 0.7,\ \gamma_{2,1} = 0.7,\ \gamma_{2,2} = 0.2,$$
$$\phi_{1,11} = \phi_{1,22} = \phi_{2,11} = \phi_{2,22} = 1.0,\ \phi_{1,12} = \phi_{2,12} = 0.3,$$
$$\psi_{11} = \cdots = \psi_{19} = \psi_{21} = \cdots = \psi_{29} = \psi_{\delta 11} = \psi_{\delta 21} = 0.5.$$

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,

- 400 observations are simulated 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:
  (1) 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. For a mixture of SEMs with two components, we have 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.
  (2) 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 show that the most two significant differences between the two components are samples corresponding to $\mu_{1,5}$ and $\mu_{1,6}$.
- Bayesian estimates are obtained using the permutation sampler with the identifiability constraint $\mu_{1,5} < \mu_{2,5}$.


μ_{12}
μ_{13}
μ_{14}
μ_{15}
μ_{16}
μ_{17}
μ_{18}
μ_{19}
μ_{11}
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

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

A small portion of the ICPSR data set is used as an illustrative example:

- The data from the United Kingdom with a sample size 1,484 are used,
- Eight variables (116, 117, 180, 132, 96, 255, 254, and 252) related to respondents' job and home life are considered.
- After deleting the missing data, the remaining sample size is 824.

Assume k is known, here 2. but in real data application, the appropriate k is unknown.

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 $k = 1,\ 2$, $\mathbf{\Pi}_k = \mathbf{0},\ \mathbf{\Psi}_{\delta k} = \psi_{\delta k},\ \mathbf{\Gamma}_k = (\gamma_{k,1}, \gamma_{k,2})$,

$$\mathbf{\Lambda}_k^T = \begin{bmatrix} 1 & \lambda_{k,21} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{k,42} & \lambda_{k,52} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{k,73} & \lambda_{k,83} \end{bmatrix}, \ \mathbf{\Phi}_k = \begin{bmatrix} \phi_{k,11} & \phi_{k,12} \\ \phi_{k,21} & \phi_{k,22} \end{bmatrix}; \tag{13}$$

and $\mathbf{\Psi}_k = \text{diag}(\psi_{k1}, \cdots, \psi_{k8})$. The total number of unknown parameters is 56.

The following hyperparameters are used:

- $\alpha = 1$, $\boldsymbol{\mu}_0 = \bar{\mathbf{y}}$, $\boldsymbol{\Sigma}_0 = \mathbf{S}_y^2/2.0$, $\rho_0 = 5$, 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; $\mathbf{H}_{0ykm} = \mathbf{I}$, $\mathbf{H}_{0\omega kl} = \mathbf{I}$; $\boldsymbol{\Lambda}_{0km} = \tilde{\boldsymbol{\Lambda}}_{0km}$, and $\boldsymbol{\Lambda}_{0\omega kl} = \tilde{\boldsymbol{\Lambda}}_{0\omega kl}$ for all k, m, and l, where $\tilde{\boldsymbol{\Lambda}}_{0km}$ and $\tilde{\boldsymbol{\Lambda}}_{0\omega kl}$ are the initial estimates of $\boldsymbol{\Lambda}_{0km}$ and $\boldsymbol{\Lambda}_{0\omega kl}$ obtained using noninformative prior distributions.

We first use MCMC samples simulated with the random permutation sampler to find an identifiability constraint, and observe that $\mu_{1,1} < \mu_{2,1}$ is a suitable one. The Bayesian estimates of the structural parameters and their standard error estimates are reported in Table 7.3, which shows that there are at least two components which have different sets of Bayesian parameter estimates.

	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

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, together with the path sampling procedure, will be introduced.

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

$$B_{10} = \frac{P(\mathbf{Y}|M_1)}{P(\mathbf{Y}|M_0)}. \tag{14}$$

In computing the Bayes factor through path sampling, 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:

$$M_1: \quad \mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\pi} \overset{D}{=} \sum_{k=1}^{K} \pi_k f_k(\mathbf{Y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{15}$$

$$M_0: \quad \mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\pi}^* \overset{D}{=} \sum_{k=1}^{c} \pi_k^* f_k(\mathbf{Y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{16}$$

where $1 \leq c < K$. They can be linked up by a path M_t as follows:

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

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

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

By differentiation with respect to t, we have

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

where $f_k(\cdot)$ is written as follows.

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

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

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

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

The same portion of the ICPSR data set is reanalyzed to illustrate the path sampling procedure. We wish to find out whether there are some mixture models that are better than the 2-component mixture SEM.

The hyperparameter values are assigned as follows:

- $\alpha = 1$; $\boldsymbol{\mu}_0 = \bar{\mathbf{y}}$, $\boldsymbol{\Sigma}_0 = \mathbf{S}_y^2/2$, 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 $l = 1, \cdots, q_1$;
- $\{\alpha_{0\epsilon k}, \beta_{0\epsilon k}\}$, 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;
- $\boldsymbol{\Lambda}_{0km} = \tilde{\boldsymbol{\Lambda}}_{0km}$, $\boldsymbol{\Lambda}_{0\omega kl} = \tilde{\boldsymbol{\Lambda}}_{0\omega kl}$ for all k, m, and l, where $\tilde{\boldsymbol{\Lambda}}_{0km}$ and $\tilde{\boldsymbol{\Lambda}}_{0\omega kl}$ are the corresponding Bayesian estimates obtained through a single component model with noninformative prior distributions.

Let M_k denotes the mixture model with k components, the estimated logarithms of Bayes factors are equal to:

$$\widehat{\log B_{21}} = 75.055, \quad \widehat{\log B_{32}} = 4.381,$$
$$\widehat{\log B_{43}} = -0.824, \quad \widehat{\log B_{53}} = -1.395.$$

Hence, a mixture model with three components should be chosen. Although the two-component model suggested above is a plausible model, it does not give as strong support of evidence as the three-component model. 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.

In estimation, based on the MCMC samples simulated with 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 given below.

	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

In this section, we introduce a modified mixture SEM which extends the previous mixture SEMs in three aspects:

1. To examine the effects of covariates on the probability of component membership, a multinomial logit model with covariates is incorporated to predict the unknown component membership.
2. To capture the component-specific nonlinear effects of explanatory latent variables and covariates on outcome latent variables, a nonlinear structural equation is introduced in each component
3. To incoporate nonignorable missing data for both responses and covariates. It has been demonstrated that ignoring missing data would lead to less accurate estimation and misleading model comparison results. Moreover, the assumption of 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.

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

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_{ik} f_k(\mathbf{y}_i|\boldsymbol{\theta}_k), \quad i = 1, \cdots, n, \tag{17}$$

where K, π_{ik}, and $f_k(\cdot)$ are defined as before, and π_{ik} are modeled as follows:

$$\pi_{ik} = p(z_i = k|\mathbf{x}_i) = \frac{\exp\{\boldsymbol{\tau}_k^T \mathbf{x}_i\}}{\sum_{j=1}^{K} \exp\{\boldsymbol{\tau}_j^T \mathbf{x}_i\}}, \tag{18}$$

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 $\boldsymbol{\tau}_x$. The main purpose of Equation (18) is to provide an improved model with some covariates for predicting unknown component probabilities.

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

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{19}$$

where $\boldsymbol{\mu}_k$, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\omega}_i$, $\boldsymbol{\epsilon}_i$ are defined as before.

To assess the interrelationships among $\boldsymbol{\eta}_i$, $\boldsymbol{\xi}_i$, and some fixed covariates, a general structural equation is defined as follows:

$$\boldsymbol{\eta}_i = \mathbf{B}_k \mathbf{d}_i + \boldsymbol{\Pi}_k \boldsymbol{\eta}_i + \boldsymbol{\Gamma}_k \mathbf{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i, \tag{20}$$

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 $\boldsymbol{\tau}_{kd}$; $\mathbf{B}_k$, $\boldsymbol{\Pi}_k$, $\boldsymbol{\Gamma}_k$, $\boldsymbol{\xi}_i$, $\mathbf{F}(\boldsymbol{\xi}_i)$, and $\boldsymbol{\delta}_i$ are defined as before. Let $\boldsymbol{\Lambda}_{\omega k} = (\mathbf{B}_k, \boldsymbol{\Pi}_k, \boldsymbol{\Gamma}_k)$, and $\mathbf{G}(\boldsymbol{\omega}_i) = (\mathbf{d}_i^T, \boldsymbol{\eta}_i^T, \mathbf{F}(\boldsymbol{\xi}_i)^T)^T$, (20) can be rewritten as

$$\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_{\omega k} \mathbf{G}(\boldsymbol{\omega}_i) + \boldsymbol{\delta}_i. \tag{21}$$

To deal with the missing data in mixture SEMs, we define the missing indicator vectors $\mathbf{r}_i^y = (r_{i1}^y, \cdots, r_{ip}^y)^T$ of $\mathbf{y}_i$ and $\mathbf{r}_i^d = (r_{i1}^d, \cdots, r_{im_2}^d)^T$ of $\mathbf{d}_i$ such that

$$r_{ij}^y = \begin{cases} 1 & \text{if } y_{ij} \text{ is missing} \\ 0 & \text{otherwise,} \end{cases} \quad \text{and}$$

$$r_{ij}^d = \begin{cases} 1 & \text{if } d_{ij} \text{ is missing} \\ 0 & \text{otherwise.} \end{cases}$$

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 given $\mathbf{y}_i$ and $\boldsymbol{\omega}_i$ and r_{ij}^d given $\mathbf{d}_i$ and $\boldsymbol{\omega}_i$ are independent.

Thus, conditional on the kth component,

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

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

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:

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

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

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 (18), 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 $\boldsymbol{\varphi}_x$:

$$p(\mathbf{r}_i^x|\mathbf{x}_i, \boldsymbol{\varphi}_x) = \prod_{j=1}^{m_1} \{p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\}^{r_{ij}^x} \{1 - p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\}^{1-r_{ij}^x},$$

$$\text{logit}\{p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\} = \varphi_0^x + \varphi_1^x x_{i1} + \cdots + \varphi_{m_1}^x x_{im_1} = \boldsymbol{\varphi}_x^T \mathbf{u}_i^x, \quad (22)$$

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. 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,

(III) the logistic regression model is a natural way to model the probability of missingness.

poly. regression can also be applied but no close form of solution and introduce additional latent variables nuisance.

Note that

- 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.
- The unknown parameters φ_{ky} and φ_{kd} in the above missing models are different across distinct components. Hence, the possible heterogeneity in relation to the missing mechanisms can be addressed.

Let o: observed data m: missing data.

- $\mathbf{Y} = \{\mathbf{y}_i, i = 1, \cdots, n\}$, $\mathbf{D} = \{\mathbf{d}_i, i = 1, \cdots, n\}$, $\mathbf{X} = \{\mathbf{x}_i, i = 1, \cdots, n\}$, $\mathbf{Z} = \{z_1, \cdots, z_n\}$, $\mathbf{\Omega} = \{\omega_i, i = 1, \cdots, n\}$,
- $\mathbf{y}_i = \{\mathbf{y}_{oi}, \mathbf{y}_{mi}\}$, $\mathbf{d}_i = \{\mathbf{d}_{oi}, \mathbf{d}_{mi}\}$, $\mathbf{x}_i = \{\mathbf{x}_{oi}, \mathbf{x}_{mi}\}$, 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.
- $\mathbf{Y}_o = \{\mathbf{y}_{oi}, i = 1, \cdots, n\}$, $\mathbf{R}^y = \{\mathbf{r}_i^y, i = 1, \cdots, n\}$, $\mathbf{D}_o = \{\mathbf{d}_{oi}, i = 1, \cdots, n\}$, $\mathbf{R}^d = \{\mathbf{r}_i^d, i = 1, \cdots, n\}$, $\mathbf{X}_o = \{\mathbf{x}_{oi}, i = 1, \cdots, n\}$, $\mathbf{R}^x = \{\mathbf{r}_i^x, i = 1, \cdots, n\}$, $\mathbf{F}_o = \{\mathbf{Y}_o, \mathbf{D}_o, \mathbf{X}_o, \mathbf{R}^y, \mathbf{R}^d, \mathbf{R}^x\}$, full observed variables.
- $\mathbf{Y}_m = \{\mathbf{y}_{mi}, i = 1, \cdots, n\}$, $\mathbf{D}_m = \{\mathbf{d}_{mi}, i = 1, \cdots, n\}$, $\mathbf{X}_m = \{\mathbf{x}_{mi}, i = 1, \cdots, n\}$, $\mathbf{F}_m = \{\mathbf{Y}_m, \mathbf{D}_m, \mathbf{X}_m, \mathbf{\Omega}, \mathbf{Z}\}$,
- $\boldsymbol{\pi}_k = \{\pi_{ik}, i = 1, \cdots, n\}$, $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_{K-1}\}$, $\boldsymbol{\tau} = \{\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{K-1}\}$,
- $\boldsymbol{\tau}_d = \{\boldsymbol{\tau}_{1d}, \cdots, \boldsymbol{\tau}_{Kd}\}$, $\boldsymbol{\varphi}_y = \{\boldsymbol{\varphi}_{1y}, \cdots, \boldsymbol{\varphi}_{Ky}\}$, $\boldsymbol{\varphi}_d = \{\boldsymbol{\varphi}_{1d}, \cdots, \boldsymbol{\varphi}_{Kd}\}$,
- $\boldsymbol{\vartheta}_s = \{\boldsymbol{\varphi}_y, \boldsymbol{\varphi}_d, \boldsymbol{\varphi}_x, \boldsymbol{\tau}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_d\}$, $\boldsymbol{\theta}_* = \{\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\vartheta}_s\}$, where $\boldsymbol{\vartheta}_s$ contains the parameters in the logistic regression models and those involved in the probability distributions of $\mathbf{x}_i$ and $\mathbf{d}_i$.

To obtain the Bayesian estimate of $\boldsymbol{\theta}_*$, the main task is to draw observations from $p(\boldsymbol{\theta}_*|\mathbf{F}_o)$. 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

$$p(z_i = k|\mathbf{x}_i) = \pi_{ik}, \quad k = 1, \cdots, K. \tag{23}$$

The Bayesian estimate of $\boldsymbol{\theta}_*$ can be obtained by drawing samples from the $p(\boldsymbol{\theta}_*, \mathbf{F}_m|\mathbf{F}_o)$ through MCMC methods. The random permutation sampler is used to deal with the label switching problem in the MCMC algorithm.

For the modified mixture SEM with missing data, competing models are compared with respect to

(I) different numbers of components involved in the mixture model;

(II) different missing mechanisms for the missing data.

In this section, a modified DIC 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:

$$\begin{aligned}\text{DIC} = &-4E_{\theta_*,\mathbf{F}_m}\{\log p(\mathbf{F}_o,\mathbf{F}_m|\theta_*)|\mathbf{F}_o\} \\ &+ 2E_{\mathbf{F}_m}\{\log p(\mathbf{F}_o,\mathbf{F}_m|E_{\theta_*}[\theta_*|\mathbf{F}_o,\mathbf{F}_m])|\mathbf{F}_o\},\end{aligned} \tag{24}$$

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

$$\log p(\mathbf{F}_o,\mathbf{F}_m|\theta_*) = \sum_{i=1}^{n} \log p(\mathbf{y}_i,\omega_i,\mathbf{d}_i,z_i,\mathbf{x}_i,\mathbf{r}_i^y,\mathbf{r}_i^d,\mathbf{r}_i^x|\theta_*), \quad \text{and}$$

$$
\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) \\
&\quad + \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) \\
&\quad + \log p(\mathbf{x}_i | \boldsymbol{\tau}_x) + \log p(\mathbf{r}_i^y | \mathbf{y}_i, \boldsymbol{\varphi}_{ky}, z_i = k) \\
&\quad + \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}\{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)\} \\
&\quad - \frac{1}{2}\{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))\} \\
&\quad - \frac{1}{2}\{q_2 \log(2\pi) + \log|\boldsymbol{\Phi}_k| + \boldsymbol{\xi}_i^T \boldsymbol{\Phi}_k^{-1} \boldsymbol{\xi}_i\} \\
&\quad + \log p(\mathbf{d}_i | \boldsymbol{\tau}_{kd}, z_i = k) + \left[\boldsymbol{\tau}_k^T \mathbf{x}_i - \log\{\sum_{j=1}^{K} \exp(\boldsymbol{\tau}_j^T \mathbf{x}_i)\}\right] + \log p(\mathbf{x}_i | \boldsymbol{\tau}_x) \\
&\quad + \left[(\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)\}\right] + \left[(\sum_{j=1}^{m_2} r_{ij}^d)\left(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d\right)\right. \\
&\quad \left. - m_2 \log\{1 + \exp(\boldsymbol{\varphi}_{kd}^T \mathbf{u}_i^d)\}\right] + \left[(\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)\}\right]
\end{aligned}
$$

Hence, the first expectation in (24) can be obtained below:

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

loop 2.

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

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

loop 1

Thus, the second expectation in (24) 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:

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

no need for two loops

The methodology is illustrated through a longitudinal study of polydrug use conducted in five California counties in 2004. 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,
- 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,
- explore possible heterogeneity in the data.

The observed variables:

- 'retention (Retent), y_1', which was collected at 12-month follow-up interview. It indicates the days of stay in treatment, and is our primary interest.
- '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'
- 'The age of first arrest (Agefirstarrest), y_6'.

The latent variables:

- $\{y_2, y_3\}$ are associated with the severity of drug use, they were grouped into a latent variable, 'drug severity, ξ_1',
- $\{y_4, y_5, y_6\}$ are associated with drug-related crime history, they were grouped into a latent variable, 'crime, ξ_2'.

A modified mixture SEM is adopted. Specifically,

- The measurement equation is used to identify two latent variables, 'drug severity' and 'crime'.
- The 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.
- 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 they 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.

The proposed 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.

In the analysis, 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 above.

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;
- $\mathbf{H}_{0kj},\ \tilde{\mathbf{H}}_{0kl},\ \boldsymbol{\Sigma}_{0k\tau},\ \boldsymbol{\Sigma}_{0x},\ \boldsymbol{\Sigma}_{0ky},\ \boldsymbol{\Sigma}_{0kd}$, 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 $\rho_{0k} = 13$.

After convergence, 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:

missing at random

M_4: the missing data in **y** are treated as MAR, and in **d** and **x** are treated as nonignorable;

M_5: the missing data in **d** are treated as MAR, and in **y** and **x** are treated as nonignorable;

M_6: the missing data in **x** are treated as MAR, and in **y** and **d** are treated as nonignorable;

M_7: the missing data in **y**, **d**, and **x** are all treated as MAR.

M_8: instead of using logistic regression models, the probit regression models are used to model the missing mechanisms.

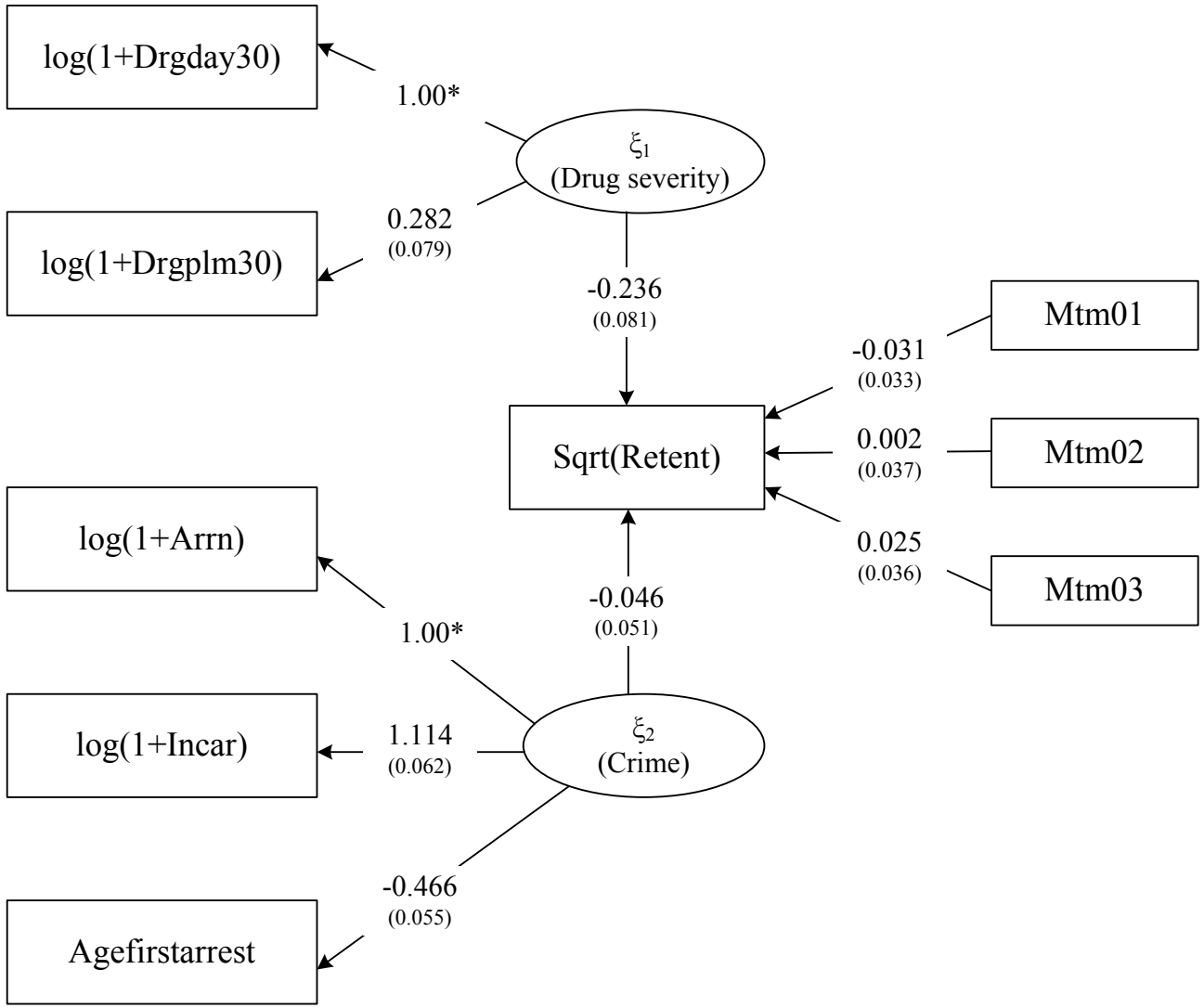
M_9: the explanatory covariates in missing data models are specified as: $\text{logit}\{p(r_{ij}^y = 1|\mathbf{y}_i, \boldsymbol{\varphi}_{ky})\} = \varphi_0^y + \varphi_j^y y_{ij}$, $\text{logit}\{p(r_{ij}^d = 1|\mathbf{d}_i, \boldsymbol{\varphi}_{kd})\} = \varphi_0^d + \varphi_j^d d_{ij}$, $\text{logit}\{p(r_{ij}^x = 1|\mathbf{x}_i, \boldsymbol{\varphi}_x)\} = \varphi_0^x + \varphi_j^x x_{ij}$.

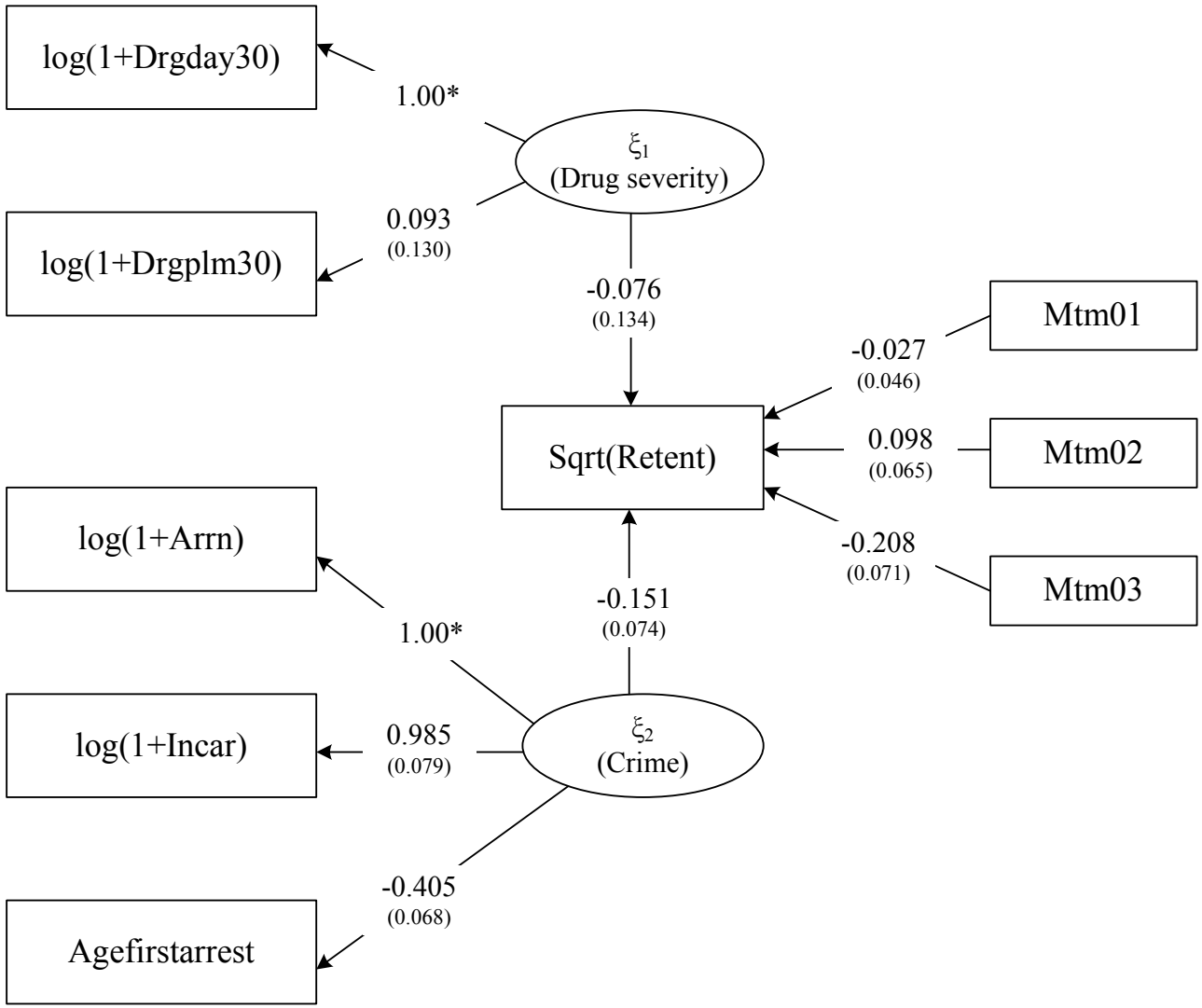
In the above model settings, M_4 to M_9 have the same number of components as the true model but have different missing mechanisms for **y**, **d**, and **x**, respectively. The modified DIC values for M_4 to M_9 were equal to

$$\text{DIC}_{M_4} = 73,458,\ \text{DIC}_{M_5} = 73,162,\ \text{DIC}_{M_6} = 73,131,$$
$$\text{DIC}_{M_7} = 74,001,\ \text{DIC}_{M_8} = 73,088,\ \text{DIC}_{M_9} = 73,862.$$

Again, M_2 with the smallest $\text{DIC}_{M_2} = 73,035$ was selected.

Based on 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 7.1.


log(1+Drgday30)
1.00*
ξ1
(Drug severity)
0.282
(0.079)
log(1+Drgplm30)
-0.236
(0.081)
Mtm01
-0.031
(0.033)
Sqrt(Retent)
0.002
(0.037)
Mtm02
0.025
(0.036)
Mtm03
log(1+Arrn)
-0.046
(0.051)
1.00*
log(1+Incar)
1.114
(0.062)
ξ2
(Crime)
-0.466
(0.055)
Agefirstarrest


log(1+Drgday30)
1.00*
ξ1
(Drug severity)
0.093
(0.130)
log(1+Drgplm30)
-0.076
(0.134)
Mtm01
-0.027
(0.046)
Sqrt(Retent)
0.098
(0.065)
Mtm02
log(1+Arrn)
-0.208
(0.071)
Mtm03
-0.151
(0.074)
1.00*
log(1+Incar)
0.985
(0.079)
ξ2
(Crime)
-0.405
(0.068)
Agefirstarrest

Par.	Component 1 Est	Component 1 SE		Component 2 Est	Component 2 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			
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

Based on the results, the following conclusions can be drawn:

(1) For components 1 and 2, the influences of 'drug severity' and 'crime' on retention have different patterns.

(2) For components 1 and 2, the effects of treatment motivation on retention are different.

(3) There are differences in other parameter estimates in the two components.

(4) 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.

(5) Many parameter estimates in φ_y, φ_d, and φ_x are substantially different from zero, indicating the necessity of the nonignorable mechanisms in the analysis of missing data.

(6) The different estimates of φ_y and φ_d in the two components imply the existence of component-specific patterns in the missing mechanisms for both responses and covariates.