Behaviormetrika
Vol.32, No.1, 2005, 71–93

A LATENT VARIABLE MODEL WITH NON-IGNORABLE MISSING DATA

Takahiro Hoshino*

A latent variable model is proposed that specifies not only the relationship between latent variables, but also the missing mechanism in which the value of the latent variables influences the frequency of missing patterns. We propose an estimation method for our model that adopts the Monte Carlo EM algorithm. Unlike previous methods, our method can be applied when the missing data assumption "Missing at random" does not hold. Moreover, our method can comprehensively explain the missing mechanism using latent variables, and the proposed estimation does not include multiple group estimation, so we can avoid the limitation present in previous studies of the number of subjects in each missing pattern. The proposed model and method are generalized for several kinds of use, such as monotone missingness. We show how to test that the missing mechanism is MAR/MCAR in this model.

We also show the validity of the estimation method in simulation studies of two kinds of missingness (non-ignorable missingness and MAR); we compared the proposed method with ML estimation under the MAR assumption and found it superior.

A read data illustration shows that the proposed method provides a feasible explanation that personality affects the missingness of some questions.

1. Introduction

In behavioral sciences, data can often be missing on several variables. Missingness implies a reduction in the information contained in the data set, but the missingness often contains some information on the population or the parameters of interest (Little & Rubin (1987) called this pattern of missingness "non-ignorable missing").

However, there are few estimation methods for modeling latent variables with non-ignorable missingness. Here, we model a situation in which the missingness of a variable is influenced by the value of the latent variables, using factor analysis and logistic regression models.

Previous studies in this area are based on Rubin's assumption that the data are "missing at random (MAR)". Finkbeiner (1979) proposed a method for the ML estimation of parameters in a factor analysis model under the MAR assumption. Muthén, Kaplan & Hollis (1987) proposed a latent variable model for data with missingness. They factorized the likelihood into the product of a function of the parameters of the data and a function of the parameters of the missing indicator given the data and latent variables. An estimation method that is valid under the MAR assumption was also proposed. Their proposed estimates were obtained by maximizing the former function by regarding the missing patterns as multiple groups. Their model is very attractive (and the proposed model here

Key Words and Phrases: Structural equation modeling, Metropolis-Hastings algorithm, Factor analysis, Bridge sampling

* Department of Interdisciplinary Statistics, Institute of Statistical Mathematics, 4–6–7 Minami-Azabu, Minato-ku, Tokyo, 106–8569, Japan.
E-mail: bayesian@jasmine.ocn.ne.jp

incorporates the advantages of their model). Nevertheless, their interest lies in (1) classifying the pattern of missing data (non-ignorable, MAR, or missing completely at random (MCAR)) in terms of the restrictions imposed in their model, and (2) comparing their estimation with traditional methods valid under MCAR. They did not propose a method of estimating parameters of the missing mechanism. Moreover, their multiple group estimation method is also problematic in that the number of observable variables must not exceed the number of observations in each pattern (see Arminger & Sobel, 1990). Lee (1986) proposed an estimation method for structural equation models with MCAR, but his method also requires multiple group estimation. Arminger & Sobel (1990) proposed an estimation method for structural equation models using the pseudo-maximum likelihood estimation method of Gourieroux, Monfort & Trognon (1984). Their method has some advantageous properties; for example, it avoids an assumption on distribution. However, this method still requires the MAR assumption for missing patterns. O'Muircheartaigh & Moustaki (1999) proposed a latent variable model in which the missingness depends on the missingness propensity latent factor.

Our aim is to propose a latent variable model for when the missingness is not random and the mechanism for the missing data is predicted by latent variables, such as the models of Muthén et al. or O'Muircheartaigh & Moustaki (1999). We propose an estimation method not only for parameters for the relationship between data and latent variables, but also for parameters related to the pattern of missing data and latent variables, using the Monte Carlo EM estimation proposed by Wei & Tanner (1990).

Section 2 describes the model assumptions of the data generating mechanism and the missing mechanism in detail. Some of the properties of the proposed method are also described. For example, our method is shown to be able to deal with monotone missing patterns without any modifications. In Section 3, we apply the Monte Carlo EM estimation method to the proposed model, because the expectation with respect to the distribution of the latent variables is not implemented analytically. Extensions of this model are proposed in Section 4, such as the hypothetical testing of the ignorability of the missing mechanism in the model. In Section 5, we perform simulation studies to show the validity of the proposed estimation method and compare the proposed method with the method proposed by Finkbeiner (1979). We apply the proposed method to real data and obtain a meaningful interpretation from the result in Section 6 and make our concluding remarks in the last section.

2. Model

Let $\boldsymbol{x}_i$ $(J \times 1)$ be a complete-data vector, and let $\boldsymbol{x}_{obs_i}$ $(J_i \times 1)$ be an observed-data vector for the i-th subject. Let $\boldsymbol{m}_i$ $(J \times 1)$ be a missing indicator vector for the i-th subject. The j-th element of $\boldsymbol{m}_i$, m_{ij} is assigned to 1 if the j-th element of $\boldsymbol{x}_i$, x_{ij} is observed for the i-th subject; otherwise m_{ij} is assigned to 0.

We assume that $\boldsymbol{x}_i$ follows the factor analysis model for measuring a vector of p-dimensional latent variables $\boldsymbol{f}_i$:

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda f}_i + \boldsymbol{\Pi y}_i + \boldsymbol{e}_i, \tag{1}$$

where $\boldsymbol{y}_i$ is the covariate vector for the i-th subject (following Muthén's LISCOMP model, Muthén, 1983, 1984), and $\boldsymbol{f}_i$ and $\boldsymbol{e}_i$ follow the multivariate normal distributions:

$$\boldsymbol{f}_i \sim N(\boldsymbol{C y}_i, \boldsymbol{\Phi}), \quad \boldsymbol{e}_i \sim N(0, \boldsymbol{\Psi}), \tag{2}$$

respectively. We also assume that $\boldsymbol{f}$ given that $\boldsymbol{y}$ and $\boldsymbol{e}$ are uncorrelated.

By introducing a latent variable vector $\boldsymbol{f}_i$, the likelihood L is written as

$$L = \prod_i^N p(\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i | \boldsymbol{y}_i, \boldsymbol{\vartheta}, \boldsymbol{\varphi}) = \prod_i^N \int\int p(\boldsymbol{x}_i, \boldsymbol{f}_i | \boldsymbol{y}_i, \boldsymbol{\vartheta}) p(\boldsymbol{m}_i | \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{f}_i, \boldsymbol{\varphi}) d\boldsymbol{f}_i d\boldsymbol{x}_{miss_i}, \tag{3}$$

where $\boldsymbol{\vartheta}$ is the vector of the parameters with respect to the joint distribution of the complete-data vector and the latent variable vector, and $\boldsymbol{\varphi}$ is the parameter vector related to the missing mechanism (Little and Rubin, 1987).

We assume that the distribution of the missing indicator vector $\boldsymbol{m}_i$ with $\boldsymbol{x}_i$ and $\boldsymbol{f}_i$ is

$$p(\boldsymbol{m}_i | \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{f}_i, \boldsymbol{\varphi}) = p(\boldsymbol{m}_i | \boldsymbol{x}_{obs_i}, \boldsymbol{y}_i, \boldsymbol{f}_i, \boldsymbol{\varphi}), \tag{4}$$

This implies that the latent variable vector and observed portion of the data contain all the information that the complete-data vector contains about the parameters of the missing mechanism (Eqn. (4)). This assumption is made by O'Muircheartaigh & Moustaki (1999), while their model uses a latent variable that influences only the missingness (see Section 4 [1]).

This assumption is justified because the informative portion of the complete-data vector is expected to concentrate on the latent variable vector in latent variable models, and because the distribution of the data with the latent variables given is equivalent to the error distribution. In the latent variable models, the values of the error variables are not assumed to contain any information.

Following Little & Rubin (1987), non-ignorable missingness differs from MAR in that the missingness is explained not only by the observed portion of the complete data vector but also by the missing portion of the complete data vector. In the factor analysis model, we assume that factors predict both the observed and missing portions of the complete data vector (i.e., this equals the complete data vector). If factors predict the complete data vector, then factors can also predict the non-ignorable missingness. (We will introduce an extended version of this assumption later in this section.)

It is also obvious that $\boldsymbol{x}_{obs}$ is not necessary for predicting the missing mechanism if the model fit of the factor analysis model (Eqn. (1)) is good, then $\boldsymbol{x}$ is predicted well by $\boldsymbol{f}$ and $\boldsymbol{y}$, and the correlations between factors and observables will be large.

Under Eqn. (4), the likelihood (Eqn. (3)) can be expressed as follows:

$$\begin{aligned} L &= \prod_i^N p(\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i | \boldsymbol{y}_i, \boldsymbol{\vartheta}, \boldsymbol{\varphi}) \\ &= \prod_i^N \int\int p(\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{f}_i, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Pi}, \boldsymbol{\Psi}) p(\boldsymbol{m}_i | \boldsymbol{x}_{obs_i}, \boldsymbol{y}_i, \boldsymbol{f}_i, \boldsymbol{\varphi}) p(\boldsymbol{f}_i | \boldsymbol{y}_i, C, \Phi) d\boldsymbol{f}_i d\boldsymbol{x}_{miss_i} \end{aligned}$$

$$= \prod_{i}^{N} \int p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_i, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Pi\Psi}) p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{f}_i, \varphi) p(\boldsymbol{f}_i|\boldsymbol{y}_i, C, \Phi) d\boldsymbol{f}_i. \tag{5}$$

Note that for the subjects with missingness for all variables $p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_i, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Pi}, \boldsymbol{\Psi}) = 1$, hence the information about the number of those subjects contributes not only to the estimation of $\boldsymbol{\varphi}$ but also to that of $\boldsymbol{\vartheta}$ through the latent variable vector f.

Let $\boldsymbol{K}_i$ ($l_i \times l$ dim) be a matrix, such that $\boldsymbol{x}_{obs_i} = \boldsymbol{K}_i \boldsymbol{x}_i$. Using the $\boldsymbol{K}_i$ matrix, it is easily shown that the distribution of $\boldsymbol{x}_{obs_i}$ with $\boldsymbol{f}_i, \boldsymbol{y}_i$, and $\boldsymbol{m}_i$ given, $p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_i, \boldsymbol{y}_i, \boldsymbol{\Lambda}, \boldsymbol{\Pi}, \boldsymbol{\Psi})$ is the J_i-dimensional multivariate normal distribution:

$$\boldsymbol{x}_{obs_i}|\boldsymbol{f}_i, \boldsymbol{y}_i, \boldsymbol{\Lambda}, \boldsymbol{\Pi}, \boldsymbol{\Psi} \sim N(\boldsymbol{K}_i(\boldsymbol{\Lambda f}_i + \boldsymbol{\Pi y}_i + \boldsymbol{\mu}), \boldsymbol{K}_i \boldsymbol{\Psi} \boldsymbol{K}_i^t). \tag{6}$$

For the missing indicator variable m_{ij}, we assume the logistic regression model in which the explanatory variable is $\boldsymbol{f}_i$,

$$p(\boldsymbol{m}_i|\boldsymbol{f}_i, \boldsymbol{y}_i) = \prod_{j=1}^{J} \left(\frac{exp(\alpha_j + \boldsymbol{f}_i^t \boldsymbol{\beta}_j + \boldsymbol{y}_i^t \gamma_j + \boldsymbol{x}_{obs_i}^t \boldsymbol{K}_i \boldsymbol{\omega}_j)}{1 + exp(\alpha_j + \boldsymbol{f}_i^t \boldsymbol{\beta}_j + \boldsymbol{y}_i^t \gamma_j + \boldsymbol{x}_{obs_i}^t \boldsymbol{K}_i \boldsymbol{\omega}_j)} \right)^{m_{ij}} \left(\frac{1}{1 + exp(\alpha_j + \boldsymbol{f}_i^t \boldsymbol{\beta}_j + \boldsymbol{y}_i^t \gamma_j + \boldsymbol{x}_{obs_i}^t \boldsymbol{K}_i \boldsymbol{\omega}_j)} \right)^{1-m_{ij}}, \tag{7}$$

where $\boldsymbol{m}_i = (m_{i1}, m_{i2}, \cdots, m_{iJ})$ and $\boldsymbol{\omega}_j$ contains elements of regression coefficients on the missing indicator of x_j. For example, ω_{13} is the regression coefficient of the observed portion of x_1 on the missing indicator of x_3.

In this model, the k-th element of $\boldsymbol{\beta}_j$, β_{jk} represents the magnitude of the influence of the k-th element of the latent variable f_k on the frequency of missingness on the j-th element of x, x_j. We also define $\boldsymbol{\gamma}$ as the coefficient vector for covariate y. Each α_j represents the baseline of the frequency of missing data on x_j.

The model of Muthén et al. adopts the Probit model for the missing indicator variables. We adopt the logistic regression model to take advantage of the logical connection to the generalized linear model (see Moustaki & Knott, 2000). The logistic model has some advantageous properties for extending the model (see Section 4).

Some properties of this model are as follows:

1 The proposed method is free from the relatively stronger restriction in multiple group estimation that the number of subjects must exceed the number of observable variables in each missing pattern.

2 This model is easily extended to structural equation models by modeling the latent variable vector. For example, the LISREL model can be introduced by letting,

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda_\eta} & 0 \\ 0 & \boldsymbol{\Lambda_\xi} \end{pmatrix}, \quad f = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{pmatrix} \quad and \quad \boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{8}$$

where $\boldsymbol{\zeta}$ follows the multivariate normal distribution with mean 0 and the diagonal covariance matrix.

3 We can easily deal with categorical variables in this model. Let $\boldsymbol{c}_i$ be the categorical variable vector of the i-th person. If the j-th variable of $\boldsymbol{c}_i$, c_{ij} has K_j categories, we define the new categorical variable c^*_{ij} by regarding missing categories as the K_j+1-th category. Therefore, we express the joint distribution of $x_{obi}, \boldsymbol{m}_i$ and c^*_{ij} with $\boldsymbol{y}_i$ given as follows:

$$p(\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, c^*_{ij}|\boldsymbol{y}_i) = \int p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{f}_i, \boldsymbol{y}_i)p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_i, \boldsymbol{y}_i)p(c^*_{ij}|\boldsymbol{f}_i, \boldsymbol{y}_i)p(\boldsymbol{f}_i|\boldsymbol{y}_i)d\boldsymbol{f}_i. \tag{9}$$

If the original categorical variable c_{ij} is nominal, Bock's nominal response model (1972) can be applied to the distribution of $p(c^*_{ij}|\boldsymbol{f}_i, \boldsymbol{y}_i)$. If the original c_{ij} is ordinal, the ordinal logistic regression model (McCullagh 1980) is applied. Instead of defining the new variable c^*, we can simply introduce the missing indicator mc such that $mc_{ij} = 1$ if c_{ij} is observed in the i-th subject; otherwise $mc_{ij} = 0$. If c_{ij} is not observed, then $p(c_{ij}|\boldsymbol{f}_i, \boldsymbol{y}_i)$ is not included in the likelihood.

4 We can apply this model to monotone missing patterns directly. Let the missing pattern on a certain set of variable vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ be monotone, i.e., whenever $\boldsymbol{x}_i$ is missing, $\boldsymbol{x}_j$ is also missing for all $i < j$ (for $i, j = 1, \cdots, p$: see, e.g., Schafer, 1997). To deal with monotone missingness, let $mc_{1\cdots p}$ be the missing indicator variable with $p+1$ categories such that $mc_{1-p} = q+1$ if the variable vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ are observed up to $\boldsymbol{x}_q$. If data are missing for all variables, then $mc_{1-p} = 1$. From the notation above, the ordinal logistic regression model is also applicable to the monotone missingness by regarding mc_{1-p} as an ordinal categorical variable.

3. Inference method—MCEM algorithm

In the proposed model, it is very difficult to estimate all of the parameters for the data generating model and for the missing mechanism. Standard latent variable models with both continuous and ordered categorical variables can be estimated using multistage estimation procedures (for example, Muthén 1985; Lee, Poon & Bentler, 1992). However, this kind of estimation strategy cannot be applied to our model. Therefore we use the EM algorithm for this problem.

In the EM algorithm, the expectation of the logarithm of the complete likelihood $p(\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{f}|\boldsymbol{y})$ with respect to $p(\boldsymbol{f}|\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{y})$ is calculated (E-step), and is maximized with respect to the parameters (M-step) (Dempster, Laird & Rubin, 1977).

To estimate $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\varphi})$ in this model, the expectation $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^r)$ that we should maximize in the r-th iteration of the EM algorithm is,

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^r) &= E_{\boldsymbol{\theta}^r}[\log p(\boldsymbol{m}, \boldsymbol{x}_{obs}, \boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{m}, \boldsymbol{x}_{obs}] \\ &= \int \log p(\boldsymbol{m}, \boldsymbol{x}_{obs}, \boldsymbol{f}|\boldsymbol{y}, \boldsymbol{\theta})p(\boldsymbol{f}|\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{y}, \boldsymbol{\theta}^r)d\boldsymbol{f}. \end{aligned} \tag{10}$$

This cannot be calculated analytically, so we use the Monte Carlo EM (MCEM) algorithm (Wei & Tanner, 1990) to estimate the parameters of this model. In the MCEM method, the expectation of E-step is approximated by the Monte Carlo mean:

$$\hat{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^r)=\frac{1}{L}\sum_{l=1}^{L}\log p(\boldsymbol{m},\boldsymbol{x}_{obs},\boldsymbol{f}^l|\boldsymbol{y},\boldsymbol{\theta}), \tag{11}$$

where $\boldsymbol{f}^l$ is the l-th random sample from $p(\boldsymbol{f}|\boldsymbol{x}_{obs},\boldsymbol{m},\boldsymbol{y},\boldsymbol{\theta}^r)$. The maximum θ of $\hat{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^r)$ in each iteration is found using the Newton-Raphson algorithm. Let $\boldsymbol{\theta}^{r_t}$ be the t-th value of $\boldsymbol{\theta}$ in the r-th M-step. Then, the equation for updating is as follows:

$$\boldsymbol{\theta}^{r_{t+1}}=\boldsymbol{\theta}^{r_t}-a^t\times\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\hat{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{r_t})\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{r_t}}\times\left[\frac{\partial}{\partial\boldsymbol{\theta}}\hat{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{r_t})\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{r_t}}, \tag{12}$$

until the convergence criterion is satisfied (where a^t is a number less than 1), and let θ^r be the convergent value of $\boldsymbol{\theta}$ in the r-th M-step.

For notational simplicity, let $\boldsymbol{\theta}_M$ be the parameter vector of missing mechanism, $\boldsymbol{\varphi}$. Let also $\boldsymbol{\theta}_X$ be the parameter vector of the distinct elements in $\boldsymbol{\Lambda}$,$\boldsymbol{\Pi}$,$\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$, and let $\boldsymbol{\theta}_F$ be the vector of the distinct elements in $\boldsymbol{C}$ and $\boldsymbol{\Phi}$.

Then, the complete log-likelihood is divided as follows:

$$\log p(\boldsymbol{m},\boldsymbol{x}_{obs},\boldsymbol{f}|\boldsymbol{y})=\log p(\boldsymbol{m}|\boldsymbol{x}_{obs},\boldsymbol{f},\boldsymbol{y},\boldsymbol{\theta}_M)+\log p(\boldsymbol{x}_{obs}|\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}_X)+\log p(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{\theta}_F), \tag{13}$$

Using the notation given below,

$$E_M=\int\log p(\boldsymbol{m}|\boldsymbol{x}_{obs},\boldsymbol{f},\boldsymbol{y},\boldsymbol{\theta}_M^k)p(\boldsymbol{f}|\boldsymbol{x}_{obs},\boldsymbol{m},\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{f}, \tag{14}$$

$$E_X=\int\log p(\boldsymbol{x}_{obs}|\boldsymbol{y},\boldsymbol{f},\boldsymbol{\theta}_X^k)p(\boldsymbol{f}|\boldsymbol{x}_{obs},\boldsymbol{m},\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{f}, \tag{15}$$

$$and\quad E_F=\int\log p(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{\theta}_F^k)p(\boldsymbol{f}|\boldsymbol{x}_{obs},\boldsymbol{m},\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{f}, \tag{16}$$

the expectation E can be written as, $Q(\boldsymbol{\theta},\boldsymbol{\theta}^k)=E_M+E_X+E_F$

[1] M-step

At step M, we need to maximize $Q(\boldsymbol{\theta},\boldsymbol{\theta}^r)$ with respect to $\boldsymbol{\theta}$.

$$\frac{\partial}{\partial\boldsymbol{\theta}_A}Q(\boldsymbol{\theta},\boldsymbol{\theta}^r)=\frac{\partial E_A}{\partial\boldsymbol{\theta}_A}=\int\frac{\partial l_A}{\partial\boldsymbol{\theta}_A}p(\boldsymbol{f}|\boldsymbol{x}_{obs},\boldsymbol{m},\boldsymbol{\theta}^r)d\boldsymbol{f}\quad(A=M,X\ or\ F). \tag{17}$$

$$\frac{\partial^2}{\partial\boldsymbol{\theta}_A\partial\boldsymbol{\theta}_B^t}Q(\boldsymbol{\theta},\boldsymbol{\theta}^r)=0\quad if\ A\neq B.$$
$$=\frac{\partial^2 E_A}{\partial\boldsymbol{\theta}_A\partial\boldsymbol{\theta}_A^t}=\int\frac{\partial^2 l_A}{\partial\boldsymbol{\theta}_A\partial\boldsymbol{\theta}_A^t}p(\boldsymbol{f}|\boldsymbol{x}_{obs},\boldsymbol{m},\boldsymbol{y},\boldsymbol{\theta}^r)d\boldsymbol{f}\quad if\ A{=}B. \tag{18}$$

Therefore, we can divide step M into three parts E_M, E_X, and E_F and maximize them separately. Shi & Lee (2000) proposed the MCEM estimation for latent variable models with continuous and polytomous variables, in which the expectation is divided in the same way. However, our model requires the Newton-Raphson iteration in each step M because of the complexity of the model.

1. E_M
   The maximization method for parameters is basically equivalent to the inference method in the logistic regression model.
   Then, for each j, the first and second partial derivatives of l_M with respect to $\boldsymbol{\beta}_j^*$ are as follows:
$$\frac{\partial l_M}{\partial \boldsymbol{\beta}_j^*} = \boldsymbol{F}^{*t}(\boldsymbol{m}_j - \boldsymbol{p}_j) \tag{19}$$
$$\frac{\partial^2 l_M}{\partial \boldsymbol{\beta}_j^* \partial \boldsymbol{\beta}_j^{*t}} = \boldsymbol{F}^{*t}\boldsymbol{W}_j\boldsymbol{F}^* \tag{20}$$
   where $\boldsymbol{p}_j = (p_{1j}, \cdots p_{Nj})^t$, $\boldsymbol{\beta}_j^{*t} = [\boldsymbol{\beta}_j^t, \boldsymbol{\gamma}_j^t, \boldsymbol{\omega}_j^t, \alpha_j]$, $\boldsymbol{f}_i^* = (\boldsymbol{f}_i^t, \boldsymbol{y}_i^t, \boldsymbol{x}_{obs_i}^t\boldsymbol{K}_i, 1)^t$, $\boldsymbol{F}^* = (\boldsymbol{f}_1^*, \cdots, \boldsymbol{f}_N^*)^t$, $\boldsymbol{m}_j = (m_{1j}, \cdots, m_{Nj})^t$, and $\boldsymbol{W}_j$ is a diagonal matrix with the i-th diagonal element $(p_{ij}(1-p_{ij}))$.
2. E_X
   For maximization, we use the scoring method. The scoring method requires the first and second differentials given below (vec operator: see, e.g., Magnus & Neudecker 1988):
$$\frac{\partial l_X}{\partial vec(\boldsymbol{\Lambda}^*)} = \sum_{i=1}^{N} vec(\boldsymbol{K}_i^t\boldsymbol{A}_i(\boldsymbol{x}_{obs_i} - \boldsymbol{K}_i\boldsymbol{\Lambda}^*\boldsymbol{f}_i^*)\boldsymbol{f}_i^{*t}). \tag{21}$$
$$\frac{\partial l_X}{\partial vec(\boldsymbol{\Psi})} = \frac{1}{2}\sum_{i=1}^{N} vec(\boldsymbol{K}_i^t(\boldsymbol{I} - \boldsymbol{A}_i\boldsymbol{S}_i)\boldsymbol{A}_i\boldsymbol{K}_i). \tag{22}$$
$$\frac{\partial^2 l_X}{\partial vec(\boldsymbol{\Lambda}^*)\partial(vec(\boldsymbol{\Lambda}^*))^t} = \sum_{i=1}^{N}(\boldsymbol{f}_i^*\boldsymbol{f}_i^{*t}) \otimes (\boldsymbol{K}_i^t\boldsymbol{A}_i\boldsymbol{K}_i). \tag{23}$$
$$\frac{\partial^2 l_X}{\partial vec(\boldsymbol{\Psi})\partial(vec(\boldsymbol{\Psi}))^t} = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{K}_i^t\boldsymbol{A}_i\boldsymbol{K}_i) \otimes (\boldsymbol{K}_i^t(2\boldsymbol{A}_i\boldsymbol{S}_i - \boldsymbol{I})\boldsymbol{A}_i\boldsymbol{K}_i). \tag{24}$$
$$\frac{\partial^2 l_X}{\partial vec(\boldsymbol{\Lambda}^*)\partial(vec(\boldsymbol{\Psi}))^t} = \sum_{i=1}^{N}[\boldsymbol{f}_i^*(\boldsymbol{K}_i\boldsymbol{\Lambda}^*\boldsymbol{f}_i^* - \boldsymbol{x}_{obs_i})^t \otimes \boldsymbol{K}_i^t][\boldsymbol{A}_i\boldsymbol{K}_i \otimes \boldsymbol{A}_i\boldsymbol{K}_i]. \tag{25}$$
   where $\boldsymbol{S}_i = (\boldsymbol{K}_i\boldsymbol{\Lambda}^*\boldsymbol{f}_i^* - \boldsymbol{x}_{obs_i})(\boldsymbol{K}_i\boldsymbol{\Lambda}^*\boldsymbol{f}_i^* - \boldsymbol{x}_{obs_i})^t$, $\boldsymbol{A}_i = (\boldsymbol{K}_i\boldsymbol{\Psi}\boldsymbol{K}_i^t)^{-1}$, *and* $\boldsymbol{\Lambda}^* = (\boldsymbol{\Lambda}, \boldsymbol{\Pi}, \boldsymbol{\mu})$.
   Note that information on the subjects with missing data for all variables does not contribute to the estimation of $\boldsymbol{\theta}_X$, so such subjects ($l_i = 0$) are not included in the summation above.
3. E_F
   Maximization is equivalent to the estimation of the multivariate regression model under the assumption that the error vector is normally distributed. Therefore, we can calculate the estimates of $\boldsymbol{C}$ and $\boldsymbol{\Psi}$ using the closed form:
$$\hat{\boldsymbol{C}}^r = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{F}_l^t\boldsymbol{Y}(\boldsymbol{Y}^t\boldsymbol{Y})^{-1}.\hat{\boldsymbol{\Phi}}^r = \frac{1}{L \times N}\sum_{l=1}^{L}(\boldsymbol{F}_l^t - \boldsymbol{Y}\hat{C}^{r^t})^t(\boldsymbol{F}_l^t - \boldsymbol{Y}\hat{\boldsymbol{C}}^{r^t}). \tag{26}$$

where $\boldsymbol{F} = (\boldsymbol{f}_1^t, \cdots, \boldsymbol{f}_N^t)^t, \boldsymbol{Y} = (\boldsymbol{y}_1^t, \cdots, \boldsymbol{y}_N^t)^t$ and $\boldsymbol{F}_l$ is the factor score matrix using the l-th random sample. Note that the first and second differentials are also needed to calculate the Fisher information.

[2] Fisher Information and Wald Statistics
The method used to calculate the Fisher information from the EM algorithm is given by Louis (1982). By letting $I_c(\boldsymbol{\theta} : \boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{f}|\boldsymbol{y})$ be the matrix of the negative of the second-order partial derivatives of the complete-data log likelihood function with respect to the elements of $\boldsymbol{\theta}$, and $\boldsymbol{S}_c(\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{f}|\boldsymbol{y} : \hat{\boldsymbol{\theta}})$ be the gradient vector of the complete-data log likelihood function, the estimate of the Fisher information matrix of the observed-data likelihood is,

$$\begin{aligned}\boldsymbol{I}(\hat{\boldsymbol{\theta}} : \boldsymbol{x}_{obs}, \boldsymbol{m}) = & E_{\boldsymbol{\theta}}[\boldsymbol{I}_c(\boldsymbol{\theta} : x_{obs}, \boldsymbol{m}, |\boldsymbol{y})|\boldsymbol{x}_{\mathbf{obs}}, \boldsymbol{m}, \boldsymbol{y}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ & -E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{f}|\boldsymbol{y} : \hat{\boldsymbol{\theta}})\boldsymbol{S}_c^t(\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{f}|\boldsymbol{y} : \hat{\boldsymbol{\theta}})|\boldsymbol{x}_{obs}, \boldsymbol{m}, \boldsymbol{y}]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \end{aligned} \quad (27)$$

In this model, $\boldsymbol{S}_c$ and $\boldsymbol{I}_c$ are expressed as follows:

$$\boldsymbol{S}_c = \left(\frac{\partial l_M}{\partial \boldsymbol{\beta}^{*t}} \quad \frac{\partial l_X}{\partial (vec(\boldsymbol{\Lambda}^*))^t} \quad \frac{\partial l_X}{\partial (vec(\boldsymbol{\Psi}))^t} \quad \frac{\partial l_F}{\partial (vec(\boldsymbol{\Phi}))^t} \right)^t \quad (28)$$

$$I_c = -\begin{pmatrix} \frac{\partial^2 l_M}{\partial vec(\boldsymbol{\beta}^*)\partial(vec(\boldsymbol{\beta}^*))^t} & 0 & 0 & 0 \\ 0 & \frac{\partial^2 l_X}{\partial vec(\boldsymbol{\Lambda}^*)\partial(vec(\boldsymbol{\Lambda}^*))^t} & \frac{\partial l_X}{\partial vec(\boldsymbol{\Lambda}^*)\partial(vec(\boldsymbol{\Psi}))^t} & 0 \\ 0 & \frac{\partial l_X}{\partial(vec(\boldsymbol{\Lambda}^*))^t\partial(vec(\boldsymbol{\Psi}))} & \frac{\partial^2 l_X}{\partial vec(\boldsymbol{\Psi})\partial(vec(\boldsymbol{\Psi}))^t} & 0 \\ 0 & 0 & 0 & \frac{\partial^2 l_F}{\partial vec(\boldsymbol{\Phi})\partial(vec(\boldsymbol{\Phi}))^t} \end{pmatrix} \quad (29)$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*t}, \cdots, \boldsymbol{\beta}_J^{*t})^t$.

[3] Sampling from the distribution $p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i)$
It is necessary to calculate the expectations E_M, E_X and E_F to estimate the parameters. To do this, we must draw samples from $p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i)$. Since it is difficult to draw samples from this distribution directly, we use the Independence sampler (Tierney, 1994), which is a variant of the Metropolis-Hastings algorithm (see, e.g., Chen, Shao & Ibrahim, 2000).
To apply the Independence sampler algorithm to this estimation, we let the proposal distribution be $p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)$. There are two main reasons for this. First, $p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)$ is expected to approximate the target distribution $p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i)$ well, so the acceptance rate is also expected to be large. The other reason will be mentioned later.
Let $\boldsymbol{f}_i^*$ be the candidate of the $l+1$-th random sample of $\boldsymbol{f}_i$, $\boldsymbol{f}_i^{l+1}$. Then, $\boldsymbol{f}_i^*$ is accepted as $\boldsymbol{f}_i^{k+1}$ with probability δ:

$$\delta = min\left(1, \frac{p(\boldsymbol{f}_i^*|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i)p(\boldsymbol{f}_i^l|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)}{p(\boldsymbol{f}_i^*|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)p(\boldsymbol{f}_i^l|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i)}\right), \quad (30)$$

and $\boldsymbol{f}_i^{l+1} = \boldsymbol{f}_i^l$ with probability $1-\delta$.

From Eqn. (4), the following equation holds:

$$p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i) = \frac{p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{f}_i, \boldsymbol{y}_i)p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)}{p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)}. \tag{31}$$

Note that this density is not calculated analytically, because the denominator $p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i) = \int p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{f}_i, \boldsymbol{y}_i)p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i)d\boldsymbol{f}_i$. cannot be calculated.
Then, the probability δ can be rewritten as,

$$\delta = min\left(1, \frac{p(\boldsymbol{m}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{f}_i^*, \boldsymbol{y}_i)}{p(\boldsymbol{m}_i|\boldsymbol{f}_i^k, \boldsymbol{y}_i)}\right). \tag{32}$$

Note that in some simpler models, we can use the normal distribution in which the mean is the mode of the target distribution for the proposal distribution (see, e.g., Chen, Shao & Ibrahim, 2000). However, this strategy is inefficient, because both sampling from and calculating (of the mode) the target distribution are difficult in this model.
The biggest advantage of our proposal distribution is that the acceptance rate δ does not depend on $p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i, \boldsymbol{y}_i)$ (see, Eqn. (32)).
Using Bayes theorem, the distribution of $\boldsymbol{f}_i$ given $\boldsymbol{x}_{obs_i}$ and $\boldsymbol{y}_i$ can be expressed as follows:

$$p(\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i) = \frac{p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_i, \boldsymbol{y}_i)p(\boldsymbol{f}_i|\boldsymbol{y}_i)}{p(\boldsymbol{x}_{obs_i}|\boldsymbol{y}_i)}, \tag{33}$$

Then, we get the distribution of $\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i$ analytically,

$$\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i \sim N(\Sigma_{If}\boldsymbol{\Lambda}\boldsymbol{K}_i A_i(\boldsymbol{x}_{obs_i} - \boldsymbol{K}_i(\boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{y}_i)), \Sigma_{If}), \tag{34}$$

where $\Sigma_{If}^{-1} = \boldsymbol{\Lambda}\boldsymbol{K}_i^t\boldsymbol{A}_i\boldsymbol{K}_i\boldsymbol{\Lambda} + \boldsymbol{\Phi}^{-1}$, if l_i is positive, and

$$\boldsymbol{f}_i|\boldsymbol{x}_{obs_i}, \boldsymbol{y}_i \sim N(\boldsymbol{C}\boldsymbol{y}_i, \boldsymbol{\Phi}), \tag{35}$$

if l_i is 0.

[4] Determining the convergence of MCEM via Bridge-Sampling
Unlike standard implementations of EM, it is generally difficult to show convergence of the MCEM algorithm, because of the simulation variability introduced at its E-step. Wei & Tanner (1990) recommended plotting parameter values against iteration. As the proposed model contains several parameters, a practical alternative is to plot some function of the parameter values. In this MCEM algorithm, following Ming & Schilling (1996), we can use the estimate of the difference of the consecutive log-likelihood values obtained by the bridge sampling method proposed by Meng & Wong (1996) to monitor convergence.
The difference of the log-likelihood $g_K(\boldsymbol{\theta}^{r+1}, \boldsymbol{\theta}^r)$ is estimated by the identity:

$$\hat{g_K}(\boldsymbol{\theta}^{r+1}, \boldsymbol{\theta}^r) = \sum_{i=1}^{N} \log\left\{\frac{E_r[p(\boldsymbol{m}_i, \boldsymbol{x}_{obs_i}, \boldsymbol{f}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{r+1})\alpha_r(\boldsymbol{f}_i)]}{E_{r+1}[p(\boldsymbol{m}_i, \boldsymbol{x}_{obs_i}, \boldsymbol{f}_i|\boldsymbol{y}_i, \theta^r)\alpha_r(\boldsymbol{f}_i)]}\right\}, \tag{36}$$

where E_r denotes the expectation with respect to $p(\boldsymbol{f}_i|\boldsymbol{m}_i, \boldsymbol{x}_{obs_i}, \boldsymbol{y}_i, \boldsymbol{\theta}^r)$, and $\alpha_r(\boldsymbol{f}_i)$

can be chosen to satisfy some mild conditions arbitrarily (Meng & Wong, 1996). For reasons discussed by Meng & Wong (1996), we use

$$\alpha_r(\boldsymbol{f}_i) = \left[\sqrt{p(\boldsymbol{m}_i, \boldsymbol{x}_{obs_i}, \boldsymbol{f}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^r)p(\boldsymbol{m}_i, \boldsymbol{x}_{obs_i}, \boldsymbol{f}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{r+1})}\right]^{-1}. \tag{37}$$

The expectation E_r can be approximated easily by draws of factors that are obtained in the E-step.

Refer to Meng & Wong (1996) for more details of this general method.

4. Some generalizations

The proposed model can be extended to more general models using some intuitive and simple modifications. In this section, we refer to three extensions.

[1] Inclusion of factor f_m, which influences the missing mechanism only

Let the latent variable vector $\boldsymbol{f}_i$ be divided into two subvectors: $\boldsymbol{f}_{\boldsymbol{m}_i}$ and $\boldsymbol{f}_{\boldsymbol{x}_i}$. If $\boldsymbol{f}_{\boldsymbol{m}_i}$ is the latent variable vector that influences the missing mechanism only, then the joint distribution of $\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i$ can be expressed as follows:

$$\begin{aligned} p(\boldsymbol{x}_{obs_i}, \boldsymbol{m}_i|\boldsymbol{y}_i) &= \iint p(\boldsymbol{x}_{obs_i}, \boldsymbol{f}_{\boldsymbol{x}_i}, \boldsymbol{f}_{\boldsymbol{m}_i}, \boldsymbol{m}_i|\boldsymbol{y}_i) d\boldsymbol{f}_{\boldsymbol{x}_i} d\boldsymbol{f}_{\boldsymbol{m}_i} \\ &= \iint p(\boldsymbol{m}_i|\boldsymbol{f}_{\boldsymbol{x}_i}, \boldsymbol{f}_{\boldsymbol{m}_i}, \boldsymbol{y}_i) p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_{\boldsymbol{x}_i}, \boldsymbol{f}_{\boldsymbol{m}_i}, \boldsymbol{y}_i) p(\boldsymbol{f}_{\boldsymbol{x}_i}, \boldsymbol{f}_{\boldsymbol{m}_i}|\boldsymbol{y}_i) d\boldsymbol{f}_{\boldsymbol{x}_i} d\boldsymbol{f}_{\boldsymbol{m}_i} \\ &= \iint p(\boldsymbol{m}_i|\boldsymbol{f}_{\boldsymbol{m}_i}, \boldsymbol{y}_i) p(\boldsymbol{x}_{obs_i}|\boldsymbol{f}_{\boldsymbol{x}_i}, \boldsymbol{y}_i) p(\boldsymbol{f}_{\boldsymbol{x}_i}, \boldsymbol{f}_{\boldsymbol{m}_i}|\boldsymbol{y}_i) d\boldsymbol{f}_{\boldsymbol{x}_i} d\boldsymbol{f}_{\boldsymbol{m}_i}. \end{aligned} \tag{38}$$

This model specification can be done easily by simply setting the related elements of $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ at zero.

[2] Hypothesis testing of MAR/MCAR.

If "$H_0 : \boldsymbol{\beta}_j = 0$" holds, the missing patterns in the j-th data variable x_j can be regarded as ignorable, because under the hypothesis, $p(x_{obs_j}, m_j|y)$ becomes the product of $p(x_{obs_j}|y)$ and $p(m_j|x_{obs_j}, y)$, therefore, $p(m_j|x_{obs_j}, y)$; therefore, in the proposed model, this hypothesis (Ignorability) can be tested by the Wald statistic. The Wald statistic of the hypothesis concerning MAR $H_0 : \boldsymbol{\beta}_j = 0$ is,

$$W^j = \hat{\boldsymbol{\beta}}_j^{\,t} \left[I^{\boldsymbol{\beta}_j}(\hat{\theta} : \boldsymbol{x}_{obs}, m, y)\right]^{-1} \hat{\boldsymbol{\beta}}_j, \tag{39}$$

where $I^{\boldsymbol{\beta}_j}(\hat{\theta} : \boldsymbol{x}_{obs}, m, y)$ is the portion of the Fisher information related to $\boldsymbol{\beta}_j$. We can also test the hypothesis of MAR on all variables by replacing $\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}$ in the Wald statistic.

We can also implement hypothetical testing of whether the MCAR assumption holds, because the hypothesis in this model is expressed as $H_0 : \boldsymbol{\beta} = 0,\ \boldsymbol{\omega} = 0$.

[3] Missing patterns when some variables are missing simultaneously.

It is also easy for our model to express the situation in which some of the variables

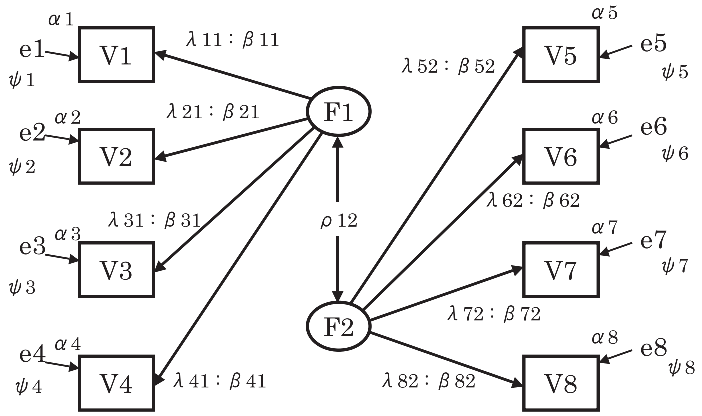

Figure 1: Model and parameters for the simulation study

are missing simultaneously. For example, assume that $\boldsymbol{x}_1$ is always missing if and only if $\boldsymbol{x}_2$ is missing. We can model this situation by packing two missing indicators m_1 and m_2 into one new missing indicator, for example denoting m_{1+2}.

5. Simulation study

In this section, we analyze some generated data sets that follow our proposed model, to show the validity of the proposed estimation. We report three simulation studies: in the first, the missing mechanism is MAR and in the others, the MAR assumption does not hold; the second study assumes that the factors influence the missing mechanism, while in the third study missingness depends on not only the factors but also the observed portion of the data x_{obs}.

Our interest lies in both the degree of the bias of estimates of the proposed method, and in the degree of reduction of the bias compared with previous methods.

To escape the influence of the variability of data generation, we generated 100 data sets, and analyzed each set using two methods: the proposed MCEM and the method of Finkbeiner (1979) under the assumption of MAR. The MAR estimation by Muthén et al. (1987) is very attractive, but was not attempted here because each set of data had several missing patterns. Consequently, the number of subjects may be less than the number of variables in some of the missing patterns. Note that the number of missing patterns was $2^8 = 256$ in the simulation studies. Therefore, the method of Muthén et al. method cannot be applied to the simulation studies. Note also that Finkbeiner's method does not include subjects with missing data for all variables because under MAR these cases contribute no information, while the proposed method can include such subjects (the average frequency of these subjects is about 6%).

We generated data sets with eight variables $x_1, \cdots, x_8$ measuring two latent variables f_1 and f_2. The model of data generation and relevant parameters are shown in Figure 1.

The parameters $\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}$, and $\boldsymbol{\psi}$ are named according to the model specifications in

Section 2. ω_j is the coefficient vector for x_{obs} on x_j according to Section 4 [1]. The variances of f_1 and f_2 are set at 1, and the correlation between f_1 and f_2 is named ρ_{12}. The parameters $\alpha_1, \cdots, \alpha_8$ can be regarded as "the baseline parameters for missing", in that these parameters determine the baseline frequency of missing. This assumption is the usual "MAR assumption" adopted in several previous studies. In the simulation study, we imposed the condition that the variables x_1, x_2, x_5, and x_6 are observed in every subject to make use of the observed portion of data in predicting missingness. Moreover, we assumed that the missingness of x_3, x_4 is explained by x_1, x_5, and the missingness of x_7, x_8 is explained by x_2, x_6, so there are 8 elements in $\boldsymbol{\omega}$.

Therefore, there are a total of 37 unknown parameters in this model. For simplification, however, some constant restrictions were imposed on the parameters in each simulation study. The sample sizes used in each simulation study were 400 and 1000.

In these studies, the true values of $\alpha_1, \cdots \alpha_8$ were set at $1.61, 0.693$, or 0, which correspond to a missing data frequency of $80, 66.7$, or 50%, respectively.

The number of sampling times of the latent variable vectors, L, was set at 1000. The convergence criterion for the Newton-Raphson algorithm in each M-step was 0.01, and the convergence criterion for the EM iterations was set at 0.0001.

5.1 Simulation study under the MAR assumption

To create MAR missing data, we constrained $\boldsymbol{\beta} = 0$, so that missingness did not depend on the latent variables f, but on the observed portion of the data, x_{obs}. Note that since the assumption that $\boldsymbol{\beta} = 0$ and $\boldsymbol{\omega} = 0$ leads to MCAR missing data, we did not generate data under this assumption. Our method can use variables that are observed only in some subjects, but the above set-up is used to simplify data generation. Therefore, there were a total of 29 unknown parameters in this simulation. We report the true values of the parameters and four estimates calculated from the estimation results for 100 data sets under three sample sizes in Tables 1, 2, 3 and 4. These show the (1) means of the estimates (MEANS), (2) means of the standard errors (SE), (3) Monte Carlo standard deviation of the estimates (MCSD), and (4) coverage (Cov), respectively.

The Monte Carlo standard deviation of the estimates is reported to check how well the estimated standard errors track the true standard errors. To compute "Coverage", the number of data sets is calculated in which the estimated interval from the 5th to the 95th percentile covers the true parameter. "Coverage" is defined as the proportion out of 100 replicates for which this event occurred, following Song & Lee (2002).

Note that estimates of the parameters for the missing mechanism (e.g., $\boldsymbol{\omega}13$; elements with an asterisk in Tables 1–2) are not calculated by Finkbeiner's method. To compare the accuracy of the proposed method with that of Finkbeiner, the RMS (root mean squares between the true values and corresponding estimates calculated from 100 replications) was calculated in each data set for the parameters that can be estimated using Finkbeiner's method (i.e., parameters in the missing mechanism were excluded). When $N = 400$, the sum of RMS for the parameters of the proposed model over 100 replicated data sets was 44.85, while that for Finkbeiner's method was 62.10. When $N = 1000$, the sum of RMS

Table 1: Estimates of the parameters (MAR, $N = 400$)

		MCMCEM				Finkbeiner			
	true value	Means	SE	MCSD	Cov	Means	SE	MCSD	Cov
$\lambda 11$	2	2.0519	0.0801	0.0938	0.9100	2.1032	0.0784	0.0901	0.8800
$\lambda 21$	2	2.0813	0.0892	0.0941	0.8800	2.0681	0.0802	0.0883	0.8700
$\lambda 31$	1.5	1.5091	0.1213	0.1280	0.9000	1.4906	0.1113	0.1203	0.8900
$\lambda 41$	1	1.0730	0.1145	0.1209	0.8800	1.0403	0.1036	0.1079	0.9000
$\lambda 52$	1	0.9893	0.0834	0.0920	0.8900	1.0246	0.0801	0.0903	0.9000
$\lambda 62$	1.5	1.4981	0.1085	0.1103	0.9000	1.4773	0.0928	0.1082	0.8600
$\lambda 72$	2	2.0234	0.1502	0.1662	0.9000	2.0501	0.1303	0.1294	0.8800
$\lambda 82$	2	2.0532	0.1593	0.1601	0.8900	2.0351	0.1414	0.1402	0.8900
$\alpha 3$	0.693	0.6841	0.0699	0.0772	0.9200	*	*	*	*
$\alpha 4$	0	0.0230	0.0578	0.0637	0.9300	*	*	*	*
$\alpha 7$	1.61	1.5882	0.0601	0.0653	0.8600	*	*	*	*
$\alpha 8$	1.61	1.6205	0.0495	0.0561	0.9000	*	*	*	*
$\omega 13$	0.5	0.4807	0.0305	0.0333	0.9100	*	*	*	*
$\omega 53$	0.3	0.2973	0.0501	0.0572	0.9200	*	*	*	*
$\omega 14$	0.3	0.2966	0.0479	0.0558	0.8900	*	*	*	*
$\omega 54$	0.5	0.2930	0.0611	0.0554	0.9000	*	*	*	*
$\omega 27$	0.5	0.5151	0.0510	0.0658	0.8700	*	*	*	*
$\omega 67$	0.3	0.2965	0.0487	0.0502	0.9100	*	*	*	*
$\omega 28$	0.3	0.5029	0.0481	0.0479	0.9300	*	*	*	*
$\omega 68$	0.5	0.4710	0.0347	0.0449	0.8800	*	*	*	*
$\psi 1$	1	0.9801	0.0901	0.1002	0.8700	0.9602	0.0886	0.0993	0.8700
$\psi 2$	1	0.9910	0.0903	0.0899	0.9100	1.0443	0.0802	0.0904	0.8700
$\psi 3$	0.8	0.7793	0.1002	0.1103	0.8800	0.8103	0.0891	0.0881	0.9000
$\psi 4$	0.5	0.5033	0.0843	0.0922	0.9100	0.4956	0.0814	0.0765	0.9100
$\psi 5$	0.5	0.4954	0.0794	0.0807	0.9000	0.4907	0.0733	0.0810	0.8900
$\psi 6$	0.8	0.8103	0.0700	0.0895	0.9100	0.8088	0.0728	0.0852	0.8800
$\psi 7$	1	0.9589	0.1317	0.1330	0.8700	0.9601	0.1231	0.1228	0.8600
$\psi 8$	1	1.1006	0.1308	0.1294	0.8800	1.1203	0.1197	0.1310	0.8400
$\rho 12$	0.4	0.4039	0.0441	0.0481	0.9100	0.4105	0.0458	0.0501	0.8800

for the parameters of the proposed model over 100 replicated data sets was 10.73 versus 23.01 for Finkbeiner's method. From these tables and the results in terms of the RMS, we can claim that our estimation method works well, even under MAR missingness. However, when the number of subjects is large ($N = 1000$), the estimates with the proposed method and Finkbeiner's method are very close to each other. We also applied the Wald test for hypothetical testing of the ignorability proposed in Section 4 to each generated data set. The null hypothesis was that the missing mechanism is MAR. The mean p-value calculated from the 100 data sets for 400 observations was 0.0913, and the MAR assumption was rejected in 5 out of 100 data sets. The smallest p-value was 0.00192. This result also supports the validity of the proposed estimation method. In this study, the average acceptance rate was 0.416, which indicates the relatively good approximation of the proposal distributions to the target distributions. For 35 data sets out of 200 replications, we observed that the MCEM algorithm did not converge within 100 iterations, although in these cases the algorithm stopped within 200 iterations.

Table 2: Estimates of the parameters (MAR, $N = 1000$)

		MCMCEM				Finkbeiner			
	true value	Means	SE	MCSD	Cov	Means	SE	MCSD	Cov
$\lambda 11$	2	2.0269	0.0361	0.0420	0.9000	2.0540	0.0352	0.0408	0.9000
$\lambda 21$	2	2.0564	0.0400	0.0420	0.8700	2.0401	0.0363	0.0401	0.9100
$\lambda 31$	1.5	1.4972	0.0608	0.0661	0.9100	1.5001	0.0502	0.5440	0.9200
$\lambda 41$	1	0.9916	0.0599	0.0618	0.9000	1.0211	0.4700	0.0491	0.9000
$\lambda 52$	1	1.0463	0.0401	0.0443	0.8700	1.0224	0.0362	0.0412	0.9000
$\lambda 62$	1.5	1.5070	0.0489	0.0506	0.9200	1.4933	0.0409	0.0508	0.8800
$\lambda 72$	2	1.9948	0.0801	0.0821	0.9300	2.0245	0.0612	0.0597	0.8800
$\lambda 82$	2	2.0091	0.0785	0.0830	0.9000	2.0151	0.0629	0.0633	0.9000
$\alpha 3$	0.693	0.6882	0.0330	0.0404	0.8900	*	*	*	*
$\alpha 4$	0	−0.0893	0.0260	0.3010	0.8600	*	*	*	*
$\alpha 7$	1.61	1.6403	0.0267	0.0289	0.8700	*	*	*	*
$\alpha 8$	1.61	1.6580	0.0234	0.0257	0.8500	*	*	*	*
$\omega 13$	0.5	0.4874	0.0138	0.0151	0.8700	*	*	*	*
$\omega 53$	0.3	0.2956	0.0232	0.0262	0.9100	*	*	*	*
$\omega 14$	0.3	0.3038	0.0214	0.0243	0.9000	*	*	*	*
$\omega 54$	0.5	0.3002	0.0280	0.0277	0.9200	*	*	*	*
$\omega 27$	0.5	0.4989	0.0231	0.0258	0.9100	*	*	*	*
$\omega 67$	0.3	0.3031	0.0217	0.0224	0.9200	*	*	*	*
$\omega 28$	0.3	0.2980	0.0215	0.0230	0.9100	*	*	*	*
$\omega 68$	0.5	0.5082	0.0160	0.0190	0.9100	*	*	*	*
$\psi 1$	1	0.9992	0.0440	0.0409	0.8900	0.9774	0.0392	0.0448	0.8600
$\psi 2$	1	1.0000	0.0409	0.0402	0.9100	0.9900	0.0361	0.0403	0.8800
$\psi 3$	0.8	0.8013	0.0528	0.0533	0.9000	0.7805	0.0405	0.0416	0.8700
$\psi 4$	0.5	0.4989	0.0510	0.0499	0.9100	0.5081	0.0370	0.0388	0.8900
$\psi 5$	0.5	0.4996	0.0359	0.0361	0.9200	0.4964	0.0319	0.0370	0.9000
$\psi 6$	0.8	0.8010	0.0322	0.0400	0.8900	0.8043	0.0346	0.0338	0.9000
$\psi 7$	1	1.0103	0.0714	0.0721	0.8700	0.9928	0.0562	0.0549	0.9100
$\psi 8$	1	0.9935	0.0761	0.0750	0.8600	1.0882	0.0550	0.0591	0.8800
$\rho 12$	0.4	0.4002	0.0208	0.0215	0.9200	0.4097	0.0211	0.0209	0.8800

5.2 Simulation study when the MAR assumption does not hold (I)

The model of data generation and relevant parameters are the same as that used in Subsection 5.1, but we assume that missingness depends on the latent variables f (i.e., $\boldsymbol{\beta} \neq 0$), so that the MAR assumption does not hold. In this simulation study, the assumption of non-ignorable missingness is

Missingness does not depend on the observed portion of data, i.e., $\boldsymbol{\omega} = 0$.

All eight variables have a chance to suffer from missingness predicted by the factors. In this simulation study, there were 33 unknown parameters.

We report the true values of the parameters and four estimates explained in the last subsection calculated from the estimation results for 100 data sets under two sample sizes (Table 3–4).

When $N = 400$, the sum of RMS for the model parameters over 100 data sets was 46.48 in the proposed model versus 483.63 in Finkbeiner's method, while the respective values for $N = 1000$ were 27.07 versus 331.51. From these tables and the results in terms

Table 3: Estimates of the parameters (non-ignorable I, $N = 400$)

		MCMCEM				Finkbeiner			
	true value	Means	SE	MCSD	Cov	Means	SE	MCSD	Cov
$\lambda 11$	2	2.0194	0.0988	0.1124	0.9300	2.2102	0.0941	0.1212	0.7700
$\lambda 21$	2	1.9892	0.1172	0.1311	0.9200	1.9818	0.0912	0.1302	0.9300
$\lambda 31$	1.5	1.5128	0.1203	0.1309	0.8900	1.3940	0.0734	0.1196	0.7500
$\lambda 41$	1	0.9898	0.1283	0.1334	0.8800	1.1930	0.0912	0.1304	0.7800
$\lambda 52$	1	0.9901	0.1291	0.1311	0.9500	1.2013	0.0922	0.1315	0.7300
$\lambda 62$	1.5	1.5201	0.1228	0.1293	0.8600	1.4035	0.0751	0.1183	0.8400
$\lambda 72$	2	1.9919	0.1128	0.1373	0.9500	1.8776	0.0921	0.1225	0.8200
$\lambda 82$	2	2.0096	0.0989	0.1281	0.9200	1.9302	0.0937	0.1147	0.8900
$\alpha 1$	1.61	1.5992	0.0729	0.0832	0.9400	*	*	*	*
$\alpha 2$	1.61	1.6301	0.0801	0.0883	0.8500	*	*	*	*
$\alpha 3$	0.693	0.7019	0.0791	0.0787	0.9300	*	*	*	*
$\alpha 4$	0	-0.1093	0.0801	0.0922	0.8900	*	*	*	*
$\alpha 5$	0	0.1816	0.0842	0.0911	0.8000	*	*	*	*
$\alpha 6$	0.693	0.7102	0.0817	0.0854	0.9500	*	*	*	*
$\alpha 7$	1.61	1.5592	0.0729	0.0941	0.9200	*	*	*	*
$\alpha 8$	1.61	1.6091	0.0731	0.0826	0.9200	*	*	*	*
$\beta 11$	0	0.0490	0.0318	0.0366	0.9000	*	*	*	*
$\beta 21$	0.5	0.5123	0.0518	0.0616	0.9300	*	*	*	*
$\beta 31$	0.5	0.4992	0.0526	0.0599	0.9500	*	*	*	*
$\beta 41$	0.5	0.4946	0.0640	0.0633	0.9400	*	*	*	*
$\beta 52$	0.5	0.5113	0.0599	0.0662	0.8900	*	*	*	*
$\beta 62$	0.5	0.4961	0.0518	0.0601	0.9300	*	*	*	*
$\beta 72$	0.5	0.5021	0.0522	0.0592	0.9500	*	*	*	*
$\beta 82$	0	-0.0180	0.0388	0.0408	0.9400	*	*	*	*
$\psi 1$	1	0.9719	0.1201	0.1301	0.8500	0.8891	0.1013	0.1307	0.8400
$\psi 2$	1	1.0218	0.1245	0.1244	0.9400	1.1003	0.1089	0.0912	0.8800
$\psi 3$	0.8	0.7836	0.0911	0.0992	0.8600	0.6891	0.0833	0.0981	0.7700
$\psi 4$	0.5	0.4971	0.0912	0.0881	0.9300	0.4298	0.0761	0.1003	0.8600
$\psi 5$	0.5	0.5001	0.0893	0.0895	0.9400	0.4195	0.0759	0.1032	0.8800
$\psi 6$	0.8	0.8112	0.0924	0.0982	0.9200	0.7592	0.0851	0.0936	0.8800
$\psi 7$	1	0.9920	0.1239	0.1193	0.9400	1.1512	0.1090	0.1535	0.8100
$\psi 8$	1	1.1033	0.1198	0.1301	0.8400	1.1207	0.1009	0.1335	0.8300
$\rho 12$	0.4	0.4108	0.0527	0.0588	0.8900	0.4428	0.0520	0.0606	0.8800

of RMS, we can claim that our estimation method works very well for non-ignorable missing data, compared with Finkbeiner's method. Also note that the estimates using Finkbeiner's method do not approach the true value, even when the number of subjects is large ($n = 1000$). Conversely, the estimates using the MCEM algorithm decrease the RMS drastically when the sample size is large. As mentioned above, the proposed method exceeds ML under the MAR assumption in that:

1 The estimation of the parameters of the factor analysis model ($\boldsymbol{\vartheta}$, e.g., λ_{11}) is accurate,

2 The parameters of the missing mechanism ($\boldsymbol{\varphi}$, i.e., $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) can be estimated.

We also applied the Wald test for hypothetical testing, in which the null hypothesis is that the missingness is MAR. The mean p-value calculated from the 100 data sets for 400

Table 4: Estimates of the parameters (non-ignorable I, $N = 1000$)

		MCMCEM				Finkbeiner			
	true value	Means	SE	MCSD	Cov	Means	SE	MCSD	Cov
$\lambda11$	2	2.0091	0.0540	0.0551	0.9100	2.2301	0.0456	0.0703	0.7500
$\lambda21$	2	2.0123	0.0491	0.0544	0.9000	2.1035	0.0477	0.0692	0.8900
$\lambda31$	1.5	1.4946	0.0534	0.0549	0.9200	1.4426	0.0387	0.0557	0.8600
$\lambda41$	1	0.9936	0.0493	0.0497	0.9100	0.8935	0.0434	0.0651	0.8400
$\lambda52$	1	1.0140	0.0663	0.0688	0.8900	0.9035	0.0489	0.0609	0.8600
$\lambda62$	1.5	1.4988	0.0678	0.0702	0.9100	1.4114	0.0371	0.0573	0.8700
$\lambda72$	2	1.9893	0.0608	0.0777	0.9100	1.8631	0.0465	0.0615	0.8200
$\lambda82$	2	1.9880	0.0483	0.0523	0.8900	1.8720	0.0454	0.0661	0.8400
$\alpha1$	1.61	1.6088	0.0392	0.0404	0.9100	*	*	*	*
$\alpha2$	1.61	1.6100	0.0368	0.0406	0.9200	*	*	*	*
$\alpha3$	0.693	0.6907	0.0410	0.0448	0.9200	*	*	*	*
$\alpha4$	0	−0.0777	0.0394	0.0393	0.8900	*	*	*	*
$\alpha5$	0	−0.0818	0.0455	0.0501	0.9300	*	*	*	*
$\alpha6$	0.693	0.7003	0.0472	0.0485	0.9100	*	*	*	*
$\alpha7$	1.61	1.6181	0.0421	0.0433	0.8700	*	*	*	*
$\alpha8$	1.61	1.6093	0.0398	0.0395	0.9000	*	*	*	*
$\beta11$	0	0.0402	0.0178	0.0200	0.9200	*	*	*	*
$\beta21$	0.5	0.4945	0.0244	0.0256	0.9100	*	*	*	*
$\beta31$	0.5	0.5031	0.0301	0.0300	0.9000	*	*	*	*
$\beta41$	0.5	0.4981	0.0352	0.0366	0.8800	*	*	*	*
$\beta52$	0.5	0.5062	0.0310	0.0318	0.8900	*	*	*	*
$\beta62$	0.5	0.5010	0.0289	0.0301	0.9000	*	*	*	*
$\beta72$	0.5	0.5042	0.3010	0.0311	0.9100	*	*	*	*
$\beta82$	0	−0.0300	0.0176	0.0180	0.8800	*	*	*	*
$\psi1$	1	0.9904	0.0701	0.0740	0.9100	0.8864	0.0554	0.0631	0.8700
$\psi2$	1	1.0100	0.0611	0.0702	0.9200	0.8567	0.0521	0.0486	0.8400
$\psi3$	0.8	0.8120	0.0518	0.0601	0.9300	0.6925	0.0426	0.0473	0.7100
$\psi4$	0.5	0.5080	0.0443	0.0489	0.8700	0.4936	0.0372	0.0470	0.8900
$\psi5$	0.5	0.5028	0.0417	0.0409	0.9000	0.4677	0.0391	0.0504	0.8400
$\psi6$	0.8	0.8085	0.0423	0.0580	0.9200	0.8656	0.0419	0.0522	0.8600
$\psi7$	1	1.0773	0.0595	0.0603	0.8700	0.9154	0.0515	0.0693	0.8300
$\psi8$	1	1.0704	0.0576	0.0599	0.8800	0.9436	0.0520	0.0701	0.8700
$\rho12$	0.4	0.3986	0.0266	0.0268	0.9100	0.3856	0.0281	0.0334	0.9000

observations was 0.0129, and the MAR hypothesis was accepted in 2 out of 100 data sets. The largest p-value was 0.0918. These results also support the validity of the proposed estimation method. In this simulation study, the average acceptance rate was 0.308. The relatively high acceptance rate indicates the efficiency of the proposed Metropolis-Hastings algorithm. To confirm the convergence of the EM iteration sequence, we monitored plots of the value of the likelihood and some parameters in each EM iteration. For example, Figures 2 and 3 plot the estimates of λ_{11} and β_{41} for particular data, respectively, where the abscissa in each figure shows the number of iterations.

The estimated difference between the consecutive log-likelihoods in particular data is also plotted as a function of the iteration calculated by Bridge sampling in Section 3 [4].

From these plots, we can conclude that the sequences of estimates can quickly become stable and the process can also converge very rapidly. In 45 out of 200 replications, the

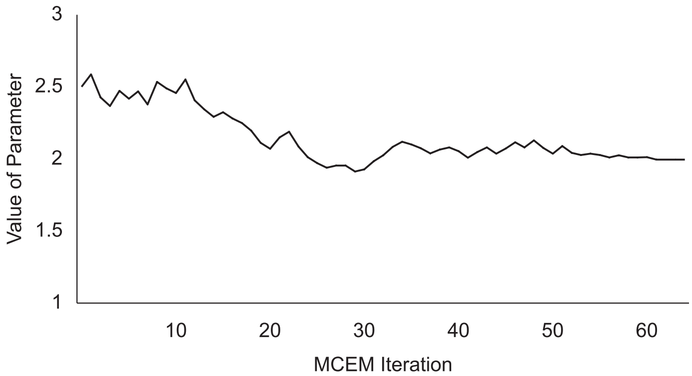

Figure 2: The behavior of EM sequences of $\lambda 11$

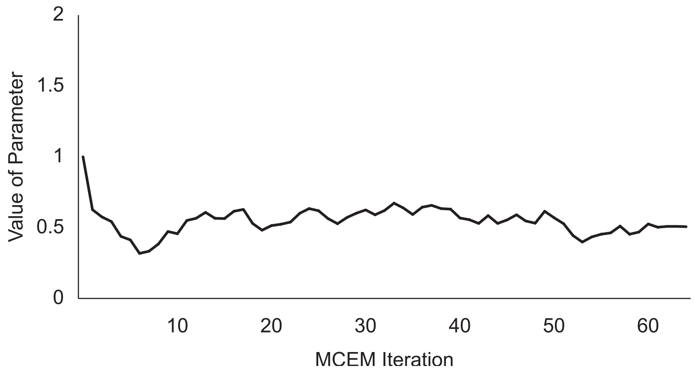

Figure 3: The behavior of EM sequences of $\beta 41$

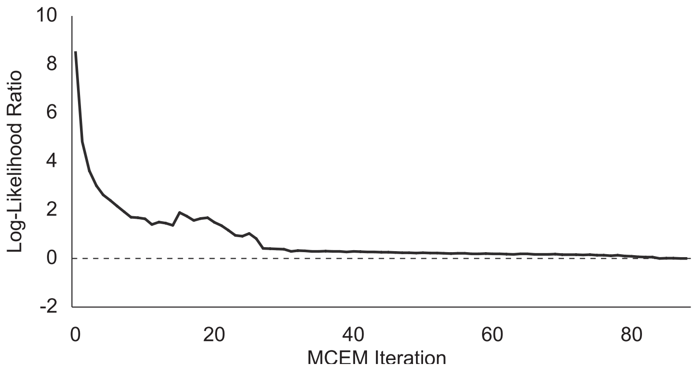

Figure 4: The log-likelihood ratio against iteration

MCEM algorithm did not converge within 100 iterations. Nevertheless, for these 45 data sets, the algorithm stopped within 300 iterations.

5.3 Simulation study when the MAR assumption does not hold (II)

The model of data generation and relevant parameters are the same as used in Subsection 5.1, but we assume that missingness depends on the latent variables f (i.e., $\boldsymbol{\beta} \neq 0$), so that the MAR assumption does not hold. In this model, the assumption of non-ignorable missingness is:

Missingness depends not only on f, but also on x_{obs}, i.e., $\boldsymbol{\omega} \neq 0$.

In the simulation study, we imposed the condition that variables x_1, x_2, x_5, and x_6 are observed in every subject to make use of the observed portion of data in predicting missingness, as in Subsection 5.1. In addition, we assumed that the missingness of x_3, x_4 is explained by x_1, x_5 and latent factor f_1, and the missingness of x_7, x_8 is explained by x_2, x_6 and f_2. Then, there were a total of 37 unknown parameters. We report the true values of parameters and the four estimates explained in the last subsection calculated from the estimation results for 100 data sets under two sample sizes (Table 5–6).

Note that estimates of parameters for the missing mechanism (e.g., β_{11}; elements with an asterisk in Table 1) cannot be calculated by Finkbeiner's method. When $N = 400$, the sum of RMS for the model parameters over 100 data sets was 102.87 in the proposed model versus 646.28 in Finkbeiner's method. When $N = 1000$, the respective values were 26.29 versus 372.69. From these tables and the results in terms of RMS, we can claim that our estimation method works very well for non-ignorable missing data, compared to Finkbeiner's method. We also applied the Wald test for hypothetical testing, in which the null hypothesis is that the missingness is MAR. In the simulation study, the mean p-value calculated from the 100 data sets was 0.0104, and the hypothesis of MAR was accepted in 1 out of 100 data sets. The largest p-value was 0.0549. These results also support the validity of the proposed estimation method. The average acceptance rate was 0.2910, which indicates the efficiency of the proposed Metropolis-Hastings algorithm. In 52 out of 200 replications, the MCEM algorithm did not converge within 100 iterations, although it did stop within 300 iterations.

6. Real Data Analysis

We illustrate the applicability of the proposed method using a read data form MIDUS (Midlife in the United States; Lackman & Weaver, 1998; Mroczek & Kolarz, 1998) project. The data are based on the responses of 3,690 individuals who were asked questions on their socio-economic status and personality using the random digit dialing method.

In this paper, we use four indicators (F1–F4) of financial status and four items (N1–N4) that measure personality "neuroticism" (Costa & McCrae, 1985). Neuroticism is one of five factor personalities that many psychological studies have shown to be valid in the last two decades.

We used the following indicators of financial status:

F1 How would you rate your financial situation these days?

Table 5: Estimates of the parameters (non-ignorable II, $N = 400$)

		MCMCEM				Finkbeiner			
	true value	Means	SE	MCSD	Cov	Means	SE	MCSD	Cov
$\lambda 11$	2	2.1882	0.1125	0.1124	0.8600	1.7903	0.1045	0.1103	0.7300
$\lambda 21$	2	2.0286	0.1029	0.1311	0.8900	1.8395	0.0893	0.1002	0.7900
$\lambda 31$	1.5	1.4791	0.1672	0.1309	0.8500	1.4913	0.1558	0.1205	0.8700
$\lambda 41$	1	0.9638	0.1410	0.1334	0.8400	0.8398	0.1406	0.1670	0.8000
$\lambda 52$	1	1.0912	0.0992	0.1311	0.9100	1.2217	0.1038	0.1052	0.7200
$\lambda 62$	1.5	1.4457	0.1103	0.1293	0.8900	1.4038	0.1077	0.0953	0.8400
$\lambda 72$	2	1.9881	0.1692	0.1373	0.9300	2.0835	0.1448	0.1308	0.8700
$\lambda 82$	2	2.0123	0.1509	0.1281	0.9100	1.6793	0.1452	0.1596	0.6900
$\alpha 3$	0.693	0.7126	0.0879	0.0910	0.8300	*	*	*	*
$\alpha 4$	0	0.0772	0.0794	0.0814	0.8600	*	*	*	*
$\alpha 7$	0.693	0.6651	0.0802	0.0820	0.8300	*	*	*	*
$\alpha 8$	0	−0.0902	0.0811	0.0799	0.9100	*	*	*	*
$\omega 13$	0.5	0.4909	0.0617	0.0678	0.8900	*	*	*	*
$\omega 53$	0.3	0.3102	0.0721	0.0703	0.8800	*	*	*	*
$\omega 14$	0.3	0.2927	0.0663	0.0649	0.9000	*	*	*	*
$\omega 54$	0.5	0.4691	0.0801	0.0793	0.8700	*	*	*	*
$\omega 27$	0.5	0.5082	0.0792	0.0905	0.9000	*	*	*	*
$\omega 67$	0.3	0.3100	0.0769	0.0772	0.8900	*	*	*	*
$\omega 28$	0.3	0.2977	0.0693	0.0753	0.8900	*	*	*	*
$\omega 68$	0.5	0.4893	0.0801	0.0869	0.8800	*	*	*	*
$\beta 11$	0	−0.0112	0.0519	0.0557	0.9200	*	*	*	*
$\beta 21$	0.5	0.4887	0.0634	0.0633	0.8600	*	*	*	*
$\beta 31$	0.5	0.5079	0.0819	0.0888	0.9000	*	*	*	*
$\beta 41$	0.5	0.5066	0.0714	0.0729	0.9100	*	*	*	*
$\beta 52$	0.5	0.4918	0.0602	0.0717	0.8900	*	*	*	*
$\beta 62$	0.5	0.5112	0.0693	0.0721	0.8600	*	*	*	*
$\beta 72$	0.5	0.5114	0.0797	0.0882	0.8700	*	*	*	*
$\beta 82$	0	0.0131	0.0612	0.0707	0.8900	*	*	*	*
$\psi 1$	1	1.0222	0.1023	0.1010	0.9000	0.8831	0.0885	0.0901	0.8500
$\psi 2$	1	0.9810	0.1102	0.1119	0.8700	0.8715	0.0904	0.0977	0.8300
$\psi 3$	0.8	0.8055	0.1513	0.0145	0.8900	0.7382	0.1121	0.1081	0.8500
$\psi 4$	0.5	0.5092	0.1350	0.0151	0.8800	0.4807	0.1254	0.1119	0.8800
$\psi 5$	0.5	0.4881	0.0904	0.1089	0.8400	0.5130	0.0841	0.0907	0.9000
$\psi 6$	0.8	0.7912	0.0887	0.0889	0.9000	0.7508	0.0821	0.0895	0.8800
$\psi 7$	1	0.9811	0.1542	0.1454	0.9100	0.9695	0.1331	0.1222	0.8900
$\psi 8$	1	1.0202	0.1382	0.1433	0.8900	0.9306	0.1289	0.1180	0.8700
$\rho 12$	0.4	0.3891	0.0572	0.0494	0.8800	0.5012	0.0509	0.0662	0.8300

F2 How would you rate the amount of control your have over your financial situation these days?

F3 How difficult is it for you to pay your monthly bills?

F4 What was your personal earnings income for the past 12 months?

F1 and F2 are on an eleven-point scale, F3 is a four-point scale, and F4 can be considered a continuous variable.

The non-response percentages for F1–F4 were 1.8, 1.4, 1.5, and 5.7%, respectively.

Items N1–N4 were taken from the Neo FFI (NEO Five Factor Inventory; Costa & McCrae, 1985).

Table 6: Estimates of the parameters (non-ignorable II, $N = 1000$)

		MCMCEM				Finkbeiner			
	true value	Means	SE	MCSD	Cov	Means	SE	MCSD	Cov
$\lambda 11$	2	2.0913	0.0542	0.0577	0.8700	1.8391	0.0471	0.0532	0.7500
$\lambda 21$	2	1.9814	0.0477	0.0539	0.8900	1.8228	0.0403	0.0551	0.7700
$\lambda 31$	1.5	1.4830	0.0781	0.0602	0.8700	1.5031	0.0690	0.0582	0.8900
$\lambda 41$	1	0.9905	0.0663	0.0597	0.9000	0.8702	0.0655	0.0779	0.8100
$\lambda 52$	1	1.0523	0.0490	0.0571	0.9100	1.2331	0.0459	0.0568	0.7100
$\lambda 62$	1.5	1.5108	0.0493	0.0572	0.9200	1.4604	0.0485	0.0511	0.8400
$\lambda 72$	2	1.9915	0.0761	0.0602	0.9200	2.1037	0.0672	0.0703	0.8800
$\lambda 82$	2	2.0187	0.0662	0.0591	0.8900	1.8005	0.0648	0.0784	0.6800
$\alpha 3$	0.693	0.7011	0.0430	0.0438	0.8900	*	*	*	*
$\alpha 4$	0	0.0285	0.0361	0.0374	0.8800	*	*	*	*
$\alpha 7$	0.693	0.6889	0.0358	0.0366	0.8700	*	*	*	*
$\alpha 8$	0	−0.0306	0.0367	0.0384	0.9000	*	*	*	*
$\omega 13$	0.5	0.5085	0.0282	0.0310	0.8900	*	*	*	*
$\omega 53$	0.3	0.3029	0.0333	0.0336	0.9000	*	*	*	*
$\omega 14$	0.3	0.2931	0.0299	0.0332	0.8900	*	*	*	*
$\omega 54$	0.5	0.5108	0.0361	0.0361	0.8800	*	*	*	*
$\omega 27$	0.5	0.5041	0.0359	0.0393	0.9000	*	*	*	*
$\omega 67$	0.3	0.3075	0.0344	0.0385	0.8900	*	*	*	*
$\omega 28$	0.3	0.3081	0.0317	0.0341	0.8900	*	*	*	*
$\omega 68$	0.5	0.5083	0.0370	0.0399	0.9000	*	*	*	*
$\beta 11$	0	0.0109	0.0241	0.0251	0.8900	*	*	*	*
$\beta 21$	0.5	0.4930	0.0305	0.0310	0.8800	*	*	*	*
$\beta 31$	0.5	0.5101	0.0371	0.0407	0.8900	*	*	*	*
$\beta 41$	0.5	0.5026	0.0352	0.0379	0.9100	*	*	*	*
$\beta 52$	0.5	0.5103	0.0271	0.0308	0.8700	*	*	*	*
$\beta 62$	0.5	0.5039	0.0307	0.0318	0.8900	*	*	*	*
$\beta 72$	0.5	0.5082	0.0360	0.0394	0.9000	*	*	*	*
$\beta 82$	0	−0.0102	0.0349	0.0339	0.9000	*	*	*	*
$\psi 1$	1	1.0131	0.0455	0.0468	0.8900	0.9001	0.0412	0.0403	0.8200
$\psi 2$	1	0.9875	0.0488	0.0500	0.8800	0.8803	0.0409	0.0441	0.8000
$\psi 3$	0.8	0.8025	0.0691	0.0700	0.9100	0.7209	0.0594	0.0594	0.8500
$\psi 4$	0.5	0.5049	0.0637	0.0682	0.8900	0.4779	0.0588	0.0603	0.8600
$\psi 5$	0.5	0.5021	0.0404	0.0432	0.9100	0.5138	0.0388	0.0392	0.9100
$\psi 6$	0.8	0.7940	0.0400	0.0428	0.9000	0.7735	0.0376	0.0407	0.8100
$\psi 7$	1	0.9889	0.0689	0.0739	0.9100	0.9885	0.0603	0.0551	0.8800
$\psi 8$	1	1.0441	0.0618	0.0705	0.8900	0.9411	0.0579	0.0617	0.7900
$\rho 12$	0.4	0.4041	0.0379	0.0442	0.8800	0.4892	0.0249	0.0304	0.8400

In this illustration, we assumed that the values of F1–F4 are explained by the "financial status factor", while N1–N4 are explained by neuroticism. Moreover, we assumed that both factors may explain the missingness of each item. Table 7 shows the resulting estimates in which $\boldsymbol{\lambda}$'s are the factor loadings and β's are the regression coefficients of missingness on each factor.

For example, the $\boldsymbol{\lambda}$ of finance status factor F1 was 0.831. The results show that both the latent variable financial status and "neuroticism" affect the missingness of the indicators of financial status. However, the missingness of N1–N4 is explained by the neuroticism factor, but not by the financial status factor. We also show that the correlation of the two

Table 7: Estimates of the real data illustration

λ			β		
Value	finance	neuroticism	Missingness	finance	neuroticism
F1	0.831	0*	F1	0.428	0.466
F2	0.779	0*	F2	0.441	0.473
F3	0.792	0*	F3	0.402	0.407
F4	0.844	0*	F4	0.509	0.493
N1	0*	0.679	N1	0.073	−0.309
N2	0*	0.701	N2	0.110	−0.365
N3	0*	0.684	N3	0.087	−0.312
N4	0*	0.704	N4	0.121	−0.405

Note: 0* is fixed at zero

latent variables is relatively low ($\phi_{12} = 0.135$).

7. Conclusion

This paper proposed a latent variable model that expresses not only the relationship between observable and latent variables but also the missing mechanism explained by the latent variables, and discussed an application of the Monte Carlo EM algorithm to making inferences in the proposed model. In the proposed model, since analytical expectation in the E-step is impossible, sampling from the conditional distribution of factors is implemented for Monte Carlo expectation. We cannot draw samples from the conditional distribution directly, but Independence sampling proposed by Tierney, a variant of the Metropolis-Hastings algorithm, and an ingenious set-up of the proposal distribution enable us to do it.

Our proposed estimation method has some advantages over previous methods. The proposed method does not require the MAR assumption. Previous methods excluded subjects with missing data for all variables whereas our method can include these subjects. Moreover, we can avoid the limitation with respect to the number of subjects in each missing pattern. Since there are usually a very large number of missing patterns in real data, the proposed method could be used to handle a variety of data sets with missing data. Moreover, the proposed method can also test the hypothesis of MAR/MCAR because our model includes the missing mechanism and estimates its parameters.

Using simulation studies, our estimation method was compared with a traditional method and confirmed to work well. It should be emphasized that the proposed method produces more accurate estimates for larger samples, whereas the bias and least square errors of the traditional method do not decrease much in a large sample.

The difference between the results of our proposed method and those of the traditional method is larger when missingness is predicted not only by factors but also by the observed portion of the data x_{obs}.

Remember that the proposed model deals with a situation in which the latent variables (and the observed portion of the data) predict the missing patterns. In a theoretical sense, however, the factors contain all the information that the data contains, and the portion of observables that the factors cannot explain is an "error" and has no information. Errors predict nothing. The reason we assume that the latent factors explain the missing patterns is that the complete data follow factor analysis. When data suffers from non-ignorable missingness, the most difficult problem is to restore the missing portion of data x_{miss} using the observed portion of data x_{obs} and missing patterns. If the complete data x follows the factor analysis model, we can restore (and integrate out) x_{miss} using the proposed model specification, but we do not claim that this approach is applicable to general settings.

References

Arminger, G., & Sobel, M.E. (1990). Pseudo-Maximum Likelihood Estimation of Mean and Covariance Structures with Missing Data. *Journal of the American Statistical Association*, **85**, 195–203.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.

Costa, P.T., & McCrae, R.R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Chen, M-H., Shao, Q-M., & Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Ser. B*, **39**, 1–38.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, **44**, 409–420.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, **52**, 681–700.

Lackman, M.E., & Weaver, S.L. (1998). The sence of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social psychology*, **74**, 763–773.

Lee, S-Y. (1986). Estimation for structural equation models with missing data. *Psychometrika*, **51**, 93–99.

Lee, S-Y., Poon, W.Y., & Bentler, P.M. (1992). Structural equation models with polytomous variables. *Psychometrika*, **57**, 89–106.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, **44**, 226–233.

Magnus, J.R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics.* Chichester: John Wiley & Sons.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **42**, 109–142.

Meng, X-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of Bridge Sampling. *Journal of the American Statistical Association*, **91**, 1254–1267.

Meng, X-L., & Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity. *Statistica Sinica*, **6**, 831–860.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, **65**, 391–411.

Mroczek, D.K., & Kolarz, C.M. (1998). The effect of age on positive and negative affect: A devepolmental perspective on happiness. *Journal of Personality and Social psychology*, **75**, 1333–1349.

Muthén, B. (1983). latent variable structural equation model with a comprehensive measurement model. *Journal of Econometrics*, **22**, 43–65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**, 115–132.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, **52**, 431–462.

O'Muircheartaigh & Moustaki (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales, *Journal of the Royal Statistical Society, Ser. A*, **162**, 177–194.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Shafer, J.L. (1997). *Analysis of incomplete multivariate data.* Boca Raton: Chapman & Hall.

Shi, J.Q., & Lee, S-Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, Ser. B*, **62**, 77–87.

Song, X-Y & Lee, S-Y. (2002). Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika*, **67**, 261–288.

Tierney, L. (1994). Markov Chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701–1762.

Wei, G.C.G, & Tanner, M.A. (1990). A Monte Carlo implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm. *Journal of the American Statistical Association*, **85**, 699–704.

(Received June 15 2004, Revised December 22 2004)