Published in final edited form as:

Stat Med. 2011 September 20; 30(21): . doi:10.1002/sim.4310.

Extended Mixture Factor Analysis Model with Covariates for Mixed Binary and Continuous Responses

Xinming An^a and Peter M Bentler^b

^aDepartment of Psychology, UCLA

^bDepartment of Psychology and Statistics, UCLA

Abstract

Finite mixture factor analysis provides a parsimonious model to explore latent group structures of high dimensional data. In this modeling framework, latent structures can be specified for continuous responses. However, dichotomous items are often used to define latent domains in practice. This paper proposes an extended finite mixture factor analysis model with covariates to model mixed continuous and binary responses. A Monte Carlo EM (MCEM) algorithm is used to estimate the model. In the E step, closed form solutions are not available for the conditional expectation of complete data log likelihood, so it is approximated by sample means, which are in turn generated by the Gibbs sampler from the joint conditional distribution of latent variables. To monitor the convergence of the MCEM algorithm, bridge sampling is used to calculate the log likelihood ratio of two successive iterations. A diagnostic plot of the log likelihood ratio against iterations is adopted for monitoring the convergence of the MCEM algorithm. Model comparison is based on BIC, in which the observed data log likelihood is approximated by a Monte Carlo method. The computational properties of the MCEM algorithm are investigated by simulation studies. A real data example is used to illustrate the practical usefulness of the model. Finally, limitations and possible extensions are discussed.

Keywords

finite mixture model; factor analysis model; MCEM algorithm; Gibbs sampler

1. Introduction

Most statistical models, such as regression and factor analysis, assume a population is homogeneous. In practice, however, this is often not the case. As noted by Muthén [1], if the target population consists of subgroups with different properties (e.g., different means or variances), results obtained from standard statistical models can be seriously distorted. Finite mixture models provide a natural way to solve this problem. This methodology has been applied in many major areas of statistics for different purposes, for example, cluster analysis [2], image analysis [3], survival analysis (for a comprehensive review see [4] and [5]). In the late 1990's a number of authors proposed several approaches to latent variable mixture modeling [e.g, 6, 7, 8, 9]. For example, Ghahramani and Hinton [10] proposed a finite mixture factor analysis model by combining the features of finite mixture and factor analysis models. Finite mixture factor analysis provides a parsimonious model to explore the latent group structures of high dimensional data [e.g, 11]. Applications to growth mixture modeling have become numerous, although problems in the methodology have been identified [e.g, 12, 13]. More recently, factor mixture models have been improved by many others in different directions, especially in estimation methodology [e.g, 14, 15, 16, 17, 18, 19, 20, 21]. An excellent compendium of recent work is given by Hancock and Samuelsen [22].

While the finite mixture approach is useful for continuous responses, dichotomous responses are often encountered in practice, especially in medical and social research. Muthén and Shedden [23] proposed a finite mixture factor analysis model with covariates for mixed continuous and binary responses. In this model, latent factors cannot be assumed to generate observed binary responses. However, binary items are often used to define latent domains such as math ability and depression. In this paper, we propose an extended finite mixture factor analysis model with covariates for mixed continuous and binary responses. Under our modeling framework, a very general latent structure can be specified to underlie mixed continuous and binary responses.

Muthén and Asparouhov [24] introduced a very general modeling framework which contains many statistical models, including the one we propose here, as special cases, but, as in other general modeling frameworks, it suffers from difficulties in model estimation and identification. While the authors presented the observed data likelihood and the expectation of the complete data likelihood, they did not provide any information on how to maximize these functions, which is the most interesting and challenging aspect of model estimation. Furthermore, the authors provided no discussion on how to address model identification, which is a very common problem for complex models and often leads to serious estimation difficulties if not well addressed. The absence of a published technical development hinders other researchers from replicating their modeling procedures and from providing detailed analyses on advantages, limitations, or inadequacies of any practical implementation of their approach.

In contrast, all the detailed information on model estimation and identification is clearly presented in this paper. Based on these technical details, the model can be easily implemented, tested, used and further improved by others. The model setup and identification will be presented in the next section. A Monte Carlo EM (MCEM) algorithm is proposed to estimate the parameters of the model. The MCEM algorithm and its convergence diagnosis is covered in detail in section 3. Model comparison based on the BIC fit index is briefly discussed in section 4. In section 5, simulation studies are used to investigate the computational properties of the MCEM algorithm in recovering the model parameters and in comparison with several likelihood based estimation methods implemented in Mplus 5.0 [25]. A real data example is also used to illustrate the practical usefulness of this model. The paper concludes with discussions on limitations and possible extensions in section 6.

2. The Extended Finite Mixture Model

In our model setup, the observed data include three parts: a c -dimensional continuous response variable y_c , a d -dimensional binary response variable u_d , and a q -dimensional covariates x . We assume that there is a d -dimensional continuous latent response variable y_d underlying u_d . Let $y = (y_c, y_d)$, $c + d = p$. Then our mixture factor analysis model with covariates can be expressed by the following three equations,

$$P(u_{dj}=1|\eta_i)=P(y_{dj}>0|\eta_i), \quad j=1, \dots, d. \quad (1)$$

$$y_i=b+\Lambda_y\eta_i+\varepsilon_i, \quad (2)$$

$$\eta_i=Ac_i+\Gamma_\eta x_i+\zeta_i. \quad (3)$$

Equation (1) uses the latent response formula to connect the observed binary response with the underlying continuous latent response. As compared with the generalized linear model formulation, the latent response formulation is easier for interpretation, identification and estimation, especially when EM type algorithms are used [26, 27, 28]. Equation (2) specifies the factor analysis model. In equation (2), b is the intercept, Λ_y is a $p \times m$ factor loading matrix, η_i is a vector of latent factors, and ε_i is a residual vector that is uncorrelated with other variables in the model and is normally distributed with mean zero and diagonal covariance matrix Σ . Because η_{ij} , $j = 1, \dots, p$ are independent and normally distributed, conditional on latent factor η_i , latent response y_{dj} , $j = 1, \dots, d$ in equation (1) are also independent and normally distributed. In equation (3), A is an $m \times K$ matrix whose columns a_k , $k = 1, \dots, K$ represent subgroup mean differences on latent factors η_i , Γ_η is an $m \times q$ regression coefficients matrix for covariates x_i , c_i is a K -dimensional latent vector with $c_{ik} = 1$ or 0, according to whether subject i belongs to the k th latent class, and ζ_i is an m -dimensional residual vector uncorrelated with other variables, and normally distributed with mean zero and covariance matrix Ψ . The mixture proportions of latent classes are represented by $\pi = (\pi_1, \dots, \pi_K)$. Let $\theta = \{b, \Lambda_y, \Sigma, A, \Gamma_\eta, \Psi, \pi\}$ be a set of all the parameters in the model.

This model is not identified without enough restrictions applied on the parameter space. First, to identify the factor analysis model, some parameters in Λ_y and Ψ need to be fixed. In exploratory analysis, matrix Ψ is fixed as the identity matrix and Λ_{ij} is fixed at 0, for $j \neq i$. In confirmatory analysis, the researcher will specify enough zeros in Λ_y such that factor rotation will be fixed. To fix factor scales, we restrict one element in each column of Λ_y to be 1. Second, suppose we have a $m \times k$ matrix B whose columns are the same and equal to a vector α . Adding this matrix to A is equivalent to changing the scale of subgroup means on η . Thus we have the following equations,

$$\eta_i = (A+B)c_i + \Gamma_\eta x_i + \zeta_i = Ac_i + \alpha + \Gamma_\eta x_i + \zeta_i. \quad (4)$$

and

$$y_i = b + \Lambda_y \alpha + \Lambda_y (Ac_i + \Gamma_\eta x_i + \zeta_i) \quad (5)$$

Equation (4) and (5) indicate that matrix A and intercept b are interdependent, so they can not be identified separately. To solve this problem, we fix the first column of A at zero and let the rest columns of A and b be free to be estimated. Last, let μ_{dj} and σ_{dj} represent the mean and standard deviation for the distribution of y_{dj} conditional on η_i . Then equation (1) can be expanded as

$$P(u_{dj}=1|\eta_i) = P(y_{dj} > 0|\eta_i) = P\left(\frac{y_{dj} - \mu_{dj}}{\sigma_{dj}} > \frac{-\mu_{dj}}{\sigma_{dj}}|\eta_i\right) = \Phi\left(\frac{\mu_{dj}}{\sigma_{dj}}\right), \quad (6)$$

where Φ is the cumulative distribution function for a standard normal distribution. If we multiply the numerator and denominator of term $\frac{y_{dj} - \mu_{dj}}{\sigma_{dj}}$ by the same constant, equation (6) still holds. This indicates that the scale of latent response variable y_d is not identified. To identify the scale for y_{dj} , $j = 1, \dots, d$, we fix their variance, σ_{dj}^2 , to be 1.

To use MCEM to estimate the model, we expand the observed data, $\{y_c, u_d\}$, into complete data $\{y_c, u_d, y_d, \eta, c\}$ by including latent variables. Then complete data likelihood can be written as

$$L_c = \prod_{i=1}^n p(y_i | \eta_i, \theta) p(\eta_i | c_i, x_i, \theta) p(c_i | \pi). \quad (7)$$

The corresponding complete data log likelihood is

$$\log L_c = \sum_{i=1}^n [\log p(y_i | \eta_i, \theta) + \log p(\eta_i | c_i, x_i, \theta) + \log p(c_i | \pi)], \quad (8)$$

where

$$\sum_{i=1}^n \log p(c_i | \pi) = \sum_{i=1}^n \sum_{k=1}^K c_{ik} \log \pi_{ik}, \quad (9)$$

$$\sum_{i=1}^n \log p(\eta_i | c_i, x_i, \theta) = \sum_{i=1}^n \left[-\frac{m}{2} \log(2\pi) - \frac{1}{2} \log |\Psi| - \frac{1}{2} (\eta_i - A c_i - \Gamma_\eta x_i) \Psi^{-1} (\eta_i - A c_i - \Gamma_\eta x_i)^T \right], \quad (10)$$

and

$$\sum_{i=1}^n \log p(y_i | \eta_i, \theta) = \sum_{i=1}^n \left[-\frac{p}{2} \log(2\pi) - \frac{1}{2} \log \left| \sum \left(-\frac{1}{2} (y_i - b - \Lambda_y \eta_i) \sum^{-1} (y_i - b - \Lambda_y \eta_i)^T \right) \right| \right]. \quad (11)$$

Equation (8) indicates that the complete data log likelihood can be divided into three parts. We can calculate the conditional expectation for each part and maximize them separately. The first part, expressed by equation (9), is a linear function of latent variable c_{ik} , so its conditional expectation can be easily obtained by replacing c_{ik} by its conditional expectation $E(c_{ik} | y_{ci}, u_{di}, x_i)$.

The last two parts, expressed by equations (10) and (11), are equivalent to the log likelihood of the following two multivariate linear models. The first one is

$$Y = \tilde{\eta} \tilde{\Lambda}_y + \varepsilon, \quad (12)$$

where $Y_{(n \times p)} = \{y_1^T, \dots, y_n^T\}^T$, $\tilde{\eta}_{(n \times (m+1))} = \{(1, \eta_1^T), \dots, (1, \eta_n^T)\}^T = (1\eta)$ (where 1 is an n by 1 vector, η is a n by p matrix, and its ith row is η_i),

$\tilde{\Lambda}_{y((m+1) \times p)} = \{b^T, \Lambda_y^T\}^T$, $\varepsilon = \{\varepsilon_1^T, \dots, \varepsilon_n^T\}^T$, and each ε_i follows a p-dimensional normal distribution with mean 0 and variance covariance matrix Σ . The second linear model is

$$\eta = \tilde{X} \tilde{\Gamma}_\eta + \zeta, \quad (13)$$

where $\tilde{X}_{(n \times (K+q))} = \{(c_1^T, x_1^T), \dots, (c_n^T, x_n^T)\}^T = (CX)$ (the ith row of C is c_i and the ith row of X is x_i), $\tilde{\Gamma}_{\eta((K+q) \times m)} = \{A^T, \Gamma_\eta^T\}$, and $\zeta = \{\zeta_1, \dots, \zeta_n\}^T$. Each ζ_i follows an m-dimensional normal distribution with mean 0 and variance covariance matrix Ψ .

Based on the results for the multivariate linear model, MLEs for $\tilde{\Lambda}_y$ and Σ are

$$\hat{\Lambda}_y = (\tilde{\eta}^T \tilde{\eta})^{-1} \tilde{\eta}^T Y \quad (14)$$

and

$$\widehat{\Sigma} = \text{diag} \left(\frac{1}{n} (Y^T Y - Y^T \tilde{\eta} (\tilde{\eta}^T \tilde{\eta})^{-1} \tilde{\eta}^T Y) \right) = \text{diag} \left(\frac{1}{n} (Y^T Y - Y^T \tilde{\eta} \hat{\Lambda}_y) \right), \quad (15)$$

where

$$\tilde{\eta}^T \tilde{\eta} = \begin{pmatrix} 1^T 1 & 1^T \eta \\ \eta^T 1 & \eta^T \eta \end{pmatrix}, \quad (16)$$

$$\tilde{\eta}^T Y = (1^T Y, \eta^T Y)^T. \quad (17)$$

MLEs for Γ_{η} and Ψ are

$$\hat{\Gamma}_{\eta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \eta \quad (18)$$

and

$$\hat{\Psi} = \left(\frac{1}{n} (\eta^T \eta - \eta^T \tilde{X} \hat{\Gamma}_{\eta}) \right), \quad (19)$$

where

$$\tilde{X}^T \tilde{X} = \begin{pmatrix} C^T C & C^T X \\ X^T C & X^T X \end{pmatrix}, \quad (20)$$

$$\tilde{X}^T \eta = (C^T \eta, X^T \eta)^T. \quad (21)$$

Finally, if c_i , $i = 1, \dots, n$ were observed, the MLE for π would be

$$\hat{\pi} = \sum_{i=1}^n c_i / n = C^T 1 / n. \quad (22)$$

Because MLEs can be expressed as linear functions of $Y^T Y$, $Y^T 1$, $\eta^T 1$, $\eta^T \eta$, $\eta^T Y$, $C^T 1$, $C^T C$, $C^T X$, $C^T \eta$, and $X^T \eta$, they are complete data sufficient statistics for this model. Calculating the conditional expectation of (8) is equivalent to calculating the conditional expectation of these sufficient statistics.

3. Model Estimation

3.1. The Monte Carlo EM Algorithm

The EM algorithm maximizes the conditional expectation of the complete data log likelihood (8) given observed data y_c, u_d , and x . In the E step, we calculate the conditional expectations of sufficient statistics with respect to the joint conditional distribution of y_d, η and c . Since this joint conditional distribution is unknown, the conditional expectations of sufficient statistics can not be obtained analytically. Because this joint conditional distribution is hard to sample directly, we use the Gibbs sampler to draw random numbers from this joint conditional distribution, and approximate the conditional expectations by the sample means.

With initial values for model parameters, $\theta^{(0)} = \{b^{(0)}, \Lambda_y^{(0)}, \Sigma^{(0)}, A^{(0)}, \Gamma_\eta^{(0)}, \Psi^{(0)}, \pi^{(0)}\}$, at iteration $(t + 1)$ the MCEM algorithm works as follows,

E step—Given current parameter estimates, samples from the joint conditional distributions, $[(y_{di}, c_i, \eta_i) | y_{ci}, u_{di}, x_i], i = 1, \dots, n$, are obtained by using the Gibbs sampler (see appendix A for details). Then conditional expectations of sufficient statistics are approximated by corresponding sample means from equations (47) to (57) in Appendix B.

M step—Parameter estimates are updated by equation (49) to (57) in Appendix C.

Iteratively, the above E and M steps form the Monte Carlo EM algorithm.

3.2. Monitor the Convergence of MCEM Algorithm

Because of variations due to Monte Carlo simulation introduced in the E step, the log likelihood of an MCEM sequence will fluctuate around a stationary point even on convergence. Thus convergence diagnosis of the MCEM algorithm is more difficult than a deterministic EM, which can be terminated when the difference of observed data log likelihood between two consecutive iterates is within a desired level. To monitor the convergence of the MCEM algorithm, we plot the observed data log likelihood ratio against iterations. If the plot shows a curve decreasing from a relatively large value to zero and then fluctuating around zero within the extent that can be expected from the simulation size, then the MCEM can be treated as approximately converged. Although the observed data log likelihood can not be obtained easily with enough accuracy, bridge sampling [29] provides an efficient way to calculate the ratio of observed data log likelihoods between two iterations.

Using bridge sampling, the observed data log likelihood ratio for subject i can be calculated as follows

$$\hat{r}(\theta_1, \theta_2) = \frac{\frac{1}{n_1} \sum_{l=1}^{n_1} \left[L(\theta_2 | y_{di}, y_{ci}, x_i, \eta_{i,l}^{(1)}) / L(\theta_1 | y_{di}, y_{ci}, x_i, \eta_{i,l}^{(1)}) \right]^{1/2}}{\frac{1}{n_2} \sum_{l=1}^{n_2} \left[L(\theta_1 | y_{di}, y_{ci}, x_i, \eta_{i,l}^{(2)}) / L(\theta_2 | y_{di}, y_{ci}, x_i, \eta_{i,l}^{(2)}) \right]^{1/2}}, \quad (23)$$

where θ_h includes all the parameters for model $h, h = 1, 2, \{\eta_{i,l}^{(h)}\}, l=1, \dots, n_h$ are random samples from the conditional distribution of $f(\eta_i | \theta_h, y_{di}, y_{ci}, x_i), h = 1, 2$, and

$$L(\theta_h|y_{di}, y_{ci}, x_i, \eta_{i,l}^{(1)}) = \left[\sum_{k=1}^K \pi_k f(y_i|\eta_i) f_k(\eta_i|x_i, c_{ik}=1) \right]. \quad (24)$$

Let

$$L_{11}=L(\theta_1|y_{di}, y_{ci}, x_i, \eta_{i,l}^{(1)}), L_{21}=L(\theta_2|y_{di}, y_{ci}, x_i, \eta_{i,l}^{(1)}), L_{12}=L(\theta_1|y_{di}, y_{ci}, x_i, \eta_{i,l}^{(2)}), L_{22}=L(\theta_2|y_{di}, y_{ci}, x_i, \eta_{i,l}^{(2)}).$$

. Then

$$\hat{r}(\theta_1, \theta_2) = \frac{\frac{1}{n_1} \sum_{l=1}^{n_1} [L_{21,l}/L_{11,l}]^{1/2}}{\frac{1}{n_2} \sum_{l=1}^{n_2} [L_{12,l}/L_{22,l}]^{1/2}}. \quad (25)$$

Finally, the log likelihood ratio can be estimated as

$$\hat{g}(\theta_2, \theta_1) = \sum_{i=1}^{n_0} \log \hat{r}(\theta_1, \theta_2). \quad (26)$$

4. Model Comparison

Selecting the number of latent classes in mixture models is an important but difficult task [5, ch. 6]. In this paper, we use the Bayesian information criterion (BIC) to compare models with different number of latent classes. BIC can be defined by

$$BIC = -2 \ln L(\hat{\theta}_h) + s * \ln(n), \quad (27)$$

where $\hat{\theta}_h$ is the final set of estimates for all parameters in model h , $\ln L(\hat{\theta}_h)$ is the observed data log likelihood for model h , s is the number of parameters, and n is the number of subjects. Given models with different number of latent classes, the one with the smallest value of BIC is preferred. Since s and n are already available, the key part for BIC is to calculate the observed data log likelihood. For our model, it can be expressed as follows

$$L(\theta_h|y_{di}, y_{ci}, x_i) = \sum_{i=1}^n \ln \left\{ \sum_{k=1}^K \pi_{(k,h)} \left[\int f(y_i|\eta_i, \theta_h) f(\eta_i|x_i, c_{ik}=1) d\eta_i \right] \right\}. \quad (28)$$

Notice that integrations are involved in the observed data log likelihood. Here, the Monte Carlo method is used to approximate these integrations. Suppose we have L samples, $\{\eta_{ik,1}, \dots, \eta_{ik,L}\}$, from $f(\eta_i|x_i, c_{ik}=1)$, $i = 1, \dots, n$; $k = 1, \dots, K$, then the observed data log likelihood can be approximated as follows

$$L(\theta_h|y_{di}, y_{ci}, x_i) = \sum_{i=1}^n \ln \left\{ \sum_{k=1}^K \pi_{(k,h)} \frac{1}{L} \sum_{l=1}^L f(y_i|\eta_{(ik,l)}, \theta_h) \right\}. \quad (29)$$

5. Numerical Illustrations

In this section, simulation studies are used to investigate the performance of the MCEM algorithm in recovering the model's parameters and in comparison with several likelihood based estimation methods implemented in Mplus 5.0. Also, to illustrate the practical usefulness of our model, we apply it to a real data set. By using our model, subjects from

this real data set are clustered into different subgroups based on their perceived social support and social desirability. Then subgroup differences in drug abuse and HIV risk behaviors are investigated, and some interesting differences are observed.

5.1. Simulation Study I

The purpose of this section is to empirically evaluate the properties of the MCEM algorithm for model estimation. In this simulation study, we have three continuous covariates, two latent classes and two latent factors. The first latent factor is defined by five dichotomous items and the second factor is defined by five continuous items. So totally we have ten observed responses, out of which five are dichotomous and five are continuous. Covariates are centered and fixed during all the simulations. For this simulation setup, the corresponding Gibbs sampler mixes very fast with autocorrelation decreasing to near zero for a lag of five iterations (figure 1). To be more conservative, the first 200 iterations are discarded as burn-in iterations and then random samples are collected after each 20 iterations until we have enough samples. The number of random samples starts at 150 for each subject, and increases 5 after each iteration.

A few diagnosis runs with 1000 subjects are conducted for the MCEM algorithm to determine the number of iterations for it to achieve convergence. A plot of observed data log likelihood ratio against iterations is used to diagnose convergence and is presented in figure 2. The left part of the figure gives the overall plot, while the right part gives more detail on later iterations. From figure 2 we can see that the log likelihood ratio drops very fast at the beginning and then becomes flat. After 80 iterations, log likelihood ratios become stable and the values are very small, so we can consider the MCEM algorithm as converged around 80 iterations. In the following simulation study, 100 iterations are used to produce MLE for parameters.

To evaluate the accuracy of parameter estimates, we repeat the simulation 30 times with 1000 subjects each time and obtained 30 sets of estimates. The bias of parameter estimation, which is the difference between true value and the mean of 30 estimates, and root mean square error (RMSE) of 30 estimates are calculated as a measure of accuracy. These values are included in Table 1. Table 1 indicates that the proposed MCEM algorithm can recover the true parameters satisfactorily. Estimation for parameters associated with continuous responses are better than estimations for corresponding parameters associated with binary responses. This can be explained by the fact that binary responses contain much less information compared with continuous responses.

5.2. Simulation study II

As pointed out by one of the reviewers, Mplus [25] has implemented several approaches to estimate the model. According to Muthén and Muthén [30], these estimation algorithms include fully MCMC Bayes, and three other likelihood based approaches based on either Gauss-Hermite or Monte Carlo integrations. These likelihood approaches include (1) expectation maximization (EM); (2) accelerated EM (EMA); and (3) direct optimization of the observed-data log likelihood. However, technical details about these estimation algorithms are not publicly available, so we can only empirically evaluate these algorithms by simulation studies. Because the fully MCMC Bayes approach is a totally different modeling framework and involves numerous special issues, it will not be covered here.

A number of simulation studies covering various situations were conducted to investigate the computational properties of these likelihood-based estimation approaches implemented in Mplus. These simulation studies suggest that (1) direct optimization of the observed data log likelihood is computationally much less efficient than EM type algorithms; (2) the

Monte Carlo based EM type algorithm implemented in Mplus is much less stable or accurate than those based on (adaptive) Gauss-Hermite integration. When the Monte Carlo integration is used, the log likelihood is more likely to decrease and the algorithm has more difficulty converging; (3) the accelerated EM (EMA) algorithm converges much faster than the EM algorithms while still enjoying stable convergence properties. In summary, the (adaptive) Gauss-Hermite quadrature based EMA algorithm is the best estimation approach implemented in Mplus for the model we propose here, and it will be used as a benchmark in comparison with our MCEM algorithms.

In the following studies, we use several simulations to compare our MCEM algorithm with the (adaptive) Gauss-Hermite quadrature based EMA (G-H EMA) algorithm implemented in Mplus. The comparisons mainly focus on estimation accuracy. Because computational efficiency in terms of computation time can be affected by many factors, this will not be compared here. It has been widely recognized that Monte Carlo integration is more efficient than numerical quadrature when the number of integrations is large.

These simulation studies include different latent factor structures, including simple, quasi-simple and bifactor models. The same starting values are used to ensure a fair comparison. All the computations in Mplus are terminated normally with the default convergence criterion. Estimation accuracy is measured by root mean square error (RMSE) across all the model parameters, and they are presented in Table 2. Results in Table 2 indicate that when the latent factor structure is simple, estimation generated by both algorithms is very good with RMSE around 0.06. However, as the latent factor structure becomes more complicated, estimates generated by the G-H EMA algorithm become worse as compared to the MCEM algorithm. For the quasi-simple structure model, the RMSE for G-H EMA is almost two times larger than the RMSE for MCEM, and similarly it is almost three times larger for the bifactor structure.

5.3. Real Data Example

In this section we use a real data example to illustrate our model in practical research. The target population in this data set is homeless people. Stein et al. [31] used this data set to study the associations among situational and psychosocial variables related to health, health care, and HIV/AIDS risk behaviors among homeless men. Following this study, we want to see whether subjects could be clustered into meaningful subgroups based on their perceived social support and social desirability. Perceived social support is measured by 5 continuous items. Social desirability is measured by 5 dichotomous items. Perceived social support and social desirability are assumed to be correlated. Educational and religious levels are included as covariates and are centered. Thus the model we proposed here is a mixture of two factor simple structure CFA models with covariates. To compare models with different number of latent classes, observed data log likelihoods and BICs are calculated with 1000 random samples, and are shown in Table 3. In Table 3 we can see that a model with 3 latent classes has the smallest BIC, so this model should be selected compared with the other three models. Based on this model, subjects are classified into three subgroups based on their posterior probability of belonging to each subgroup. There are 147 subjects in the first group, 189 in the second group and 283 in the third group. The first group is used as a baseline group. Compared with the first group, the second group has a mean that is 0.23 lower on perceived social support but a mean that is 1.22 higher on social desirability; the third group has roughly the same level of perceived social support (the difference is only 0.06) but a mean that is 2.64 greater on social desirability. Thus the main differences among these three subgroups involve social desirability. The effects of educational and religious levels on perceived social support are very small with regression coefficients 0.004 and 0.001 respectively. As for social desirability, educational and religious levels have minor negative effects with the regression coefficients -0.10 and -0.02 respectively.

Next we want to know whether these subgroups have any other meaningful differences besides social desirability. To answer this question, we compare the groups on drug abuse and HIV risk behaviors, variables that were not part of the class determination. Several interesting differences are observed. From group 1 to 3, the proportion that used alcohol in the last 6 months are 0.62, 0.58, and 0.45 respectively. The difference between group 1 and 2 is not statistically significant, but the difference between group 3 and group 1 and the difference between group 3 and group 2 are statistically significant. Marijuana use in the last 6 months is also statistically different among these three groups. The proportion of those using marijuana in the last 6 months decreases from group 1 to group 3. These results indicate that alcohol use and marijuana use are negatively correlated with social desirability. Overall 72 percent of homeless people included in this study are current smokers. The proportion of current smoker are 0.72, 0.67 and 0.75 for group 1, 2 and 3 respectively. But these differences are not statistically different.

The proportion of subjects using cocaine and heroin in the last 6 months also differs among groups. There is statistically significantly more use in groups 1 and 2 than in group 3. There are also more subjects in the first two groups who injected cocaine or heroin in the last 6 months, as compared to the third group. The difference is statistically significant for heroin injection but not significant for cocaine injection.

6. Discussion

In this paper an advanced finite mixture factor analysis model with covariates was proposed. Under our modeling framework, latent structures can be specified for mixed continuous and binary responses. Simulation studies indicate that MCEM algorithm can recover the true model parameters satisfactorily, and outperforms estimation algorithms implemented in Mplus in terms of estimation accuracy when the latent factor structure becomes more complex. Since the technical details are not publicly available for these methods implemented in Mplus, it is worth to note that these difference maybe due to the implementation not the methods. The real data example suggests that our model can provide efficient and useful modeling techniques for high dimensional mixed data. By treating random growth rates as latent factors, this modeling framework can also be used to model binary longitudinal data with heterogenous growth rates.

Besides these positive properties, this model also has its own limitations. First, under this modeling framework, the assumption that some latent structural parameters (Λ and Ψ) are the same for different subgroups can not be easily released. Building a more general model without these assumptions will be practically useful, but model estimation, identification and interpretation will be more challenging. Second, simulation study indicates that the proposed MCEM algorithm has a very slow convergence rate, mainly because of the large fraction of missing information. Accelerating techniques such as Nesting EM and Parameter Expansion EM can be used to speed up the convergence rate. If it makes sense, it may also be valuable to fix class membership for some cases to speed up convergence. Third, in practical research, ordinal and nominal responses are also very common, so an extension to include ordinal or nominal responses would be useful and should be relatively easy.

References

1. Muthén B. Latent variable modeling in heterogeneous populations. *Psychometrika*. 1989; 54(4): 557–585.
2. Baek J, McLachlan G, Flack L. Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. *IEEE transactions on pattern analysis and machine intelligence*. 2010; 32:1298–1309. [PubMed: 20489231]

3. Hinton G, Dayan P, Revow M. Modeling the manifolds of images of handwritten digits. *Neural Networks, IEEE Transactions on*. 2002; 8(1):65–74.
4. Boehning D, Seidel W, Alfo M, Garell B, Patilea V, Walther G. Advances in mixture models. *Computational Statistics & Data Analysis*. 2007; 51(11):5205–5210.10.1016/j.csda.2006.10.025
5. McLachlan, G.; Peel, D. *Finite Mixture Models*. John Wiley & Sons; 2000.
6. Arminger G, Stein P, Wittenberg J. Mixtures of conditional mean- and covariance-structure models. *Psychometrika*. 1999; 64(4):475–494.
7. Dolan C, van der Maas H. Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*. 1998; 63(3):227– 253.
8. Jedidi K, Jagpal H, DeSarbo W. STEMM: A general finite mixture structural equation model. *Journal of Classification*. 1997; 14(1):23–50.
9. Yung YF. Finite mixtures in confirmatory factor-analysis models. *Psychometrika*. 1997; 62(3):297–330.
10. Ghahramani, Z.; Hinton, G. Tech Rep CRG-TR-96-1. University of Toronto; 1997. The EM algorithm for mixtures of factor analyzers.
11. Lubke G, Muthén B. Investigating population heterogeneity with factor mixture models. *Psychological Methods*. 2005; 10:21–39. [PubMed: 15810867]
12. Bauer DJ. Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*. 2007; 42(4):757–786.
13. Hipp J, Bauer D. Local solutions in the estimation of growth mixture models. *Psychological Methods*. 2006; 11(1):36–53.10.1037/1082-989X.11.1.36 [PubMed: 16594766]
14. Fokoue E. Mixtures of factor analyzers: an extension with covariates. *Journal of Multivariate Analysis*. 2005; 95(2):370–384.10.1016/j.jmva.2004.08.004
15. Lee S, Song X. Bayesian model selection for mixtures of structural equation models with an unknown number of components. *British Journal of Mathematical and Statistical Psychology*. 2003; 56(1):145–165. [PubMed: 12803828]
16. Hoshino T. Bayesian inference for finite mixtures in confirmatory factor analysis. *Behaviormetrika*. 2001; 28(1):37–63.
17. McLachlan, G.; Peel, D. Mixtures of factor analyzers. *Proceedings of the Seventeenth International Conference on Machine Learning*; Citeseer. 2000.
18. McLachlan G, Peel D, Bean R. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*. 2003; 41(3–4):379–388.
19. Utsugi A, Kumagai T. Bayesian analysis of mixtures of factor analyzers. *Neural Computation*. 2001; 13(5):993–1002. <http://dx.doi.org/10.1162/08997660151134299>. [PubMed: 11359641]
20. Zhao JH, Yu PLH. Fast ML estimation for the mixture of factor analyzers via an ECM algorithm. *IEEE Transactions on Neural Networks*. 2008; 19(11):1956–1961.10.1109/TNN.2008.2003467 [PubMed: 19000964]
21. Zhou X, Liu X. The EM algorithm for the extended finite mixture of the factor analyzers model. *Computational Statistics & Data Analysis*. 2008; 52(8):3939–3953.10.1016/j.csda.2008.01.023
22. Hancock, G.; Samuelsen, K. *Advances in latent variable mixture models*. Information Age Publishing; 2007.
23. Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*. 1999; 55(2):463–469. [PubMed: 11318201]
24. Muthén, B.; Asparouhov, T. *Growth mixture modeling: Analysis with non-Gaussian random effects*. Vol. chap 6. Chapman & Hall/CRC Press; 2009. p. 143-165.
25. Muthén, Mplus 5.0: Computer software. Los Angeles: 2008.
26. McCulloch C. Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*. 1994; 89 (425):330–335.
27. Chan JSK, Kuk AYC. Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*. 1997; 53 (1):86–97.
28. Rabe-Hesketh, S.; Skrondal, A. Generalized linear-mixed effects models. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. *Longitudinal Data Analysis*. Boca Raton: CRC Press; 2009.

29. Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*. 1996; 6:831– 860.
30. Muthén, Muthén. *Mplus User's Guide*. 6. Los Angeles, CA: 2010.
31. Stein JA, Nyamathi AM, Zane JJ. Situational, Psychosocial, and Physical Health-Related Correlates of HIV/AIDS Risk Behaviors in Homeless Men. *American Journal of Mens Health*. 2009; 3(1):25–35. [10.1177/1557988307307862](https://doi.org/10.1177/1557988307307862)
32. Bartlett M. An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*. 1951:107–111.

A. Gibbs Sampler

With parameter estimation from the t th iteration of MCEM algorithm,

$\theta^{(t)} = \{b^{(t)}, \Lambda_y^{(t)}, \sum^{(t)}, A^{(t)}, \Gamma_\eta^{(t)}, \Psi^{(t)}, \pi^{(t)}\}$, the Gibbs sampler works as follows. To simplify equations, we will drop the superscripts for current parameter estimates. First we initialize the Gibbs sampler

- generate $c_i^{(0)}$ from its marginal distribution $p(c_i) \sim \text{Mul}(\pi_i)$;
- generate $\eta_i^{(0)}$ from $p(\eta_i^{(0)} | c_i, x_i) \sim N_m(Ac_i + \Gamma_\eta x_i, \Psi)$.

Next we run the following steps iteratively, and in the l th iteration of the Gibbs sampler

1. generate $y_{di}^{(l)}$ from $p(y_{di} | \eta_i) \sim N_d(\Lambda_y \eta_i, \Sigma)$, let $y_i^{(l)} = \{y_{ci}, y_{di}^{(l)}\}$.
2. generate $c_i^{(l)}$ from $p(c_i | y_i^{(l)}, x_i) \sim \text{Mul}(\tau_i^{(l)})$, where

$$\tau_{ik}^{(l)} = P(c_{ik} = 1 | y_i^{(l)}, x_i) = \pi_{ik} N_p[\Lambda_y(\alpha_k + \Gamma_\eta x_i), \Lambda_y \Psi \Delta_y^T + \sum] / p(y_i^{(l)} | x_i) \quad (30)$$

and

$$p(y_i^{(l)} | x_i) = \sum_{k=1}^K \pi_{ik} N_p[\Lambda_y(\alpha_k + \Gamma_\eta x_i), \Lambda_y \Psi \Delta_y^T + \sum]. \quad (31)$$

3. generate $\eta_i^{(l)}$ from $p(\eta_i | y_i^{(l)}, c_i^{(l)}, x_i) \sim N_m(\mu_\eta^{(l)}, \sum_\eta^{(l)})$, where

$$\mu_\eta^{(l)} = Ac_i^{(l)} + \Gamma_\eta x_i + \Lambda_y \Psi (\Lambda_y \Psi \Delta_y^T + \sum)^{-1} (y_i^{(l)} - \Lambda_y [Ac_i^{(l)} + \Gamma_\eta x_i]) \quad (32)$$

and

$$\sum_\eta^{(l)} = \Psi - (\Lambda_y \Psi)^T (\Lambda_y \Psi \Delta_y^T + \sum)^{-1} \Lambda_y \Psi. \quad (33)$$

Rearranging the terms for μ_η , we get

$$\mu_\eta^{(l)} = [I - \Lambda_y \Psi (\Lambda_y \Psi \Delta_y^T + \sum)^{-1} \Lambda_y] (Ac_i^{(l)} + \Gamma_\eta x_i) + (\Lambda_y \Psi)^T (\Lambda_y \Psi \Delta_y^T + \sum)^{-1} y_i^{(l)}. \quad (34)$$

Using the Woodbury's identity [32], we have

$$[I - \Lambda_y \Psi (\Lambda_y \Psi \Lambda_y^T + \Sigma)^{-1} \Lambda_y] = [I + \Psi \Lambda_y^T \Sigma^{-1} \Lambda_y]^{-1} = [\Psi^{-1} + \Lambda_y^T \Sigma^{-1} \Lambda_y]^{-1} \Psi^{-1}. \quad (35)$$

Let $V_{m \times m} = [\Psi^{-1} + \Lambda_y^T \Sigma^{-1} \Lambda_y]^{-1}$, and $W_{p \times p} = (\Lambda_y \Psi \Lambda_y^T + \Sigma)^{-1}$ then

$$\mu_\eta = V \Psi^{-1} (A c_i^{(l)} + \Gamma_\eta x_i) + (\Lambda_y \Psi)^T W y_i^{(l)} \quad (36)$$

and

$$\sum_\eta = V. \quad (37)$$

Iteratively, above steps form the Gibbs sampler, and samples generated by Gibbs sampler form a Markov Chain. To obtain samples much closer to the targeted distribution with less correlation, we will discard some of the initial samples (for example the first 1000 iterations) and take samples after every, say 10th, iteration. Finally we can get approximately

independent samples, $\{[y_{di}^{(l)}, c_i^{(l)}, \eta_i^{(l)}], l = 1, \dots, L\}$, from the joint conditional distributions $p(y_{di}, c_i, \eta_i | y_{ci}, u_{di}, x_i), i = 1, \dots, n$

B. Conditional Expectation of Sufficient Statistics

With samples, $\{[y_{di}^{(l)}, c_i^{(l)}, \eta_i^{(l)}], l = 1, \dots, L\}$, from the joint conditional distributions $p(y_{di}, c_i, \eta_i | y_{ci}, u_{di}, x_i), i = 1, \dots, n$, we approximate the conditional expectation of $Y^T Y, Y^T 1, \eta^T 1, \eta^T \eta, \eta^T Y, C^T 1, C^T C, C^T X, C^T \eta$, and $X^T \eta$ by corresponding sample means. In the following, to simplify notation we will use $E(S)$ for $E(S | y_c, u_d, x)$.

1. for $i = 1, \dots, n$, and $k = 1, \dots, K$,

$$p_{ik} = E[c_{ik}] = P(c_{ik} = 1 | u_{di}, x_i) = \sum_{l=1}^L c_{ik}^{(l)} / L, \quad (38)$$

$$S_c = E[C^T 1] = E \left[\sum_{i=1}^n c_i \right] = \sum_{i=1}^n (p_i), \quad (39)$$

$$S_{cc} = E[C^T C] = E \left[\sum_{i=1}^n c_i c_i^T \right] = \sum_{i=1}^n \text{diag}(p_i), \quad (40)$$

where $p_i = (p_{i1}, \dots, p_{ik})$.

$$S_{cx} = E[C^T X] = E \left[\sum_{i=1}^n c_i x_i^T \right] = \sum_{i=1}^n p_i x_i^T. \quad (41)$$

2. for $i = 1, \dots, n$,

$$S_{1\eta} = E[1^T \eta] = \sum_{i=1}^n E[\eta_i] = \sum_{i=1}^n \left[\sum_{l=1}^L \eta_i^{(l)} / L \right], \quad (42)$$

$$S_{c\eta} = E[C^T \eta] = \sum_{i=1}^n E[c_i \eta_i^T] = \sum_{i=1}^n \left[\sum_{l=1}^L c_i^{(l)} (\eta_i^{(l)})^T / L \right], \quad (43)$$

$$S_{x\eta} = E[X^T \eta] = \sum_{i=1}^n E[x_i \eta_i^T] = \sum_{i=1}^n \left[\sum_{l=1}^L x_i^{(l)} (\eta_i^{(l)})^T / L \right], \quad (44)$$

$$S_{\eta\eta} = E[\eta^T \eta] = \sum_{i=1}^n E[\eta_i \eta_i^T] = \sum_{i=1}^n \left[\sum_{l=1}^L \eta_i^{(l)} \cdot (\eta_i^{(l)})^T / L \right]. \quad (45)$$

3. for $i = 1, \dots, n$, let $y_i^{(l)} = \{y_{ci}, y_{di}\}$, then we have

$$S_{\eta y} = E[\eta^T y] = \sum_{i=1}^n E[\eta_i y_i^T] = \sum_{i=1}^n \left[\sum_{l=1}^L \eta_i^{(l)} (y_i^{(l)})^T / L \right], \quad (46)$$

$$S_{1y} = E[1^T y] = \sum_{i=1}^n E[y_i] = \sum_{i=1}^n \left[\sum_{l=1}^L y_i^{(l)} \right], \quad (47)$$

$$S_{yy} = E[y^T y] = \sum_{i=1}^n E[y_i y_i^T] = \sum_{i=1}^n \left[\sum_{l=1}^L y_i^{(l)} (y_i^{(l)})^T / L \right]. \quad (48)$$

C. Updating Equations for the M step

1. Update $\tilde{\Lambda}_y$, and Σ by

$$\hat{\tilde{\Lambda}}_y^{(t+1)} = (S_{\tilde{\eta}\tilde{\eta}})^{-1} S_{\tilde{\eta}y}, \quad (49)$$

$$\widehat{\Sigma}^{(t+1)} = \text{diag} \left(\frac{1}{n} (S_{yy} - S_{\tilde{\eta}y}^T \hat{\tilde{\Lambda}}_y^{(t+1)}) \right), \quad (50)$$

where

$$S_{\tilde{\eta}\tilde{\eta}} = \begin{pmatrix} 1^T 1 & S_{1\eta} \\ S_{1\eta}^T & S_{\eta\eta} \end{pmatrix} \quad (51)$$

and

$$S_{\tilde{\eta}Y} = (S_{1Y}, S_{\eta T_Y}) \quad (52)$$

2. Update $\tilde{\Gamma}_\eta = \{A^T, \Gamma_\eta^T\}$, and Ψ by

$$\hat{\tilde{\Gamma}}_\eta^{(t+1)} = (S_{\tilde{x}\tilde{x}})^{-1} S_{\eta\tilde{x}}^T, \quad (53)$$

$$\hat{\Psi}^{(t+1)} = \text{diag} \left(\frac{1}{n} (S_{\eta\eta} - S_{\eta\tilde{x}} \hat{\tilde{\Gamma}}_\eta^{(t+1)}) \right), \quad (54)$$

where

$$S_{\tilde{x}\tilde{x}} = \begin{pmatrix} S_{cc} & S_{cx} \\ S_{cx}^T & S_{xx} \end{pmatrix} \quad (55)$$

and

$$S_{\eta\tilde{x}} = (S_{\eta c}, S_{\eta x}). \quad (56)$$

3. Update the mixture proportion π by

$$\hat{\pi}^{(t+1)} = S_c / n. \quad (57)$$

Figure 1.
Auto Correlation of Random Samples Generated by Gibbs Sampler

Figure 2.
Convergence Diagnosis Plot for Simulation

Table 1

Bias and Error for Simulation Study I

Para.	True	Bias	RMSE	Para.	True	Bias	RMSE
b_1	0	0.027	0.102	σ_6	0.35	0.000	0.028
b_2	0	0.023	0.096	σ_7	0.45	0.000	0.029
b_3	0	0.020	0.092	σ_8	0.55	0.007	0.022
b_4	0	0.028	0.128	σ_9	0.45	0.000	0.025
b_5	0	0.009	0.115	σ_{10}	0.35	0.002	0.016
b_6	0	0.010	0.056	A_{12}	2.5	0.102	0.127
b_7	0	0.018	0.042	A_{22}	3.0	0.014	0.057
b_8	0	0.012	0.058	Γ_{11}	1.3	0.033	0.075
b_9	0	0.024	0.063	Γ_{12}	0.9	0.008	0.040
b_{10}	0	0.030	0.073	Γ_{21}	1.3	0.038	0.066
λ_{21}	0.85	0.008	0.062	Γ_{22}	1.3	0.005	0.28
λ_{31}	0.95	0.024	0.058	ψ_{11}	0.8	0.032	0.107
λ_{41}	1.15	0.028	0.056	ψ_{12}	0.3	0.003	0.055
λ_{51}	1.25	0.095	0.068	ψ_{21}	0.3	0.003	0.055
λ_{72}	0.75	0.000	0.014	ψ_{22}	0.8	0.009	0.050
λ_{82}	0.95	0.002	0.018	π_1	0.3	0.000	0.006
λ_{92}	1.15	0.004	0.014	π_2	0.7	0.000	0.006
λ_{102}	1.35	0.005	0.014				

Table 2

Comparison of Estimation Accuracy in Terms of Root Mean Square Error in Simulation Study II

Estimation Algorithms	Simple Structure	Quasi-Simple Structure	Bifactor
G-H EMA (Mplus)	0.06149	0.1811	0.5333
Monte Carlo EM	0.06795	0.1062	0.1893

Table 3

Determination of Number of Homeless Classes

Number of Latent Classes	Log Likelihood	Number of Parameters	BIC
1	-5747.73	33	11707.59
2	-5694.84	34	11608.24
3	-5674.97	37	11587.78
4	-5671.28	40	11599.68