PSYCHOMETRIKA—VOL. 66, NO. 1, 133–152
MARCH 2001

A BAYESIAN ANALYSIS OF FINITE MIXTURES IN THE LISREL MODEL

HONG-TU ZHU AND SIK-YUM LEE

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG

In this paper, we propose a Bayesian framework for estimating finite mixtures of the LISREL model. The basic idea in our analysis is to augment the observed data of the manifest variables with the latent variables and the allocation variables. The Gibbs sampler is implemented to obtain the Bayesian solution. Other associated statistical inferences, such as the direct estimation of the latent variables, establishment of a goodness-of-fit assessment for a posited model, Bayesian classification, residual and outlier analyses, are discussed. The methodology is illustrated with a simulation study and a real example.

Key words: Bayesian analysis, finite mixtures, LISREL models, Gibbs sampler, conditional distributions, goodness-of-fit assessment, Bayesian classification, residual and outlier analyses.

1. Introduction

In general, a finite mixture model arises with a population which is a mixture of M components with associated probability densities $\{f_m, m = 1, \ldots, M\}$ and mixing proportions $\{\pi_m, m = 1, \ldots, M\}$. Such a situation is very common in many areas of applied statistics such as statistical pattern recognition, classification and clustering; see the survey paper by Redner and Walker (1984), and the excellent book by Titterington, Smith and Makov (1985), among others. In the literature, a variety of statistical methods have been proposed to analyze finite mixture models. Examples include the method of moments (Day, 1969; Lindsay, 1989; Lindsay & Basak, 1993), Bayesian and quasi-Bayes methods (Crawford, DeGroot, Kadane, & Small, 1992; Diebolt & Robert, 1994; Richardson & Green, 1997; Robert, 1996; Smith & Makov, 1978), the discriminant analysis approach (Do & McLachlan, 1984; Ganesalingam & McLachlan, 1981), and the maximum likelihood (Day, 1969; Hasselblad, 1966; Hathaway, 1985).

On the other hand, models for establishing substantive theory in behavioral and social sciences usually involve causal effects and correlations among manifest variables and the latent variables that cannot be measured by one single operationalization. Structural equation modeling (SEM) is an important method in finding the appropriate model and estimating the causal effects and the correlations, see Bentler (1983), Browne (1984), Jöreskog (1978), and Yuan and Bentler (1997), among others. Now, SEM represents a widely used multivariate method in behavioral, health and social sciences; see for examples, Bollen and Long (1993), Byrne (1994), Hoyle (1995) and the references therein. The most important factor accounting for the popularity of SEM is clearly due to the availability of the efficient computer softwares such as EQS (Bentler, 1992) and LISREL VIII (Jöreskog & Sörbom, 1996). Hence, a major trend of research in the field has been devoted to enlarge the scope of applicability of the LISREL model to non-standard situations; see for example, Lee and Poon (1992), Muthén (1989) and van Buuren (1997), among others.

Recently, a few important contributions on the maximum likelihood (ML) analysis of finite mixtures in structural equation models have been established. Jedidi, Jagpal and DeSarbo (1997a)

This research was supported by a Hong Kong UGC Earmarked grant CUHK 4026/97H. The authors are indebted to the Editor, the Associate Editor, and three anonymous reviewers for constructive comments in improving the paper, and also to ICPSR and the relevant funding agency for allowing the use of the data. The assistance of Michael K.H. Leung and Esther L.S. Tam is gratefully acknowledged.

Requests for reprints should be sent to S.Y. Lee, Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T. HONG KONG. E-Mail: sylee@sparc2.sta.cuhk.edu.hk

0033-3123/2001-1/1998-0694-A $00.75/0

analyzed the finite mixtures of multivariate regression and simultaneous equation models; while Jedidi, Jagpal and DeSarbo (1997b) considered a general finite mixtures of structural equation models and discussed the problem of model selection. In the estimation, both of the above articles used the EM algorithm (Dempster, Laird & Rubin, 1977) by treating the allocation variables as missing data. Yung (1997) considered the finite mixtures in the confirmatory factor analysis model and implemented an approximate scoring algorithm and an EM algorithm to solve the likelihood equation. He concluded that the approximate scoring algorithm is better than his EM algorithm which treated the allocation variables and the latent factor scores as missing. Some aspects of the maximum likelihood estimation have also been given by Dolan and van der Maas (1998). In their application of the EM algorithm, Jedidi et al. (1997b), did not give standard errors estimates. Yung pointed out that if the components are not well-separated, the appropriateness of the approximate information matrix in getting standard error estimates is questionable. Under this situation, the Hessian matrix of the log-likelihood function, which requires heavy computation to evaluate, is recommended for estimating the standard error estimates.

In this article, we propose a Bayesian approach as an alternative to the ML approach. Our development is based on finite mixtures in the well-known LISREL model. This complicated model is handled by augmenting the observed data of the manifest variables with the hypothetical missing data that associated with the latent variables and the allocation variables. On the basis of conjugate type prior distributions, the well-known Gibbs sampler (Geman & Geman, 1984) is implemented to generate a sequence of observations from the appropriate joint posterior distribution. As we will see later, the proposed Bayesian approach gives the joint Bayesian estimates of the mixing proportions, means of the underlying components, structural parameters in the covariance matrices, and direct latent variables scores corresponding to each individual. Standard errors estimates can be obtained easily. Still, obtaining point estimates is not the end of the our analysis. The simulated observations from the Gibbs sampler provide useful information for further statistical analyses such as the construction of confidence intervals, hypotheses testing, etc. To keep our presentation within a suitable length, we only discuss the issues on a goodness-of-fit assessment of a posited model, Bayesian classification, residual and outlier analyses in this paper.

The paper is organized as follows. Section 2 describes the finite mixtures in the LISREL model. The Bayesian theory for analysis of this model is presented in section 3. Statistical analyses based on the simulated observations from the Gibbs sampler are discussed in section 4. To illustrate the proposed procedure, results obtained from a simulation study and a real example are reported in section 5. A discussion is given in section 6. Some technical details are given in the Appendices.

2. Finite Mixtures in the LISREL Model

Suppose that the distribution of the $p \times 1$ random vector y is given by the probability density function of the form

$$f(y|\theta) = \sum_{m=1}^{M} \pi_m f_m(y|\mu_m, \theta_m), \tag{1}$$

where M is a given integer, π_m is the unknown mixing proportion such that $\pi_m > 0$ and $\pi_1 + \cdots + \pi_M = 1.0$, $f_m(y|\mu_m, \theta_m)$ is the multivariate normal density function with an unknown mean vector μ_m and a general covariance structure $\Sigma_m = \Sigma_m(\theta_m)$ that depends on an unknown parameter vector θ_m, and θ is the parameter vector that contains all unknown parameters in π_m, μ_m and θ_m, $m = 1, \ldots, M$.

The following well-known LISREL model for the random vector y conditional on the mth component is considered in this paper. The measurement equation of the model is given by

$$y = \mu_m + \Lambda_m \xi_m + \epsilon_m, \tag{2}$$

where μ_m is the intercept vector, Λ_m is the $p \times q$ factor loading matrix, ξ_m is the random vector of latent variables, and ϵ_m is the random vector of error measurements which is distributed according to $N[0, \Psi_m]$ and independent with ξ_m, where Ψ_m is a diagonal matrix. Moreover, let $\xi_m = (\xi_{m1}^T, \xi_{m2}^T)^T$; the structural equation of the model, which describes the causal effects among the latent variables, is defined as

$$\xi_{m1} = \Pi_m \xi_{m1} + \Gamma_m \xi_{m2} + \delta_m, \tag{3}$$

where ξ_{m1} and ξ_{m2} are $q_1 \times 1$ and $q_2 \times 1$ subvectors of ξ_m respectively, δ_m is a random vector that is independent with ξ_m, Π_m and Γ_m are unknown parameter matrices such that $\Pi_{0m}^{-1} = (I - \Pi_m)^{-1}$ exists and $|\Pi_{0m}|$ is independent with Π_m. Let the distributions of ξ_{m2} and δ_m be $N[0, \Phi_{2m}]$ and $N[0, \Psi_{\delta m}]$, respectively; where $\Psi_{\delta m}$ is a diagonal matrix. The parameter vector θ_m contains the free unknown parameters in Λ_m, Π_m, Γ_m, Φ_{2m}, $\Psi_{\delta m}$ and Ψ_m. The covariance structure of ξ_m is given by

$$\Sigma_{\xi m} = \begin{bmatrix} \Pi_{0m}^{-1}(\Gamma_m \Phi_{2m} \Gamma_m^T + \Psi_{\delta m})(\Pi_{0m}^{-1})^T & \Pi_{0m}^{-1} \Gamma_m \Phi_{2m} \\ \Phi_{2m} \Gamma_m^T (\Pi_{0m}^{-1})^T & \Phi_{2m} \end{bmatrix}, \tag{4}$$

and $\Sigma_m(\theta_m) = \Lambda_m \Sigma_{\xi m} \Lambda_m^T + \Psi_m$. Any of these unknown parameter matrices can be set invariant across components.

The identification of the model is an important problem in the analysis. For special models, the problem can be solved by imposing appropriate constraints on the unknown parameters. Based on substantive and theoretical considerations, these constraints may be defined by fixing certain parameters at preassigned known values or by imposing some linear or nonlinear constraints among the parameters. For clear discussion of the statistical method, we assume that the identification problem has been taken care of by various methods (see, e.g., Jedidi et al. 1997a; Jöreskog & Sörbom, 1996) and the underlying LISREL model is identified.

According to Redner and Walker (1984) and Yung (1997), three different types of observations can be sampled from the following distinct schemes: (a) The observations $\{y_i, i = 1, \ldots, N\}$ are sampled independently from the whole population. (b) For each m, predetermined N_m observations $\{y_{mj}, j = 1, \ldots, N_m\}$ are sampled from the m-th component separately and independently; hence $N = N_1 + \cdots + N_M$. (c) As in Scheme (a), the component memberships of the observations are identified after being drawn, so that all N_m are observed. Data obtained from Schemes (b) and (c) can be analyzed as a multiple-sample problem as described in LISREL VIII (Jöreskog & Sörbom, 1996), see also Lee and Tsui (1982). Under Scheme (a), the sample sizes of the components are unknown, hence the corresponding analysis is more difficult and has received the most attention in the literature. For brevity, we assume the observations are all obtained under the sampling Scheme (a), but the proposed methodology can be extended easily to the general mixed sampling scheme by adding appropriate terms to the expressions.

3. Bayesian Analysis of the Model

Let θ_{ym} be the vector of unknown parameters in Λ_m and Ψ_m, and $\theta_{\xi m}$ be the vector of unknown parameters in Π_m, Γ_m, Φ_{2m} and $\Psi_{\delta m}$. Let μ, π, θ_y and θ_ξ be the vectors that contain the unknown parameters in $\{\mu_1, \ldots, \mu_M\}$, $\{\pi_1, \ldots, \pi_M\}$, $\{\theta_{y1}, \ldots, \theta_{yM}\}$ and $\{\theta_{\xi 1}, \ldots, \theta_{\xi M}\}$, respectively; then $\theta = (\mu, \pi, \theta_y, \theta_\xi)$.

It is natural to introduce a group label z_i for the ith observation y_i as a latent *allocation variable*, and assume that it is independently drawn from the following distribution:

$$p(z_i = m) = \pi_m, \qquad \text{for } m = 1, \ldots, M. \tag{5}$$

Moreover, let $\mathbf{Y} = (y_1, \ldots, y_n)$ be the observed data matrix, $\mathbf{X} = (\xi^1, \ldots, \xi^n)$ be the matrix of latent vectors; and $\mathbf{Z} = (z_1, \ldots, z_n)$ be the matrix of allocation variables.

In a standard Bayesian analysis, we require to evaluate the complicated posterior distribution $p[\theta|\mathbf{Y}]$. However, if $\mathbf{Z}$ is observed,- the component of every y_i can be identified and the mixture model becomes the familiar multiple group model. In addition, if $\mathbf{X}$ is observed, the LISREL model will become the linear simultaneous equation model which is also comparatively easy to handle. Hence, in our Bayesian analysis, the observed data $\mathbf{Y}$ will be augmented with the latent data $\mathbf{X}$ and $\mathbf{Z}$ in the posterior analysis. In the following, we will concentrate on $p(\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y})$, the posterior distribution of $(\theta, \mathbf{X}, \mathbf{Z})$ given $\mathbf{Y}$. Since this distribution involves high dimensional integrals, direct evaluation of it is still intractable. Hence, a Gibbs sampler (Geman & Geman, 1984) is used to generate a sequence of observations from $p(\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y})$, and the Bayesian solution is then obtained on the basis of this sequence of observations. The basic algorithm of the Gibbs sampler is briefly given as below. At the r-th iteration with current values $\theta^{(r)}$, $\mathbf{X}^{(r)}$ and $\mathbf{Z}^{(r)}$:

Step (a): Generate $(\mathbf{X}^{(r+1)}, \mathbf{Z}^{(r+1)})$ from $p(\mathbf{X}, \mathbf{Z}|\mathbf{Y}, \theta^{(r)})$;
Step (b): Generate $\theta^{(r+1)}$ from $p(\theta|\mathbf{Y}, \mathbf{X}^{(r+1)}, \mathbf{Z}^{(r+1)})$.

Since $p(\mathbf{X}, \mathbf{Z}|\mathbf{Y}, \theta) = p(\mathbf{Z}|\mathbf{Y}, \theta)p(\mathbf{X}|\mathbf{Y}, \mathbf{Z}, \theta)$, Step (a) can be further decomposed into the following two steps:

Step (a.1): Generate $\mathbf{Z}^{(r+1)}$ from $p(\mathbf{Z}|\mathbf{Y}, \theta^{(r)})$;
Step (a.2): Generate $\mathbf{X}^{(r+1)}$ from $p(\mathbf{X}|\mathbf{Y}, \theta^{(r)}, \mathbf{Z}^{(r+1)})$.

Note that $p(\mathbf{Z}|\mathbf{Y}, \theta)$ is simpler than $p(\mathbf{Z}|\mathbf{Y}, \theta, \mathbf{X})$ and it does not involve $\mathbf{X}$.

3.1. Conditional Distributions

Since

$$p(\mathbf{Z}|\mathbf{Y}, \theta) = \prod_{i=1}^{n} p(z_i|y_i, \theta) \propto \prod_{i=1}^{n} p(z_i|\pi)p(y_i|z_i, \theta);$$

it can be shown that

$$p(z_i = m|y_i, \theta) = \frac{\pi_m f_m(y_i|\mu_m, \theta_m)}{f(y_i|\theta)}. \tag{6}$$

where $f_m(y_i|\mu_m, \beta_m)$ is the probability density function of $N[\mu_m, \Sigma_m(\theta_m)]$. Hence, simulating a $\mathbf{Z}$ from $p(\mathbf{Z}|\mathbf{Y}, \theta^{(r)})$ is not difficult. Because ξ^i are mutually independent and y_i are also mutually independent with given z_i, we have

$$p(\mathbf{X}|\mathbf{Y}, \theta, \mathbf{Z}) \propto p(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mu, \theta_y)p(\mathbf{X}|\mathbf{Z}, \theta_\xi) = \prod_{i=1}^{n} p(y_i|\xi^i, z_i, \mu, \theta_y)p(\xi^i|z_i, \theta_\xi).$$

Let $\Omega_m = \Sigma_{\xi m}^{-1} + \Lambda_m^T \Psi_m^{-1} \Lambda_m$; it follows from the definition of the model that (also see Appendix A):

$$p(\xi^i|y_i, z_i = m, \mu, \theta_y) \sim N[\Omega_m^{-1} \Lambda_m^T \Psi_m^{-1}(y_i - \mu_m), \Omega_m^{-1}]. \tag{7}$$

As a result, Step (a) can be completed on the basis of two simple and familiar distributions.

As it stands, the conditional distribution $p(\theta|\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ with given $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ required in Step (b) of the Gibbs sampler is very complicated; however, this difficulty can be overcome by assuming the following mild conditions on the prior distribution of θ. Clearly, the prior distribution of the mixing proportion π can be assumed to be independent with the prior distributions of μ, θ_y and θ_ξ. Like many Bayesian analyses (see, among others, Arminger & Muthén,

1998; Shi & Lee, 1998) the prior distribution of the mean vector μ can be assumed to be independent with the prior distributions of the parameters θ_y and θ_ξ in the covariance structures. Moreover, when $\mathbf{X}$ is given, the parameters in $\theta_{ym} = \{\Lambda_m, \Psi_m\}$ are the only parameters involved in the linear regression model (2) with the manifest variables in y; while the parameters in $\theta_{\xi m} = \{\pi_m, \Gamma_m, \Phi_{2m}, \Psi_m\}$ are the parameters involved in the other simultaneous equation model (3) with the latent variables. Hence, for convenience, we assume that the prior distributions of θ_y and θ_ξ are independent. Hence, $p(\theta) = p(\pi, \mu, \theta_y, \theta_\xi) = p(\pi)p(\mu)p(\theta_y)p(\theta_\xi)$. Moreover, from the definition of the model and the properties of $\mathbf{X}$, $\mathbf{Z}$ and θ, we have $p(\mathbf{Z}|\theta) = p(\mathbf{Z}|\pi)$, $p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, \theta) = p(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mu, \theta_y)p(\mathbf{X}|\mathbf{Z}, \theta_\xi)$. As a result, the joint distribution of all these random quantities can be expressed as

$$\begin{aligned} p(\theta, \mathbf{Z}, \mathbf{X}, \mathbf{Y}) &= p(\theta)p(\mathbf{Z}|\theta)p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, \theta) \\ &= p(\pi)p(\mu)p(\theta_y)p(\theta_\xi)p(\mathbf{Z}|\pi)p(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mu, \theta_y)p(\mathbf{X}|\mathbf{Z}, \theta_\xi). \end{aligned} \tag{8}$$

The prior distribution of π can be taken as the symmetric Dirichlet distribution; that is, $\pi \sim D(\alpha, \ldots, \alpha)$ with probability density function given by

$$p(\pi) = \frac{\Gamma(M\alpha)}{\Gamma(\alpha)^M}\pi_1^\alpha \cdots \pi_M^\alpha,$$

where $\Gamma(\cdot)$ is the Gamma function. Since $p(\mathbf{Z}|\pi) \propto \prod_{m=1}^M \pi_m^{n_m}$, it follows from (8) that the full conditional distribution for the weights in π remains Dirichlet in form:

$$p(\pi|\cdot) \propto p(\pi)p(\mathbf{Z}|\pi) \propto \prod_{m=1}^M \pi_m^{n_m+\alpha}, \tag{9}$$

where n_m is the total number of i such that $z_i = m$. Thus, $p(\pi|\cdot)$ is distributed as $D(\alpha + n_1, \ldots, \alpha + n_M)$, see Gelman, Carlin, Stern and Rubin (1995).

Let $\mathbf{Y}_m$ and $\mathbf{X}_m$ be the respective submatrices of $\mathbf{Y}$ and $\mathbf{X}$, such that all the ith column with $z_i \neq m$ are deleted. It is natural to assume that for $m \neq m'$, $(\mu_m, \theta_{ym}, \theta_{\xi m})$ and $(\mu_{m'}, \theta_{ym'}, \theta_{\xi m'})$ are independent. Hence, from (8), we have

$$p(\mu, \theta_y, \theta_\xi|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto \prod_{m=1}^M p(\mu_m)p(\theta_{ym})p(\theta_{\xi m})p(\mathbf{Y}_m|\mathbf{X}_m, \mu_m, \theta_{ym})p(\mathbf{X}_m|\theta_{\xi m}), \tag{10}$$

and we can treat (10) separately with each m. With $\mathbf{Z}$ given, the original complicated problem of finite mixtures reduces to a much simpler multi-sample problem. If there are no cross-group constraints, the analysis can be carried out separately with each individual sample. The EM algorithm (Jedidi et al. 1997a, 1997b) and the AS algorithm in Yung (1997) also have this feature.

According to the suggestions given in Raiffa and Schlaifer (1961), Lindley and Smith (1972), Lee (1981), and Broemeling (1985), the following commonly used conjugate type prior distribution can be used in situations where we have rough ideas about the hyper-parameters:

$$\begin{aligned} p(\mu_m) \sim N[\mu_0, \Sigma_0], \quad p(\psi_{mk}^{-1}) \sim \Gamma[\alpha_{0k}, \beta_{0k}], \\ p(\Lambda_{ymk}|\psi_{mk}) \sim N[\Lambda_{0mk}, \psi_{mk}H_{0mk}], \quad \text{for } k = 1, \ldots, p, \end{aligned} \tag{11}$$

where $\Gamma[\alpha_{ok}, \beta_{ok}]$ is the Gamma distribution with a shape hyper-parameter $\alpha_{0k} > 0$ and a scale hyper-parameter $\beta_{0k} > 0$ (see Zellner, 1971), and Λ_{ymk} is a $r_{mk} \times 1$ row vector that contains the unknown parameters in the k-th row of Λ_m. The scalars α_{0k}, β_{0k} and μ_0, the vector Λ_{0mk}, and the matrices H_{0mk} and Σ_0 are the hyper-parameters whose values are assumed to be given. For $k \neq h$, it is assumed that $(\psi_{mk}, \Lambda_{ymk})$ and $(\psi_{mh}, \Lambda_{ymh})$ are independent. It has been pointed out by the above cited work that the conjugate type prior distributions are sufficiently flexible in most applications; and for situations with a reasonable amounts of data available, the hyper-parameters

scarcely affect the analysis. In practice, the values of the hyper-parameters can be chosen on the basis of prior information obtained from previous experience or some preliminary data analyses.

To identify the covariance model, some appropriate elements in Λ_m may be fixed to pre-assigned known values. To cope with this situation, let $C_{ym} = (c_{mkj})$ be the index matrix such that $c_{mkj} = 0$ if λ_{mkj} is known and $c_{mkj} = 1$ if λ_{mkj} is unknown, $r_{mk} = \sum_{j=1}^{q} c_{mkj}$; $\mathbf{X}_{mk} = (\xi_{mk}^{i_1*}, \ldots, \xi_{mk}^{i_{n_m}*})$ be a submatrix of $\mathbf{X}_m = (\xi_m^{i_1}, \ldots, \xi_m^{i_{n_m}})$ such that all the j-th row with $c_{mkj} = 0$ are deleted; and $\mathbf{Y}_{mk}^T = (y_{mki_1}^*, \ldots, y_{mki_{n_m}}^*)$ with

$$y_{mki_s}^* = y_{mki_s} - \mu_{mk} - \sum_{j=1}^{q} \lambda_{mkj} \xi_m^{i_s j} (1 - c_{mkj}).$$

Let $\Omega_{mk} = (H_{0mk}^{-1} + \mathbf{X}_{mk}\mathbf{X}_{mk}^T)^{-1}$, $\nu_{mk} = \Omega_{mk}[H_{0mk}^{-1}\Lambda_{0mk} + \mathbf{X}_{mk}\mathbf{Y}_{mk}]$, and $\beta_{mk} = \beta_{0k} + 2^{-1}(\mathbf{Y}_{mk}^T\mathbf{Y}_{mk} - \nu_{mk}^T\Omega_{mk}^{-1}\nu_{mk} + \Lambda_{0mk}^T H_{0mk}^{-1}\Lambda_{0mk})$. Then, it can be shown that (see Appendix A) for $k = 1, \ldots, p$,

$$\begin{aligned} p(\gamma_{mk}|\mathbf{Y}_m, \mathbf{X}_m, \mu_m) &\sim \Gamma[n_m/2 + \alpha_{0k}, \beta_{mk}], \\ p(\Lambda_{ymk}|\mathbf{Y}_m, \mathbf{X}_m, \psi_{mk}, \mu_m) &\sim N[\nu_{mk}, \psi_{mk}\Omega_{mk}], \\ p(\mu_m|\mathbf{Y}_m, \mathbf{X}_m, \Lambda_m, \Psi_m) &\sim N[(\Sigma_0^{-1} + n_m\Psi_m^{-1})^{-1}(n_m\Psi_m^{-1}\overline{B}_m + \Sigma_0^{-1}\mu_0), (\Sigma_0^{-1} + n_m\Psi_m^{-1})^{-1}], \end{aligned} \tag{12}$$

where $\gamma_{mk} = \psi_{mk}^{-1}$, and $\overline{B}_m = \sum_{i:z_i=m}(y_i - \Lambda_m\xi^i)/n_m$ with $\sum_{i:z_i=m}$ denotes the summation with respect to those i such that $z_i = m$.

Consider the conditional distribution of $\theta_{\xi m}$ that is proportional to $p(\mathbf{X}_m|\theta_{\xi m})p(\theta_{\xi m})$. Let $\mathbf{X}_m = (\mathbf{X}_{m,1}^T, \mathbf{X}_{m,2}^T)^T$ where $\mathbf{X}_{m,1}$ and $\mathbf{X}_{m,2}$ are the submatrices of $\mathbf{X}_m$ corresponding to ξ_{m1} and ξ_{m2}, respectively. Since the distribution of ξ_{m2} only involves Φ_{2m}, we have $p(\mathbf{X}_{m,2}|\theta_m) = p(\mathbf{X}_{m,2}|\Phi_{2m})$. Moreover, it is assumed that the prior distribution of Φ_{2m} is independent with the prior distribution of $\Lambda_{\xi m} = (\Pi_m, \Gamma_m)$ and $\Psi_{\delta m}$. It follows that

$$p(\mathbf{X}_m|\theta_{\xi m})p(\theta_{\xi m}) \propto [p(\mathbf{X}_{m,1}|\mathbf{X}_{m,2}, \Lambda_{\xi m}, \Psi_{\delta m})p(\Lambda_{\xi m}, \Psi_{\delta m})][p(\mathbf{X}_{m,2}|\Phi_{2m})p(\Phi_{2m})].$$

Hence, the marginal conditional densities of $(\Lambda_{\xi m}, \Psi_{\delta m})$ and Φ_{2m} can again be treated separately.

Following a standard Bayesian procedure, see also Arminger and Muthén (1998); Shi and Lee (1998); we consider a conjugate type prior distribution for Φ_{2m} with $p(\Phi_{2m}^{-1}) \sim W[R_0, \rho_0, q_2]$, where $W[\cdot, \cdot, \cdot]$ denotes the Wishart distribution with probability density function proportional to (see Zellner, 1971)

$$|R_0|^{-\rho_0/2}|\Phi_{2m}^{-1}|^{(\rho_0 - q_2 - 1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}R_0^{-1}\Phi_{2m}^{-1}\right),$$

in which ρ_0 and the positive definite matrix R_0 are the given hyper-parameters. It can be shown that (see Appendix A):

$$p(\Phi_{2m}|\mathbf{X}_{m,2}) \sim IW[(\mathbf{X}_{m,2}\mathbf{X}_{m,2}^T + R_0^{-1}), n_m + \rho_0, q_2], \tag{13}$$

where $IW[\cdot, \cdot, \cdot]$ denotes the inverted Wishart distribution with probability density function proportional to (see Zellner, 1971)

$$|\mathbf{X}_{m,2}\mathbf{X}_{m,2}^T + R_0^{-1}|^{(n_m+\rho_0)/2}|\Phi_{2m}|^{-(n_m+\rho_0+q_2+1)} \exp\left\{-\frac{1}{2}\mathrm{tr}\Phi_{2m}^{-1}(\mathbf{X}_{m,2}\,\mathbf{X}_{m,2}^T + R_0^{-1})\right\}.$$

Similarly as before, the prior distributions of $(\psi_{\delta mk}, \Lambda_{\xi mk})$ are taken as:

$$p(\psi_{\delta mk}^{-1}) \sim \Gamma[\alpha_{0\delta k}, \beta_{0\delta k}],\ p(\Lambda_{\xi mk}|\psi_{\delta mk}) \sim N[\Lambda_{0\xi mk}, \psi_{\delta mk}H_{0\xi mk}], \tag{14}$$

where $k = 1, \ldots, q_1$, $\Lambda_{\xi mk}$ is a $r_{\xi mk} \times 1$ row vector that contains the unknown parameters in the kth row of $\Lambda_{\xi m}$; $\alpha_{0\delta k}$, $\beta_{0\delta k}$, $\Lambda_{0\xi mk}$ and $H_{0\xi mk}$ are the given hyper-parameters. For $h \neq k$, $(\psi_{\delta mk}, \Lambda_{\xi mk})$ and $(\psi_{\delta mh}, \Lambda_{\xi mh})$ are assumed to be independent. To handle the fixed known parameters, let $C_{\xi m} = (c_{\xi mkj})$ be the index matrix associated with $\Lambda_{\xi m}$ and similarly defined as before, $\mathbf{X}^*_{mk}$ be the submatrix of $\mathbf{X}_m$ such that all the jth row corresponding to $c_{\xi mkj} = 0$ deleted; and $\mathbf{X}_{1mk} = (\xi^{i_1*}_{1,mk}, \ldots, \xi^{i_{n_m}*}_{1,mk})$ with

$$\xi^{i_s*}_{1,mk} = \xi^{i_s}_{1,mk} - \sum_{j=1}^{q} \lambda_{\xi mkj} \xi^{i_s j}_{m} (1 - c_{\xi mkj}).$$

Then, it can be shown that for $k = 1, \ldots, q_1$,

$$p(\Lambda_{\xi mk} | \mathbf{X}_m, \psi_{\delta mk}) \sim N[\nu_{\delta mk}, \psi_{\delta mk} \Omega_{\delta mk}]$$

and

$$p(\gamma_{\delta mk} | \mathbf{X}_m, \beta) \sim \Gamma[n_m/2 + \alpha_{0\delta k}, \beta_{\delta mk}], \tag{15}$$

where $\gamma_{\delta mk} = \psi^{-1}_{\delta mk}$, $\Omega_{\delta mk} = (H^{-1}_{0\xi mk} + \mathbf{X}^*_{mk}\mathbf{X}^{*T}_{mk})^{-1}$, $\nu_{\delta mk} = \Omega_{\delta mk}[H^{-1}_{0\xi mk}\Lambda_{0\xi mk} + \mathbf{X}^*_{mk}\mathbf{X}^*_{1mk}]$, and

$$\beta_{\delta mk} = \beta_{0\delta k} + 2^{-1}(\mathbf{X}^{*T}_{1mk}\mathbf{X}^*_{1mk} - \nu^T_{\delta mk}\Omega^{-1}_{\delta mk}\nu_{\delta mk} + \Lambda^T_{0\xi mk}H^{-1}_{0\xi mk}\Lambda_{0\xi mk}).$$

Finally, it is noted that the conditional distributions given in (9) through (15) are familiar and simple distributions. The computational burden required in simulating observations from them is light, and the algorithm is rather efficient. Some derivations of these conditional distributions are presented in the Appendix A.

3.2. *Remarks*

i. Labeling the components is an important issue in the estimation of finite mixtures. Because our whole model is invariant with respect to permutation of the labels $m = 1, \ldots, M$; for identifiability, it is important to adopt an unique labelling. Without loss of generality, we assume that the $\mu_{11} < \cdots < \mu_{M1}$, where μ_{m1} is the first element of the mean vector μ_m. Thus the joint prior distribution of μ is subject to the above inequality constraints. According to the procedure given in Richardson and Green (1997), to preserve this constraint in the Gibbs sampler $\{\mu_{m1}, m = 1, \ldots, M\}$, the full conditional is used only to generate a proposed observation. The proposed observation is accepted as a simulated observation only if the constraint is satisfied.

ii. It has been pointed out (see Diebolt & Robert, 1994; Roeder & Wasserman, 1997) that using fully noninformative prior distributions may lead to improper posterior distributions in a mixture context. So, most of the existing Bayesian analysis on mixtures of the normal distribution used the conjugate type prior distributions (see, e.g., Roeder & Wasserman, 1997). This is one of the motivations for our selection of the conjugate prior distributions in the present analysis.

iii. For the mixture model, it may be desirable to impose some simple constraints on the parameters in θ. For example, it may be interesting to see whether Λ_m, $\Lambda_{\xi m}$, Φ_{2m}, Ψ_m and $\Psi_{\delta m}$ are invariant across all components in the model. Our methodology developed here can be extended to handle simple constraints by some minor modifications in defining the prior distributions and in deriving the conditional distributions.

iv. For brevity, we do not introduce model to the mean vector. Extension of the proposed methodology to models with mean structures requires to incorporate in the Gibbs sampler an additional component that is induced by the corresponding additional parameters in the

mean structure. This can be done without much difficulty. Also, for brevity, the distribution of ξ_{m2} is taken to be $N[0, \Phi_{2m}]$. However, this assumption can be relaxed with the following minor modification of the proposed Gibbs sampler algorithm. Let ξ_{m2} be partitioned into $(\xi_{m2}^{(1)T}, \xi_{m2}^{(2)T})^T$, where $\xi_{m2}^{(1)}$ consists explanatory variables such as professional affiliation or gender which may be observed directly, and $\xi_{m2}^{(2)}$ is a latent vector with distribution $N[0, \Phi_{2m}^{(2)}]$. Then in the Gibbs sampler, $\xi_{m2}^{(1)}$ will be treated as observed data, and only $\xi_{m2}^{(2)}$ is required to be simulated. The estimation of Γ_m and $\Phi_{2m}^{(2)}$ will be the same as described in section 3.1.

4. Statistical Analyses

4.1. Bayesian Estimates

It has been shown (Geman & Geman, 1984; Geyer, 1992) that under mild conditions and for sufficiently large j, say J, the joint distribution of $(\theta^{(J)}, \mathbf{X}^{(J)}, \mathbf{Z}^{(J)})$ converges at an exponential rate to the desired posterior distribution $[\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y}]$. Hence, $[\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y}]$ can be approximated by the empirical distribution of $\{(\theta^{(t)}, \mathbf{X}^{(t)}, \mathbf{Z}^{(t)}) : t = J+1, \ldots, J+T\}$ where T is chosen to give sufficient precision to the empirical distribution. The convergence of the Gibbs sampler can be monitored by the "estimated potential scale reduction (EPSR)" values as suggested by Gelman and Rubin (1992). To obtain a more nearly independent sample, observations may be collected in cycles with indices $t = J + c, J + 2c, \ldots, J + Tc$ for some spacing c (see Gelfand & Smith, 1990). In most practical applications a small c will suffice for many statistical analyses such as getting estimates of the parameters and standard errors, see Zeger and Karim (1991), Albert and Chib (1993).

For brevity, let $\{(\theta^{(t)}, \mathbf{X}^{(t)}, \mathbf{Z}^{(t)}), t = 1, \ldots, T\}$ be the random observations of $(\theta, \mathbf{X}, \mathbf{Z})$ generated by the Gibbs sampler from $[\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y}]$. The joint Bayesian estimates of θ, allocation variables in $\mathbf{Z}$ and the latent factors in $\mathbf{X}$ can be obtained easily via the corresponding sample means of the generated observations as follows:

$$\hat{\theta} = T^{-1}\sum_{t=1}^{T}\theta^{(t)}, \quad \hat{\mathbf{X}} = T^{-1}\sum_{t=1}^{T}\mathbf{X}^{(t)}, \quad \hat{\mathbf{Z}} = T^{-1}\sum_{t=1}^{T}\mathbf{Z}^{(t)}. \tag{16}$$

Clearly, these Bayesian estimates are consistent estimates of the corresponding posterior means, see Geyer (1992). It is rather difficult to derive analytic forms for the covariance matrices $\mathrm{Var}(\theta|\mathbf{Y})$ and $\mathrm{Var}(\xi^i|\mathbf{Y})$. However, their consistent estimates can be obtained as follows:

$$\widehat{\mathrm{Var}(\theta)}|\mathbf{Y}) = (T-1)^{-1}\sum_{t=1}^{T}(\theta^{(t)} - \hat{\theta})(\theta^{(t)} - \hat{\theta})^T, \tag{17}$$

$$\widehat{\mathrm{Var}(\xi^i)}|\mathbf{Y}) = (T-1)^{-1}\sum_{t=1}^{T}(\xi^{i(t)} - \hat{\xi}^i)(\xi^{i(t)} - \hat{\xi}^i)^T, \quad i = 1, \ldots, n.$$

Hence, the standard error estimates can be obtained conveniently by the Gibbs sampler algorithm. Theoretically, since they are coming from the sample covariance matrix of a sufficiently large number of random observations from the appropriate posterior distribution of the parameter, the effect of poor separation of components may not be as serious as in the maximum likelihood estimation. But, we need further theoretical or empirical evidence to draw more definite conclusion. Other statistical inferences on θ or ξ^i, such as deriving the confidence intervals and the statistics for hypothesis testing can be achieved based on the simulated observations as well (see, e.g., Besag, Green Higdon & Mengersen, 1995; and Gilks, Richardson & Spiegelhalter, 1996). In small samples, the posterior distributions, especially for π_m, may be nonsymmetric. Under these situations, it may not be possible to construct confidence intervals.

4.2. *Posterior Predictive p-Value*

Assessing the plausibility of a proposed model is always fundamental in data analysis. Based on the idea of posterior predictive assessment given in Rubin (1984) and Meng (1994), Gelman, Meng and Stein (1996) proposed an approach for model diagnosis in a Bayesian framework. It has been shown that (see Gelman et al., 1996, and the references therein) this approach is computationally and conceptually simple, and is very useful for a wide variates of complicated situations. Further, the required computation is a byproduct of the common Bayesian simulation procedure such as the Gibbs sampler. Hence, this procedure is applied here to establish a goodness-of-fit assessment for the posited model under the null hypothesis H_0 that the true population model is a M mixtures of the LISREL model as defined in (2) and (3). The *posterior predictive p-value* for our situation is defined as

$$p_B = \Pr\left\{D(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z}) \geq D(\mathbf{Y}|\theta, \mathbf{X}, \mathbf{Z})|\mathbf{Y}, H_0\right\},$$

$$= \int I\left\{D(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z}) \geq D(\mathbf{Y}|\theta, \mathbf{X}, \mathbf{Z})\right\} p(\mathbf{Y}^{\text{rep}}, \theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y}, H_0) d\mathbf{Y}^{\text{rep}} d\theta d\mathbf{Z} d\mathbf{X},$$

where $I(\cdot)$ is an indicator function, $\mathbf{Y}^{\text{rep}}$ denotes a replication of $\mathbf{Y}$ and $D(\cdot|\cdot)$ is a discrepancy variable. The probability is taken over the following joint posterior distribution of $(\mathbf{Y}^{\text{rep}}, \theta, \mathbf{X}, \mathbf{Z})$ given H_0 and $\mathbf{Y}$:

$$p(\mathbf{Y}^{\text{rep}}, \theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y}, H_0) = p(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z}) p(\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y}).$$

The discrepancy variable is taken as

$$D(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z}) = \sum_{m=1}^{M} \sum_{i:z_i=m} (y_i^{\text{rep}} - \mu_m - \Lambda_m \xi_i)^T \Psi_m^{-1} (y_i^{\text{rep}} - \mu_m - \Lambda_m \xi_i),$$

where the distribution of $D(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z})$ is chi-square with pn degrees of freedom. The posterior predictive p-value based on this discrepancy variable is given by

$$p_B(\mathbf{Y}) = \int \Pr\left\{\chi^2(pn) \geq D(\mathbf{Y}|\theta, \mathbf{X}, \mathbf{Z})\right\} p(\theta, \mathbf{X}, \mathbf{Z}|\mathbf{Y})\, d\theta d\mathbf{Z} d\mathbf{X}.$$

By the Rao-Blackwell theorem (see Mood, Graybill & Boes, 1974)

$$\begin{aligned}\text{Var } I\{D(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z}) \geq D(\mathbf{Y}|\theta, \mathbf{X}, \mathbf{Z})\} \\ &\geq \text{Var } E[I\{D(\mathbf{Y}^{\text{rep}}|\theta, \mathbf{X}, \mathbf{Z}) \geq D(\mathbf{Y}|\theta, \mathbf{X}, \mathbf{Z})\}|\theta, \mathbf{X}, \mathbf{Z}] \\ &= \text{Var } P_r\{\chi^2(pn) \geq D(\mathbf{Y}|\theta, \mathbf{X}, \mathbf{Z})\}.\end{aligned}$$

Thus, to obtain an estimate with smaller variation, we propose to use the following *Rao-Blackwellized* type estimate of $p_B(\mathbf{Y})$:

$$\hat{p}_B(\mathbf{Y}) = T^{-1} \sum_{t=1}^{T} \Pr\left(\chi^2(pn) \geq D(\mathbf{Y}|\theta^{(t)}, \mathbf{X}^{(t)}, \mathbf{Z}^{(t)})\right). \tag{18}$$

The computation of $\hat{p}_B(\mathbf{Y})$ is straightforward, since $D(\mathbf{Y}|\theta^{(t)}, \mathbf{X}^{(t)}, \mathbf{Z}^{(t)})$ can be calculated in each iteration of the Gibbs sampler and the tail-area probability of χ^2 distribution can be obtained in any standard statistical software. A too small (or too big) $p_B(\mathbf{Y})$, say less than 0.05 (or larger than 0.95), indicates the inadequacy of the posited model. See Gelman et al. (1996), and Meng (1994) for more detailed discussions about the theoretical and practical aspects of the posterior predictive p-value.

4.3. Bayesian Classification

Apart from their role in facilitating computation, the allocation variables in $\mathbf{Z}$ also form a coherent basis for classification of the observations. Classification can either be addressed on a within-sample basis or a predictive basis. Using the 'percentage correctly classified' loss function (see Richardson & Green, 1997), the Bayes classification of an existing observation y_i and a new observation y^* are respectively given by

$$\hat{z}_i = \text{argmax}_m\{\Pr(z_i = m|\mathbf{Y})\} \quad \text{and} \quad \hat{z}^* = \text{argmax}_m\{\Pr(z^* = m|\mathbf{Y}, y^*)\}.$$

The posterior probabilities $\{\Pr(z_i = m|y_i); m = 1, \ldots, M\}$ can be directly estimated via the sample mean of the observations generated by the Gibbs sampler:

$$\Pr(z_i = m|\mathbf{Y}) \approx T^{-1} \sum_{t=1}^{T} I(z_i^{(t)} = m).$$

Predictive classification addresses the question of classifying a new observation y^*. Let the corresponding allocation variable be z^*, the Bayes classification requires to compute $\Pr(z^* = m|\mathbf{Y}, y^*)$. Inclusion of the additional datum theoretically changes the posterior distributions, and it seems that the simulation process should be rerun for each new y^*. This is obviously impractical. So, we employ the following approximation:

$$\Pr(z^* = m|\mathbf{Y}, y^*) = \int p(z^* = m|\theta, y^*)p(\theta|\mathbf{Y}, y^*)\,d\theta \approx \int p(z^* = m|\theta, y^*)p(\theta|\mathbf{Y})\,d\theta;$$

and estimate the last integral by the following sample average of the generated observations from the Gibbs sampler procedure:

$$\Pr(z^* = m|\mathbf{Y}, y^*) \approx T^{-1} \sum_{t=1}^{T} \left[\pi_m^{(t)} f_m\left(y^*|\mu_m^{(t)}, \beta_m^{(t)}\right) \Big/ \sum_{j=1}^{M} \pi_j f_j\left(y^*|\mu_j^{(t)}, \beta_j^{(t)}\right) \right].$$

4.4. Residual and Outlier Statistics

The problem of residual and outlier analyses is thoroughly studied and reviewed by Barnett and Lewis (1984), Hawkins (1980), and Cook and Weisberg (1982), among others. Chaloner and Brant (1988) developed some Bayesian outlier statistics for the linear model based on the posterior distribution of the unobserved error terms. The residual estimates are the posterior means of the unknown random errors. For the normal linear model $y_i = x_i^T\beta + \epsilon_i$, where y_i is the observed data, x_i is the covariate, and ϵ_i is the residual with distribution $N[0, \sigma^2]$; the posterior probability $\text{POS}(i; K) = \Pr(|\epsilon_i|/\sigma > K|\mathbf{Y})$ is called the posterior outlier statistic for a fixed constant K. This constant may be chosen to be a familiar number such as 2.0, 3.0 or $\Phi^{-1}(0.5 + 0.5(0.95)^{1/n})$, where $\Phi(\cdot)$ is the standard normal distribution function. It provides a formal method for identification of outliers in normal linear model.

In structural equation modeling, residuals and outliers are usually discussed in the sense that an estimated covariance term differs significantly from the empirical covariance term. As a complementary method, we now extend the approaches of Chaloner and Brant (1988), Albert and Chib (1995), Chaloner (1991), and Weiss (1994) to the finite mixtures in the LISREL model. The essential idea is to define the residual as an unobserved random error and a particular observation whose residual is far from the corresponding expected value may be regarded as an outlier. The posterior probability can be directly estimated by using the observations $\{(\theta^{(t)}, \mathbf{X}^{(t)}, \mathbf{Z}^{(t)}),\ t = 1, \ldots, T\}$ generated by the Gibbs sampler. The ith observation is regarded as the *outlier* if the corresponding posterior probability is large. Informal identification of certain cases as outliers is usually suggested by the plots of residuals (see Weiss & Lazaro, 1992). More details are given as below.

Let $\epsilon_{ij}^* = \psi_{mj}^{-1/2}(y_{ij} - \mu_{mj} - \Lambda_{mj}\xi_i)$, it can be seen from (2) and (3) that ϵ_{ij}^* can be treated as the *standardized residual* when conditional on $z_i = m$. An informal identification of outliers can be obtained by plotting the sample mean $\overline{\epsilon}_{ij}^*$ of ϵ_{ij}^* against i. The corresponding posterior outlier statistic is given by $\mathrm{POS}_\epsilon(i, j; K) = \Pr(|\epsilon_{ij}^*| > K|\mathbf{Y})$. The method described above for detecting univariate outliers requires the computation of $n(p+q)$ posterior outlier statistics in a data set. Computationally, this can be a moderate burden. An alternative way is to introduce the multivariate outlier statistics. For example, the following statistic

$$\mathrm{POS}_\epsilon(i; K_1) = \Pr(\epsilon_i^{*T}\epsilon_i^* > K_1|Y), \tag{19}$$

where $\epsilon_i^* = (\epsilon_{i1}^*, \ldots, \epsilon_{ip}^*)^T$ and K_1 is some fixed constant, can be used. Conditional on the parameters, z_i and ξ^i; $\epsilon_i^{*T}\epsilon_i^*$ is distributed as a $\chi^2(p)$ distribution. So, it makes sense to choose K_1 to be $\chi^2(p; 1-q)$ for $q = 0.01$ or 0.05, where $\chi^2(p, 1-q)$ is the $100(1-q)$ percentile of $\chi^2(p)$. Another useful approach is to consider a single value summary measure of $\mathrm{POS}_\epsilon(i, j; K)$; for example, we can introduce the following statistics:

$$\mathrm{POS}_\epsilon(i, +; K) = \sum_{j=1}^{p} \mathrm{POS}_\epsilon(i, j; K), \quad \mathrm{POS}_\epsilon(i, \max; K) = \max_{1\le j\le p} \mathrm{POS}_\epsilon(i, j; K).$$

Again, observations with large POS_ϵ values can be identified as outliers. Finally, we can investigate the QQ plot of $\overline{\epsilon}_{ij}^*$ to check the assumption of the normality; and apply the above analyses with ξ_2^* and δ_i.

5. A Simulation Study and An Example

5.1. Simulation Study

Results of a simulation study will be presented to give some ideas on the performance of the proposed Bayesian approach. The data set is generated from a mixture of two LISREL models defined in (2) and (3). Each model involves six manifest variables which are related with three latent factors $\xi_{m1} = (\xi^{(1)}, \xi^{(2)})$, and $\xi_{m2} = \xi^{(3)}$ for $m = 1, 2$. The population values of the elements in Λ_1, Λ_2, Π_1 and Π_2 are taken as:

$$\Lambda_1^T = \Lambda_2^T = \begin{bmatrix} 1.0 & 0.8^* & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.8^* & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.8^* \end{bmatrix}, \quad \Pi_1 = \begin{bmatrix} 0 & 0 \\ 0.5^* & 0 \end{bmatrix}, \quad \Pi_2 = \begin{bmatrix} 0 & 0 \\ -0.5^* & 0 \end{bmatrix}.$$

In the estimation, the one's and zero's are treated as fixed known parameters, while the parameters with asterisks are treated as unknown parameters. The true population values of the other unknown parameters are given by: $\Phi_1 = \Phi_2 = (1.0)$, $\mu_1 = 0.0 \times J_6$, $\mu_2 = 2.0 \times J_6$, $\psi_{1i} = \psi_{2i} = 0.8$ for all $i = 1, \ldots, 6$, $\psi_{\delta 1k} = \psi_{\delta 2k} = 1.0$ for all $k = 1, 2$, $\Gamma_1 = [0.6, 0.6]^T$, $\Gamma_2 = [0.6, -0.6]^T$, where J_6 is a 6×1 vector with all elements equal to 1. The following two designs with various mixing proportions are considered: $\{\pi_1 = 0.5, \pi_2 = 0.5\}$; and $\{\pi_1 = 0.3, \pi_2 = 0.7\}$. For each design, we have a 2-mixtures of LISREL model with 40 unknown parameters. Sample sizes $n = 400$ and 800 were selected, and 100 replications were completed for each combination.

Two Bayesian estimates with the following prior distributions were obtained via our approach: (I) Estimates based on conjugate priors with hyper-parameters $\{\Lambda_{0mk}, \Lambda_{0\xi mk}\}$ fixed at the true values; $\alpha = 1$, $\mu_0 = \overline{y}$, $\Sigma_0 = S_y/2$, $\rho_0 = 5$, $R_0^{-1} = [5.0]$, $\alpha_{0k} = \alpha_{0\delta k} = 10$, $\beta_{0k} = \beta_{0\delta k} = 8$ for all k; $H_{0mk} = I$ and $H_{0\xi mk} = I$ for all m and k, where $\overline{y}$ and S_y are the sample mean and the sample covariance matrix of the simulated data. This can be regarded as a situation with good prior information. (II) Estimates based on conjugate priors with hyper-parameters $\{\Lambda_{0mk}, \Lambda_{0\xi mk}\}$ equal to 2.0 times the true values; other hyperparameters are fixed at the same values as in (I).

The starting values of the unknown parameters are given by: $\pi_1 = \pi_2 = 0.5$, $\Phi_1 = \Phi_2 = 2.0$, $\mu_1 = 0.0 \times J_6$, $\mu_2 = 3.0 \times J_6$, $\psi_{1i} = \psi_{2i} = 1.2$ for $i = 1, \ldots, 6$, $\psi_{\delta 1k} = \psi_{\delta 2k} = 1.2$ for $k = 1, 2$, $\Gamma_1 = [1.0, 1.0]^T$, $\Gamma_2 = [1.0, -1.0]^T$, and $\Lambda_{1ij} = \Lambda_{2ij} = 0.0$ for all the unknown parameters in Λ_1 and Λ_2. We first conducted a few test runs as a pilot study to obtain some idea about the number of the Gibbs sampler iterations in getting convergence. We found that in all these runs, the Gibbs sampler converged in about 500–1000 iterations, where the ESPR values (Gelman & Rubin, 1992) are less than 1.2. So, for the 100 replications in the simulation, random observations were collected after $J = 1000$ iterations. Then, a total of an additional 3000 observations were collected to produce the Bayesian estimates and their standard error estimates via (16) and (17). On the basis of the 100 replications, the mean and the standard derivations (SD) of the estimates, as well as the mean of the standard errors estimates (SE) were computed. Moreover, the bias which is the difference of the true parameter and the mean of the corresponding estimates, and the root mean squares (RMS) between the estimates and the true values based on the 100 replications were computed. The results for $\pi_1 = 0.5$ are reported in Tables 1 and 2; to save space, the similar results with $\pi_1 = 0.3$ are not reported. Here, the sums of the RMS across the estimates are presented in the last row. We have the following findings from these tables: (i) As expected, Bayesian estimates with more accurate priors are better, but the differences are not significant. Hence, it seems that the requirement of accurate hyperparameters values is not crucial in the Bayesian analysis. This agrees with the results in the Bayesian literature, see, for example, Lindley and Smith (1972). (ii) Both Bayesian estimates are reasonably accurate. (iii) As expected, increasing the sample size improves the accuracy of the estimates and reduces the differences between the two types of Bayesian estimates. (iv) In most cases, the SE values are slightly smaller than the SD values. As pointed out by Dolan and van der Maas (1998), this minor difference may due to the difference in the model that featured in the simulation study. However, the SE and SD values are quite close to each other, this indicates that the standard error estimates produced by the proposed procedure are reasonable.

All computations are performed using a Sun Enterprise 4000 server. For $N = 400$, the average computing time over the four designs with different π and prior distributions is about 292 minutes for 100 replications; while for $N = 800$, the corresponding average computer time is about 640 minutes.

5.2. *An Example on "Job" and "Homelife"*

A small portion of the ICPSR data set collected in the project WORLD VALUES SURVEY 1981-1984 AND 1990-1993 (World Value Study Group, ICPSR Version) is analyzed in this example. The whole data set was collected in 45 societies around the world on broad topics such as work, the meaning and purpose of life, family life and contemporary social issues. As an illustration of our proposed method, only the data obtained from the United Kingdom with sample size 1484 were used. Eight variables in the original data set (variables 116, 117, 180, 132, 96, 255, 254 and 252) that related with respondents' job and homelife were taken as manifest variables in $y = (y^1, \ldots, y^8)^T$, see Appendix B. These variables were measured in a 10 points scale and hence were treated as continuous in this illustration. For brevity, observations with missing data were deleted and the remaining sample size is 824. The data set was first analyzed with a regular LISREL model with three latent variables which can be roughly interpreted as "job satisfaction, $\xi_1(1)$", "homelife, $\xi_2(1)$" and "job attitude, $\xi_2(2)$". The specification of the parameter matrices in the LISREL formulation are given by: $\Pi = 0$, $\Psi_\delta = \psi_\delta$, $\Gamma = (\tau_{11}, \tau_{21})$,

$$\Lambda^T = \begin{bmatrix} \lambda_{11} & \lambda_{21} & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & \lambda_{32} & \lambda_{42} & \lambda_{52} & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \lambda_{63} & \lambda_{73} & \lambda_{83} \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

and $\Psi_\epsilon = \text{diag}(\psi_{\epsilon 11}, \ldots, \psi_{\epsilon 88})$. To identify the model, λ_{11}, λ_{32}, and λ_{63} were restricted at 1.0, and all the zero elements in Λ and the off diagonal elements of Ψ_ϵ were also fixed. From the

TABLE 1.
Summary statistics for Bayesian estimates (I) with $\pi_1 = 0.5$

	$N = 400$				$N = 800$			
Para.	Bias	SE	SD	RMS	Bias	SE	SD	RMS
$\pi_1 = 0.5$	0.01	0.05	0.04	0.05	0.01	0.03	0.03	0.03
$\pi_2 = 0.5$	0.01	0.05	0.04	0.05	0.01	0.03	0.03	0.03
$\Phi_1 = 1.0$	0.09	0.24	0.25	0.26	−0.07	0.17	0.19	0.18
$\Phi_2 = 1.0$	0.08	0.21	0.21	0.23	−0.06	0.15	0.16	0.16
$\mu_1(1) = 0.0$	0.03	0.16	0.15	0.16	0.01	0.11	0.11	0.11
$\mu_1(2) = 0.0$	0.05	0.14	0.13	0.15	0.01	0.10	0.10	0.10
$\mu_1(3) = 0.0$	0.05	0.19	0.18	0.20	0.01	0.14	0.13	0.14
$\mu_1(4) = 0.0$	0.06	0.16	0.16	0.18	0.01	0.13	0.12	0.13
$\mu_1(5) = 0.0$	0.02	0.14	0.13	0.14	−0.00	0.11	0.10	0.11
$\mu_1(6) = 0.0$	0.04	0.13	0.13	0.14	0.00	0.09	0.09	0.09
$\mu_2(1) = 2.0$	0.02	0.13	0.12	0.13	0.00	0.08	0.09	0.08
$\mu_2(2) = 2.0$	0.01	0.10	0.11	0.10	0.00	0.09	0.08	0.08
$\mu_2(3) = 2.0$	−0.02	0.14	0.14	0.14	0.00	0.09	0.09	0.09
$\mu_2(4) = 2.0$	−0.01	0.12	0.12	0.12	0.00	0.08	0.08	0.08
$\mu_2(5) = 2.0$	0.03	0.11	0.12	0.12	0.00	0.08	0.08	0.08
$\mu_2(6) = 2.0$	0.02	0.10	0.11	0.10	0.00	0.08	0.08	0.08
$\psi_{\delta 1}(1) = 1.0$	−0.09	0.13	0.18	0.16	−0.09	0.12	0.14	0.16
$\psi_{\delta 1}(2) = 1.0$	−0.09	0.13	0.19	0.16	−0.08	0.15	0.15	0.16
$\psi_{\delta 2}(1) = 1.0$	−0.11	0.14	0.18	0.18	−0.08	0.12	0.14	0.15
$\psi_{\delta 2}(2) = 1.0$	−0.12	0.15	0.19	0.19	−0.07	0.14	0.16	0.15
$\Psi_1(1) = 0.8$	0.06	0.12	0.15	0.14	0.06	0.10	0.12	0.12
$\Psi_1(2) = 0.8$	−0.03	0.09	0.12	0.10	−0.02	0.08	0.10	0.09
$\Psi_1(3) = 0.8$	0.06	0.13	0.16	0.14	0.03	0.09	0.13	0.09
$\Psi_1(4) = 0.8$	−0.01	0.10	0.13	0.10	−0.02	0.08	0.10	0.08
$\Psi_1(5) = 0.8$	0.09	0.11	0.15	0.14	0.06	0.11	0.12	0.13
$\Psi_1(6) = 0.8$	−0.03	0.09	0.12	0.09	−0.01	0.09	0.09	0.08
$\Psi_2(1) = 0.8$	0.05	0.12	0.15	0.13	0.05	0.11	0.12	0.13
$\Psi_2(2) = 0.8$	0.00	0.10	0.12	0.10	0.00	0.09	0.10	0.08
$\Psi_2(3) = 0.8$	0.03	0.12	0.16	0.13	0.04	0.12	0.13	0.11
$\Psi_2(4) = 0.8$	0.01	0.09	0.13	0.09	0.00	0.08	0.11	0.08
$\Psi_2(5) = 0.8$	0.06	0.12	0.15	0.14	0.04	0.10	0.11	0.10
$\Psi_2(6) = 0.8$	0.00	0.10	0.12	0.10	−0.02	0.08	0.09	0.08
$\Lambda_1(2, 1) = 0.8$	0.07	0.09	0.12	0.11	0.06	0.08	0.09	0.10
$\Lambda_1(4, 2) = 0.8$	0.04	0.08	0.10	0.08	0.02	0.06	0.07	0.07
$\Lambda_1(6, 3) = 0.8$	0.11	0.15	0.17	0.18	0.08	0.13	0.12	0.15
$\Lambda_2(2, 1) = 0.8$	0.03	0.09	0.11	0.09	0.04	0.08	0.08	0.09
$\Lambda_2(4, 2) = 0.8$	0.01	0.07	0.08	0.07	0.02	0.05	0.06	0.06
$\Lambda_2(6, 3) = 0.8$	0.08	0.12	0.14	0.14	0.07	0.09	0.10	0.12
$\Pi_1(2, 1) = 0.5$	0.01	0.16	0.16	0.16	0.04	0.12	0.11	0.12
$\Gamma_1(1, 1) = 0.6$	0.06	0.19	0.18	0.20	0.03	0.12	0.12	0.12
$\Gamma_1(2, 1) = 0.6$	0.06	0.24	0.23	0.24	0.04	0.15	0.15	0.16
$\Pi_2(2, 1) = -0.5$	−0.02	0.17	0.18	0.17	−0.02	0.12	0.11	0.12
$\Gamma_2(1, 1) = 0.6$	0.10	0.15	0.16	0.18	0.03	0.11	0.10	0.11
$\Gamma_2(2, 1) = -0.6$	−0.11	0.21	0.24	0.24	−0.04	0.14	0.15	0.15
$\sum$ RMS				6.22				4.73

TABLE 2.
Summary statistics for Bayesian estimates (II) with $\pi_1 = 0.5$

	$N = 400$				$N = 800$			
Para.	Bias	SE	SD	RMS	Bias	SE	SD	RMS
$\pi_1 = 0.3$	0.01	0.05	0.04	0.05	0.01	0.04	0.03	0.04
$\pi_2 = 0.7$	−0.01	0.05	0.04	0.05	−0.01	0.04	0.03	0.04
$\Phi_1 = 1.0$	−0.14	0.23	0.25	0.27	−0.08	0.20	0.19	0.21
$\Phi_2 = 1.0$	−0.18	0.18	0.20	0.26	−0.12	0.16	0.15	0.20
$\mu_1(1) = 0.0$	0.04	0.13	0.15	0.14	0.01	0.13	0.11	0.13
$\mu_1(2) = 0.0$	0.05	0.14	0.14	0.15	0.02	0.12	0.10	0.12
$\mu_1(3) = 0.0$	0.06	0.20	0.19	0.21	0.03	0.15	0.13	0.15
$\mu_1(4) = 0.0$	0.06	0.18	0.17	0.19	0.04	0.13	0.12	0.13
$\mu_1(5) = 0.0$	−0.01	0.14	0.14	0.14	0.01	0.10	0.10	0.10
$\mu_1(6) = 0.0$	0.03	0.14	0.13	0.14	0.02	0.10	0.09	0.10
$\mu_2(1) = 2.0$	0.00	0.13	0.13	0.13	0.02	0.09	0.09	0.09
$\mu_2(2) = 2.0$	−0.00	0.11	0.12	0.11	0.01	0.08	0.08	0.08
$\mu_2(3) = 2.0$	−0.03	0.12	0.14	0.13	−0.02	0.10	0.09	0.10
$\mu_2(4) = 2.0$	−0.03	0.11	0.12	0.12	−0.02	0.07	0.08	0.07
$\mu_2(5) = 2.0$	0.05	0.13	0.12	0.14	0.02	0.08	0.08	0.08
$\mu_2(6) = 2.0$	0.01	0.10	0.11	0.10	0.03	0.08	0.08	0.09
$\psi_{\delta 1}(1) = 1.0$	−0.11	0.16	0.18	0.20	−0.07	0.12	0.15	0.14
$\psi_{\delta 1}(2) = 1.0$	−0.12	0.15	0.18	0.19	−0.06	0.13	0.15	0.14
$\psi_{\delta 2}(1) = 1.0$	−0.11	0.13	0.18	0.17	−0.09	0.11	0.14	0.14
$\psi_{\delta 2}(2) = 1.0$	−0.09	0.14	0.20	0.17	−0.09	0.14	0.16	0.17
$\Psi_1(1) = 0.8$	0.08	0.12	0.15	0.15	0.06	0.10	0.13	0.12
$\Psi_1(2) = 0.8$	−0.02	0.10	0.13	0.10	0.00	0.08	0.10	0.08
$\Psi_1(3) = 0.8$	0.08	0.13	0.16	0.15	0.04	0.10	0.13	0.11
$\Psi_1(4) = 0.8$	−0.03	0.10	0.13	0.10	−0.01	0.08	0.10	0.08
$\Psi_1(5) = 0.8$	0.11	0.12	0.15	0.17	0.08	0.12	0.12	0.14
$\Psi_1(6) = 0.8$	−0.04	0.10	0.12	0.11	−0.02	0.08	0.09	0.08
$\Psi_2(1) = 0.8$	0.06	0.11	0.15	0.12	0.04	0.11	0.12	0.12
$\Psi_2(2) = 0.8$	−0.03	0.09	0.13	0.10	−0.01	0.08	0.10	0.08
$\Psi_2(3) = 0.8$	0.02	0.11	0.16	0.12	0.03	0.11	0.13	0.11
$\Psi_2(4) = 0.8$	−0.00	0.09	0.13	0.09	−0.01	0.09	0.10	0.09
$\Psi_2(5) = 0.8$	0.09	0.11	0.15	0.14	0.08	0.11	0.11	0.13
$\Psi_2(6) = 0.8$	−0.02	0.11	0.12	0.12	−0.01	0.09	0.09	0.09
$\Lambda_1(2, 1) = 0.8$	0.09	0.11	0.13	0.14	0.04	0.09	0.10	0.10
$\Lambda_1(4, 2) = 0.8$	0.03	0.08	0.10	0.09	0.03	0.06	0.07	0.07
$\Lambda_1(6, 3) = 0.8$	0.16	0.19	0.18	0.24	0.09	0.11	0.13	0.14
$\Lambda_2(2, 1) = 0.8$	0.07	0.10	0.11	0.13	0.04	0.07	0.08	0.08
$\Lambda_2(4, 2) = 0.8$	0.03	0.08	0.09	0.08	0.02	0.05	0.06	0.05
$\Lambda_2(6, 3) = 0.8$	0.13	0.14	0.16	0.20	0.09	0.10	0.11	0.14
$\Pi_1(2, 1) = 0.5$	0.02	0.15	0.17	0.15	0.02	0.11	0.12	0.11
$\Gamma_1(1, 1) = 0.6$	0.10	0.17	0.19	0.20	0.06	0.12	0.13	0.13
$\Gamma_1(2, 1) = 0.6$	0.15	0.23	0.24	0.27	0.06	0.15	0.16	0.16
$\Pi_2(2, 1) = -0.5$	0.02	0.17	0.18	0.17	0.01	0.12	0.12	0.12
$\Gamma_2(1, 1) = 0.6$	0.12	0.17	0.18	0.21	0.06	0.13	0.12	0.14
$\Gamma_2(2, 1) = -0.6$	−0.14	0.20	0.26	0.25	−0.12	0.16	0.17	0.20
$\sum$ RMS				6.66				4.99

maximum likelihood solution obtained via the regular LISREL VIII (Jöreskog & Sörbom, 1996) program, the chi-squared goodness of fit statistic was equal to 34.46 with 16 degrees of freedom. The corresponding p-value is about 0.008. Hence, the model fit statistic is significant at 0.05 level. However, this conclusion may be misleading since there might be more than one single component involved.

This data set was reanalyzed via a 2-mixtures of the LISREL model, where the model for each component is defined as above. The total number of unknown parameters is 62. Bayesian estimates of the structural parameters and direct estimates of the factor scores were obtained via the Gibbs sampler. The following hyper-parameters were selected: $\alpha = 1$, $\mu_0 = \overline{y}$, $\Sigma_0 = S_y/2.0$, $\rho_0 = 5$ and $R_0^{-1} = 5I_2$; $\alpha_{0k} = \alpha_{0\delta k} = \beta_{0k} = \beta_{0\delta k} = 6$ for all k; $H_{0mk} = I$ and $H_{0\xi mk} = I$, $\Lambda_{0mk} = \tilde{\Lambda}_{0mk}$, $\Lambda_{0\xi mk} = \tilde{\Lambda}_{0\xi mk}$ for all m and k, where $\tilde{\Lambda}_{0mk}$ and $\tilde{\Lambda}_{0\xi mk}$ are obtained by some initial runs of the Gibbs sampler with the prior hyper-parameters fixed at the LISREL solution as discussed above. Based on different starting values of the parameters, three parallel sequences of observations were generated and the EPSR values were calculated. Figure 1 presents the plots of the EPSR values against the iteration numbers. We observed that the EPSR values of the parameters at the starting points were quite large, this indicates that the starting values are far away from the solution. The Gibbs sampler algorithm converged after about 1000 iterations. After the convergence of the Gibbs sampler, a total of $M = 1000$ observations with $c = 10$ were collected for analysis. The Bayesian estimates of the structural parameters and their standard errors estimates with the following starting values are reported in Table 3: $\pi_1 = \pi_2 = 0.5$, $\Lambda_{1ij} = \Lambda_{2ij} = 0.0$ for all unknown parameters in Λ_1 and Λ_2, $\mu_1 = 0.0 \times J_8$, $\mu_2 = 10.0 \times J_6$, $\psi_{\varepsilon 1i} = \psi_{\varepsilon 2i} = 1.0$ for $i = 1, \ldots, 8$, $\Gamma_1 = \Gamma_2 = (1.0, 1.0)$, $\Phi_1 = \Phi_2 = I_2$ and $\psi_{1\delta} = \psi_{2\delta} = 1.0$. The posterior predictive p-value was equal to 0.41, indicating the proposed model fits the data well. From Table 3, it can be seen that there are clearly two components which have quite different Bayesian estimates for some of the parameters.

Following the suggestion of a reviewer to give a comparison between the Bayesian estimates and the ML estimates, we implemented the EM algorithm proposed in Yung (1997) to reanalyze

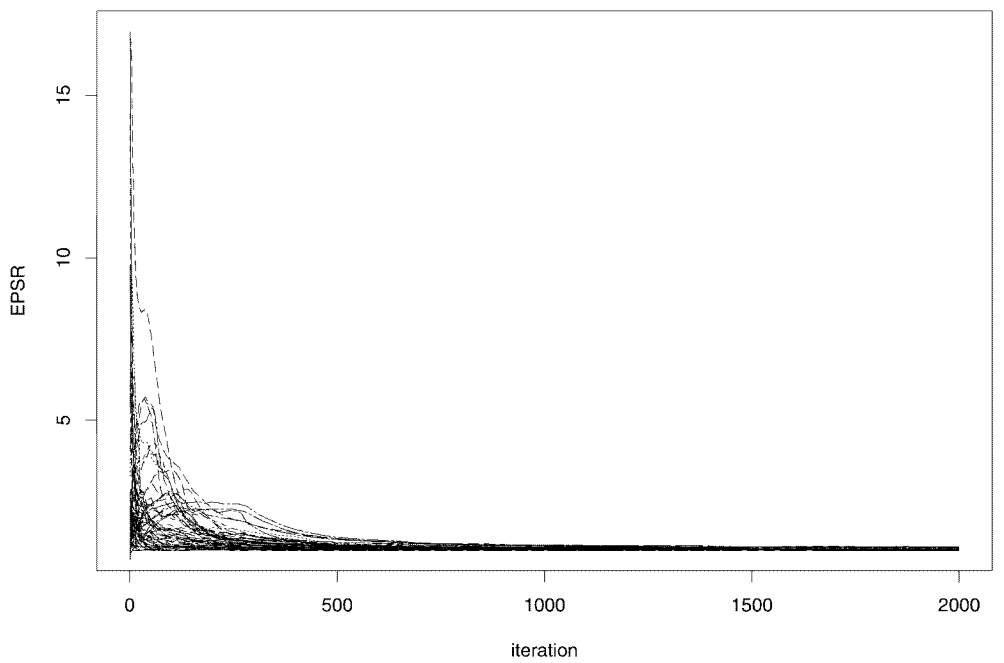

FIGURE 1.
EPSR values of all parameters from three parallel runs in the ICPSR example.

TABLE 3.
Estimates and standard errors estimates of the ICPSR example

	Bayesian estimate				ML estimate			
	Component 1		Component 2		Component 1		Component 2	
Parameter	EST	SE	EST	SE	EST	SE	EST	SE
π	0.56	0.03	0.44	0.03	0.57	0.04	0.43	0.04
μ_1	6.91	0.12	8.09	0.09	6.91	0.19	8.10	0.12
μ_2	6.30	0.15	7.90	0.14	6.32	0.21	7.89	0.15
μ_3	5.87	0.14	7.83	0.11	5.89	0.19	7.81	0.09
μ_4	7.83	0.10	8.70	0.07	7.85	0.15	8.69	0.06
μ_5	7.10	0.11	8.08	0.08	7.11	0.17	8.08	0.06
μ_6	5.42	0.14	4.01	0.16	5.43	0.15	3.98	0.13
μ_7	4.06	0.13	3.61	0.14	4.06	0.14	3.59	0.13
μ_8	5.59	0.14	4.61	0.14	5.59	0.15	4.60	0.14
λ_{11}	1*	–	1*	–	1*	–	1*	–
λ_{21}	0.49	0.11	0.86	0.13	0.38	0.07	0.73	0.19
λ_{32}	1*	–	1*	–	1*	–	1*	–
λ_{42}	1.30	0.17	0.94	0.10	1.21	0.17	0.99	0.12
λ_{52}	1.58	0.20	1.02	0.11	1.53	0.15	1.07	0.15
λ_{63}	1*	–	1*	–	1*	–	1*	–
λ_{73}	2.05	0.44	0.98	0.07	2.02	0.35	0.95	0.09
λ_{83}	1.08	0.27	0.74	0.08	1.05	0.14	0.72	0.08
τ_{11}	0.68	0.14	0.77	0.11	0.65	0.13	0.83	0.11
τ_{12}	−0.02	0.15	−0.09	0.04	−0.02	0.16	−0.10	0.04
ϕ_{11}	1.18	0.26	0.90	0.18	1.25	0.24	0.86	0.21
ϕ_{21}	−0.12	0.08	−0.28	0.15	−0.13	0.08	−0.27	0.15
ϕ_{22}	0.92	0.30	4.30	0.52	0.87	0.31	4.49	0.71
ψ_δ	3.38	0.72	0.70	0.12	4.61	0.58	0.83	0.27
$\psi_{\epsilon 11}$	1.56	0.65	0.56	0.11	0.43	0.24	0.26	0.24
$\psi_{\epsilon 22}$	6.92	0.50	2.80	0.34	7.22	0.53	2.21	0.40
$\psi_{\epsilon 33}$	4.87	0.37	1.35	0.18	4.91	0.36	1.43	0.28
$\psi_{\epsilon 44}$	2.51	0.27	0.45	0.06	2.63	0.29	0.37	0.10
$\psi_{\epsilon 55}$	1.29	0.27	0.55	0.08	1.16	0.44	0.48	0.11
$\psi_{\epsilon 66}$	6.31	0.50	1.25	0.35	6.42	0.51	1.07	0.55
$\psi_{\epsilon 77}$	2.43	0.76	1.07	0.23	2.52	1.26	1.05	0.47
$\psi_{\epsilon 88}$	6.39	0.57	3.15	0.41	6.56	0.61	3.17	0.43

the data set. The standard error estimates of the ML estimates were computed via the "missing information principle" (Louis, 1982), with the expectations evaluated using 10000 simulated observations from the appropriate posterior distributions. For the present model, closed form solution for the M-step can be obtained. The EM algorithm converged in about 599 iterations to the ML solution given in Table 3. Except the estimate of $\psi_{\varepsilon 11}$, the ML estimates are quite close to the Bayesian estimates. The standard errors estimates obtained by these two approaches are also quite similar in most cases.

For the residual and outlier analysis, all the posterior outlier statistics were calculated with $K = 3.0$. A few selected observations with large $\text{POS}_\epsilon(i, j; 3.0)$ values are presented in Table 4. From this table, it can be seen that the observation with $i = 271$ is a possible outlier, because $\text{POS}_\epsilon(i, j; 3.0)$ is larger than 0.50 and its corresponding $\text{POS}_\epsilon(i; K_1)$ is also very large, where K_1 is the 99 percentile of χ^2_8. Further interpretation and discussion are not reported to save space.

TABLE 4.
The posterior outlier statistics for the ICPSR example

Observation number i	$POS_\epsilon(i, j; 3)$ $j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$POS_\epsilon(i; 20.09)$
26	0.07	0	0	0.02	0.01	0	0.04	0	0.66
130	0	0	0	0.42	0.06	0	0	0	0.14
199	0	0.20	0	0	0	0	0	0	0.10
219	0	0	0	0	0	0	0	0.02	0.21
262	0	0	0	0.14	0	0	0	0	0.69
271	0.01	0	0	0.17	0.70	0	0	0	0.94
400	0	0	0	0	0	0	0	0.50	0.14
424	0	0	0	0.03	0.47	0	0	0	0.35
428	0	0	0	0	0.02	0	0	0	0.61
433	0.01	0.01	0	0.04	0.07	0	0	0	0.39
693	0.01	0	0	0	0	0	0	0.22	0.12
724	0.15	0	0	0	0	0	0	0	0.11

6. Discussion

Bayesian analysis of structural equation models has received a lot of attention recently. For example, Scheines, Hoijtink, and Boomsma (1999) discussed a Bayesian statistical inference about a covariance structure via posterior analysis on the basis of the sample covariance matrix. They also used the Gibbs sampler to get the estimates and the posterior predictive p-value to test the goodness-of-fit of the posited model. Since their basic model is not a finite mixtures, the underlying development is less complicated. In addition to Scheines et al. (1999), see Arminger and Muthén (1998), and Zhu and Lee (1999) for more sophisticated Bayesian analysis of nonlinear models, and Shi and Lee (1998) for models with continuous and polytomous variables.

This paper develops a Bayesian approach to analyze a finite mixtures in a LISREL model. It is shown that the algorithm on the basis of the Gibbs sampler is feasible in obtaining the Bayesian solution. In addition to the point estimate of θ, the Gibbs sampler provides useful information to conduct conveniently other statistical inferences; such as the direct estimation of the latent variables, the goodness-of-fit assessment of the posited model, Bayesian classification, residuals and outlier analyses. However, it is not our intention to conclude that the Bayesian approach with the Gibbs sampler is better than the maximum likelihood approach with the EM algorithm or other optimization procedures. To draw such a conclusion, it requires more theoretical and empirical comparisons on various aspects of the approaches which are beyond the scope of this paper. For this complicated model, there remains many open problems for future research. For example, for various algorithms, it is worthwhile to study the robustness to starting values and their performances under poor separation of components. Another important but difficult problem is on analyzing models with an unknown number of components.

Appendix A: Derivation of the Conditional Distributions

From (2) and (4), it is easily seen that

$$p(\xi^i|y_i, z_i = m, \mu, \theta_y) \propto \exp\left\{-\frac{1}{2}\left[\xi^{iT}\left(\Sigma_m^{-1} + \Lambda_m^T\Psi_m^{-1}\Lambda_m\right)\xi^i - 2\xi^{iT}\Lambda_m^T\Psi_m^{-1}(y_i - \mu_m)\right]\right\},$$

which is a normal distribution, $N[\Omega_m^{-1}\Lambda_m^T\Psi_m^{-1}(y_i - \mu_m), \Omega_m^{-1}]$.

Using the likelihood function of $\mathbf{Y}_m$, we obtain

$$p(\mu_m|\cdot) \propto \exp\left\{-\frac{1}{2}\left[(\mu_m - \mu_0)^T\Sigma_0^{-1}(\mu_m - \mu_0)\right.\right.$$

$$+\sum_{z_i=m}(y_i-\mu_m-\Lambda_m\xi^i)^T\Psi_m^{-1}(y_i-\mu_m-\Lambda_m\xi^i)\Bigg]\Bigg\}$$

$$=\exp\left\{-\frac{1}{2}[\mu_m^T(\Sigma_0^{-1}+n_m\Psi_m^{-1})\mu_m-2\mu_m^T(\Sigma_0^{-1}\mu_0+n_m\overline{B}_m)]\right\},$$

which is a normal distribution as described in (12). Moreover, the exponential term in $p(\mathbf{Y}_m|\mu_m,\Lambda_m,\Psi_m,\mathbf{X}_m)$ can be written as

$$\sum_{z_i=m}(y_i-\mu_m-\Lambda_m\xi^i)^T\Psi_m^{-1}(y_i-\mu_m-\Lambda_m\xi^i)$$

$$=\sum_{k=1}^{p}\sum_{z_i=m}\psi_{mk}^{-1}(y_{ki}-\mu_{mk}-\Lambda_{mk}^T\xi^i)^2$$

$$=\sum_{k=1}^{p}\psi_{mk}^{-1}\sum_{z_i=m}(y_{ki}^*-\Lambda_{ymk}^T\xi_k^{i*})^2$$

$$=\sum_{k=1}^{p}\psi_{mk}^{-1}\left[\sum_{z_i=m}y_{ki}^{*2}-2\Lambda_{ymk}^T\sum_{z_i=m}y_{ki}^*\xi_k^{i*}+\mathrm{tr}\left(\Lambda_{ymk}\Lambda_{ymk}^T\sum_{z_i=m}\xi_k^{i*}\xi_k^{i*T}\right)\right]$$

$$=\sum_{k=1}^{p}\psi_{mk}^{-1}\left[\mathbf{Y}_{mk}\mathbf{Y}_{mk}^T-2\Lambda_{ymk}^T\mathbf{X}_{mk}\mathbf{Y}_{mk}+\Lambda_{ymk}^T\mathbf{X}_{mk}\mathbf{X}_{mk}^T\Lambda_{ymk}\right].$$

Therefore, it follows that

$$p(\Lambda_{ym},\Psi_m^{-1}|\mathbf{Y}_m,\mathbf{X}_m,\mu_m)=\prod_{k=1}^{p}p(\Lambda_{ymk},\Psi_{mk}^{-1}|\mathbf{Y}_m,\mathbf{X}_m,\mu_m)$$

$$\propto\prod_{k=1}^{p}\left[\gamma_{mk}^{n_m/2+\alpha_{0k}-1}\exp\{-\beta_k\gamma_{mk}\}\right]\cdot\left[\gamma_{mk}^{r_{mk}/2}\exp\left\{-\frac{1}{2}\gamma_{mk}(\Lambda_{ymk}-\nu_{mk})^T\Omega_{mk}^{-1}(\Lambda_{ymk}-\nu_{mk})\right\}\right].$$

Thus, the full distributions for $(\Lambda_{ymk},\gamma_{mk})$ are the Normal-Gamma distribution (Broemeling, 1985).

From the prior distribution of Φ_m^{-1} and the distribution of $\mathbf{X}_{m,2}$, it can be shown that

$$p(\Phi_m|\mathbf{X}_{m,2})\propto|\Phi_m|^{-(n_m+\rho_0+q_2+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[\Phi_m^{-1}(\mathbf{X}_{m,2}\mathbf{X}_{m,2}^T+R_0^{-1})]\right\}.$$

Hence, $p(\Phi_m|\mathbf{X}_{m,2})$ is distributed as $IW[(\mathbf{X}_{m,2}\mathbf{X}_{m,2}^T+R_0^{-1}),n_m+\rho_0,q_2]$. It can be shown from exactly the same reasoning as before that the full conditional distributions for $(\psi_{\delta mk},\Lambda_{\xi mk})$ are distributed as given in (15).

Appendix B: Manifest Variables in the ICPSR Example

The number of the variable corresponding to the original data set is given in parenthesis at the end of each statement.

y^1: Overall, how satisfied or dissatisfied are you with your job? (V116)
y^2: How free are you to make decisions in your job? (V117)
y^3: Overall, how satisfied are you with your home life? (V 180)
y^4: How satisfied are you with the financial situation of your household? (V132)
y^5: All things considered, how satisfied are you with your life as a whole in these days? (V96)

y^6: In the long run, hard work usually bring a better life. (V255)
y^7: Competition is good. It simulates people to work hard and develop new ideas. (V254)
y^8: Individual should take more responsibility for providing for themselves. (V252)

References

Albert, J.H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669–679.

Albert, J.H., & Chib, S. (1995). Bayesian residual analysis for binary regression models. *Biometrika, 82*, 747–759.

Arminger, G., & Muthén, B.O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika, 63*, 271–300.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data* (2nd ed). Chichester: Wiley & Sons.

Bentler, P.M. (1983). Some contributions to efficient statistics for structural models: Specification and estimation of moment structures. *Psychometrika, 48*, 493–517.

Bentler, P.M. (1992). *EQS: Structural equation program manual*. Los Angeles, CA: BMDP Statistical Software.

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science, 10*, 3–66.

Bollen, K.A., & Long, J.S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Broemeling, L.D. (1985). *Bayesian analysis of linear models*. New York, NY: Marcel Dekker.

Browne, M.W. (1984). Asymptotically distribution-free methods in the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology, 37*, 62–83.

Byrne, B.M. (1994). *Structural equation modeling with EQS and EQS/Windows*, Thousand Oaks, CA: Sage.

Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika, 78*, 637–644.

Chaloner, K., & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika, 75*, 651–659.

Cook, R.D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman and Hall.

Crawford, S.L., DeGroot, M.H., Kadane, J.B., & Small, M.J. (1992). Modeling lake chemistry distributions: Approximate Bayesian methods for estimating a finite mixture model. *Technometrics, 34*, 441–453.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions, *Biometrika, 56*, 463–474.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Diebolt, J., & Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, 56*, 163–175.

Do, K., & McLachlan, G.J. (1984). Estimation of mixing proportions: A case study. *Applied Statistics, 33*, 134–140.

Dolan, C.V., & van der Maas, Han L.J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika, 63*, 227–254.

Ganesalingam, S., & McLachlan, G.J. (1981). Some efficiency results for the estimation of the mixing proportion in a mixture of two normal distributions. *Biometrics, 37*, 23–33.

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Gelman, A., Meng, X.L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistics Sinica, 6*, 733–807.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo. *Statistical Science, 7*, 473–511.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 1–19). London: Chapman and Hall.

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

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions, *Technometrics, 8*, 431–444.

Hawkins, D.M. (1980). *Identification of outliers*. London: Chapman and Hall.

Hoyle, R. (Ed.). (1995). *Structural equation modeling: Concepts, issues and applications*. Thousand Oaks, CA: Sage.

Jedidi, K., Jagpal, H.S., & DeSarbo, W.S. (1997a). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*, 39–59.

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

Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43*, 443–477.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hove and London: Scientific Software International.

Lee, S.Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika, 46*, 153–160.

Lee, S.Y., & Poon, W.Y. (1992). Two-level analysis of covariance structures for unbalanced designs with small level-one samples. *British Journal of Mathematical and Statistical Psychology, 45*, 109–124.

Lee, S.Y., & Tsui, K.L. (1982). Covariance structure analysis in several populations. *Psychometrika, 57*, 409–426.

Lindsay, B.G. (1989). Moment matrices: Applications in mixtures. *The Annals of Statistics, 17*, 722–740.

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

Lindley, D.V., & Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34*, 1–42.

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

Meng, X.L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*, 1142–1160.

Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). *Introduction to the theory of statistics* (3rd ed.). London: McGraw-Hill.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Raiffa, H., & Schlaifer, L.R. (1961). *Applied statistical secision theory* (2nd ed.). Boston, MA: Harvard, Graduate School of Business Administration.

Redner, R.A., & Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26*, 195–239.

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

Robert, C.P. (1996). Mixtures of distributions: Inference and estimation. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 163–187). London: Chapman and Hall.

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

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151–1172.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64*, 37–52.

Shi, J.Q., & Lee, S.Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51*, 233–252.

Smith, A.F.M., & Makov, U.E. (1978). A quasi-Bayes sequential procedure for mixtures. *Journal of the Royal Statistical Society, Series B, 40*, 106–112.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley & Sons.

van Buuren, S. (1997). Fitting ARMA time series by structural equation models. *Psychometrika, 62*, 215–236.

Weiss, R.E. (1994). *Residuals and outliers in Bayesian random effects models* (Tech. Rep.). Los Angeles, CA: UCLA, School of Public Health, Department of Biostatistics.

Weiss, R.E., & Lazaro, C.G. (1992). Residual plots for repeated measures. *Statistics in Medicine, 11*, 115–124.

World Values Survey, 1981–1984 and 1990–1993 (ICPSR version). (1994). Ann Arbor, MI: Institute for Social Research [producer], Inter-university Consortium for Political and Social Research [distributor].

Yuan, K.E., & Bentler, P.M. (1997). Mean and covariance structures analysis: theoretical and practical improvements. *Journal of the American Statistical Association, 92*, 767–774.

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

Zeger, S.L., & Karim, M.R. (1991). Generalized linear models with random effects: A gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79–86.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, New York, NY: Wiley & Sons.

Zhu, H.T., & Lee, S.Y. (1999). Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology, 52*, 225–242.

Manuscript received 20 NOV 1998
Final version received 6 APR 2000