PSYCHOMETRIKA—VOL. 67, NO. 2, 189–210
JUNE 2002

MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR STRUCTURAL EQUATION MODELS

SIK-YUM LEE

DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG

HONG-TU ZHU

DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF VICTORIA

The existing maximum likelihood theory and its computer software in structural equation modeling are established based on linear relationships among manifest variables and latent variables. However, models with nonlinear relationships are often encountered in social and behavioral sciences. In this article, an EM type algorithm is developed for maximum likelihood estimation of a general nonlinear structural equation model. To avoid computation of the complicated multiple integrals involved, the E-step is completed by a Metropolis-Hastings algorithm. It is shown that the M-step can be completed efficiently by simple conditional maximization. Standard errors of the maximum likelihood estimates are obtained via Louis's formula. The methodology is illustrated with results from a simulation study and two real examples.

Key words: nonlinear structural equation models, missing data, MCECM algorithm, Metropolis-Hastings algorithm, standard errors estimates.

1. Introduction

Structural equation modeling is a popular statistical method in behavioral and social sciences. The existing theory and computer software such as LISREL (Jöreskog & Sörbom, 1996) and EQS (Bentler, 1992) are developed based on models in which manifest variables and latent variables are related by linear functions. Recently, it is recognized that nonlinear relations among the variables are important in establishing more meaningful and correct models for some complex situations. For example, see Busemeyer and Jones (1993), Bollen and Paxton (1998), Jonsson (1998), Ping (1996b, 1996c), Kenny and Judd (1984), Bagozzi, Baumgartner and Yi (1992), Schumacker and Marcoulides (1998), and references therein on the importance of quadratic and interaction effects of latent factors in various applied research.

Due to the complex distribution associated with the nonlinear latent variables, methods for analyzing nonlinear structural equation models are more difficult and less familiar. Nonlinear factor analysis model with polynomial relationships was first explored by McDonald (1962, 1967a, 1967b), then followed by Etezadi-Amoli and McDonald (1983), and Mooijaart and Bentler (1986). Recently, useful methods that used LISREL (Jöreskog & Sörbom, 1996), EQS (Bentler, 1992) or the COSAN (Fraser, 1980) program have been proposed to analyze some nonlinear structural equation models with quadratic and interaction terms of latent variables; see, for example, Kenny and Judd (1984), Ping (1996a, 1996b), Jaccard and Wan (1995), and

The order of the authorship is alphabetical. This research is fully supported by a Hong Kong UGC Earmarked grant CUHK 4088/99H. The authors are indebted to the Editor, Associate Editor and anonymous reviewers for valuable comments for improving the paper; and also to ICPSR and the relevant funding agency for allowing use of the data. We thank Xin-Yuan Song, N.H. Tang and Liang Xu for helpful discussions. Assistance of Xin-Yuan Song and Michael K. H. Leung in analyzing the real examples and Esther L.S. Tam in preparing the manuscript is also 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

0033-3123/2002-2/1999-0749-A $00.75/0

Jöreskog and Yang (1996), among others. The basic approach of these contributions is to include artifical nonlinear manifest variables in the analysis to account for nonlinear relationships among variables. This approach has the following practical and theoretical difficulties (see, e.g., Arminger & Muthén, 1998): (i) It is usually tedious and difficult to derive covariances/variances among the nonlinear variables. (ii) The manifest random vector that involves nonlinear functions of variables is clearly not normal, and can be very complex. For correct statistical inference, asymptotically-distribution-free (ADF) theory (Bentler, 1983; Browne, 1984) has to be used. It is well known that (see, e.g., Bentler & Dudgeon, 1996; Hu, Bentler & Kano, 1992) ADF theory requires very large sample sizes to attain its asymptotic properties. The maximum likelihood (ML) method is also a large-sample one, but it takes a smaller sample size to attain its asymptotic properties than the ADF theory. (iii) Since the size of the weight matrix required in the ADF estimation is increased rapidly with the number of the manifest variables, one may encounter computational and storage problems in the estimation. To solve these difficulties, a Bayesian approach which utilizes some Markov chain Monte Carlo methods has been developed recently by Arminger and Muthén (1998), see also Lee and Zhu (2000).

The main objective of this article is to investigate the ML estimation of a general nonlinear structural equation model. It is well recognized that the ML approach is an important statistical method. As a general statistical procedure, its optimal properties such as consistency and efficiency have been well developed in the literature. It is also the foundation of many important statistical methods; for example, the likelihood ratio test and Bayes factor (Berger & Perrichi, 1996; Kass & Raftery, 1995), among others. However, it is not our intention here to compare the ML approach with the Bayesian approach or to claim that the ML approach is better. In the general statistical literature, there are many articles to discuss and compare these two important approaches, and it seems that a definite conclusion is not yet arrived. Like other statistical areas, our purpose is to develop the important ML method as a complementary procedure to the Bayesian approach.

Another main objective of this article is to demonstrate how recent advances in computational tools in statistics can offer reliable algorithm in computing ML estimates for complicated psychometric models such as the nonlinear structural equation model. Specifically, we treat the basic latent variables as missing data and apply the EM algorithm (Dempster, Laird, & Rubin, 1977) in getting ML estimates. Owing to the complexity of the nonlinear model, the E-step is intractable. We will demonstrate how this difficulty can be solved by using the Metropolis-Hastings (MH) algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The M-step also does not have a closed form solution. Thus, the method of conditional maximization (Meng & Rubin, 1993) is used to complete this step. A lot of computer time can be saved by conditional maximization (CM). The whole algorithm, which can be regarded as an MCECM algorithm, is rather efficient. In addition, we show how the method of bridge sampling (Meng & Wong, 1996) can be used to handle the important issue on determining convergence of the MCECM algorithm (see also Meng & Schilling, 1996). According to our knowledge, applications of the MCECM algorithm and bridge sampling have not been discussed in structural equation modeling. A method based on Louis formula (1982) in calculating standard errors of the estimations is also described.

Klein, Moosbrugger, Schermelleh-Engel, and Frandk (1997) proposed an iterative procedure in computing ML estimates for parameters in the elementary interaction model proposed by Kenny and Judd (1984). Their procedure also employed the general principle of the EM algorithm. The standard errors were obtained by evaluating the Fisher information matrix. When compared with a LISREL-based method, their simulation results showed that both methods lead to consistent estimates, but standard errors and confidence intervals calculated with their technique were more accurate. Apparently, the model considered in this paper is more general than theirs. Moreover, their computation of the E-step and the M-step, as well as the monitoring of convergence are quite different from our MCECM algorithm. Finally, their method of getting standard errors is also different.

The paper is organized as follows. Section 2 presents the maximum likelihood estimation of a non-recursive nonlinear structural equation model. Components in the implementation of the MCECM algorithm are discussed. This algorithm gives ML estimates of structural parameters in the covariance structure, and direct estimates of latent variables. Results of a simulation study and two real examples are presented in section 3. In the first real example, ML estimate will be compared with the Bayesian estimate (Arminger & Muthén, 1998) and an ad hoc estimate (Jaccard & Wan, 1995). A discussion is given in section 4. Throughout the paper, $p(\cdot|\cdot)$ and $[\cdot|\cdot]$ will denote appropriate conditional density function and conditional distribution, respectively.

2. ML Estimation of a Nonlinear Structural Equation Model

2.1. A Nonlinear Structural Equation Model

Consider the following nonlinear structural equation (NSEQ) model with a $p \times 1$ manifest random vector $y = (y^1, \ldots, y^p)^T$:

$$y = \mu + \Lambda\xi + \epsilon, \tag{1}$$

where μ is a vector of intercepts, Λ is a $p \times q$ matrix of factor loadings, $\xi = (\xi^1, \ldots, \xi^q)^T$ is a random vector of latent factors with $q < p$, ϵ is a $p \times 1$ random vector of error measurements with distribution $N[0, \Psi]$, where Ψ is diagonal and ϵ is independent with ξ. To handle more complex situations, we partition ξ as $(\xi_{(1)}^T, \xi_{(2)}^T)^T$ and further model this latent vector via the following nonlinear structural model:

$$\xi_{(1)} = \Pi\xi_{(1)} + \Gamma H(\xi_{(2)}) + \delta, \tag{2}$$

where $\xi_{(1)}$ and $\xi_{(2)}$ are $q_1 \times 1$ and $q_2 \times 1$ latent subvectors of ξ respectively; $H(\xi_{(2)}) = (h_1(\xi_{(2)}), \ldots, h_t(\xi_{(2)}))^T$ is a $t \times 1$ nonzero vector-valued function with nonzero, linearly independent differential functions $h_1, \ldots, h_t$, and $t \geq q_2$; $\Pi(q_1 \times q_1)$ and $\Gamma(q_1 \times t)$ are matrices of regression coefficients of $\xi_{(1)}$ on $\xi_{(1)}$ and $H(\xi_{(2)})$, respectively. It is assumed that $\xi_{(2)}$ and δ are independently distributed as $N[0, \Phi]$ and $N[0, \Psi_\delta]$, respectively, where Ψ_δ is a diagonal covariant matrix. If some $h_j(\xi_{(2)})$ are nonlinear, the distribution of the manifest random vector y is non-normal. Let $\Pi_0 = I_{q_1} - \Pi$, we assume that $|\Pi_0|$ is a nonzero constant independent of Π. The structural model in (2) is linear in parameter matrices Π and Γ, but may be nonlinear in the latent variables in $\xi_{(2)}$. Hence, nonlinear causal effects of latent variables in $\xi_{(2)}$ on latent variables in $\xi_{(1)}$ can be assessed. Let $\Lambda_\xi = (\Pi, \Gamma)$ and $G(\xi) = (\xi_{(1)}^T, H(\xi_{(2)})^T)^T$, then (2) can be written as

$$\xi_{(1)} = \Lambda_\xi G(\xi) + \delta.$$

Let $y = (y_{(1)}^T, y_{(2)}^T)^T$, $\mu = (\mu_{(1)}^T, \mu_{(2)}^T)^T$, $\epsilon = (\epsilon_{(1)}^T, \epsilon_{(2)}^T)^T$ and

$$\Lambda = \begin{pmatrix} \Lambda_{(1)} & 0 \\ 0 & \Lambda_{(2)} \end{pmatrix}$$

be appropriate partitions that correspond to the partition of ξ, then the measurement equations given in (1) are equivalent to

$$y_{(1)} = \mu_{(1)} + \Lambda_{(1)}\xi_{(1)} + \epsilon_{(1)},$$
$$y_{(2)} = \mu_{(2)} + \Lambda_{(2)}\xi_{(2)} + \epsilon_{(2)}.$$

These measurement equations together with the structural equation (2) define a nonlinear LISREL type model that is similar to that in Arminger & Muthén (1998).

The E-step and its associated MH algorithm in the proposed MCECM algorithm for ML estimation require the assumption that the conditional expectations of $\xi_{i(1)}\xi_{i(1)}^T$, $G(\xi_i)G(\xi_i)^T$, and $\xi_{i(1)}G(\xi_i)^T$ (see subsection 2.3) given the observed data are well defined. Hence, functions in $H(\xi_{i(2)})$ have to satisfy some additional conditions. It can be shown that functions corresponding to interaction and quadratic effects satisfy the conditions, hence almost all types of nonlinearity considered in the literature can be assessed. To save space, detailed technical discussions about precise conditions on $H(\xi_{(2)})$ are not given, they can be obtained from the authors upon request.

Another assumption of the proposed model is that $|\Pi_0|$ is a nonzero constant independent of Π so that $\partial|\Pi_0|/\partial\Pi$ is zero. Let $\xi_{(1)} = (\xi^1, \xi^2)^T$ and $\xi_{(2)} = (\xi^3, \xi^4)^T$, some examples of structural equation with nonzero 2 by 2 matrix Π that satisfies this assumption are:

$$\begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \pi & 0 \end{pmatrix} \begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix} \begin{pmatrix} \xi^3 \\ \xi^4 \\ \xi^3\xi^4 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix},$$

$$\begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} = \begin{pmatrix} 0 & \pi \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \xi^1 \\ \xi^2 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} \end{pmatrix} \begin{pmatrix} \xi^3 \\ \xi^4 \\ \xi^3\xi^3 \\ \xi^3\xi^4 \\ \xi^4\xi^4 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}.$$

As far as we know, all existing theory and illustrative examples in nonlinear structural equation modeling, including those cited in the 'Introduction', are established with $\Pi = 0$ (hence $|\Pi_0| = 1$). So the proposed NSEQ model is rather general and most of the existing models that involve interaction and quadratic terms of latent variables in $\xi_{(2)}$ are its special cases. The above structural equations allow ξ^1 to regress on ξ^2 or ξ^2 to regress on ξ^1. However, the underlying models do not allow ξ^1 to regress on ξ^2 and ξ^2 to regress on ξ^1 simultaneously, because under this situation, the matrix Π is of the type

$$\begin{pmatrix} 0 & \pi_{12} \\ \pi_{21} & 0 \end{pmatrix},$$

and $|I - \Pi|$ depends on parameters in Π. For simplicity and efficiency in completing the M-step of the MCECM algorithm, we will develop our methodology on the basis of the assumption that $|\Pi_0|$ is a constant independent of Π. However, the main structure of the MCECM algorithm can be used for handling general models without this assumption. Please refer to the Discussion section for more details.

The proposed nonlinear model is over parameterized without appropriate identification conditions. For instance, an equivalent form of model (1) is $y = \mu + \Lambda^*\xi^* + \epsilon$, where $\Lambda^* = \Lambda R$ and $\xi^* = R^{-1}\xi$ for any nonsingular matrix R. One common method to solve the identification problem in structural equation modeling is to fix appropriate elements in Λ at some known values so that the only possible choice of R is the identity matrix. Similarly, appropriate elements in Λ_ξ may also be fixed at known values if necessary. These fixed known values are not treated as parameters and required some attention in the implementation of the MCECM algorithm, see subsection 2.4.

2.2. MCECM Algorithm

Let $\mathbf{Y} = \{y_1, \ldots, y_n\}$ be the observed data matrix corresponding to a random sample obtained from a population with the NSEQ model defined in (1) and (2), $\mathbf{Z} = (\xi_1, \ldots, \xi_n)$ be the matrix of latent factors, and θ be the structural parameter vector that contains all unknown distinct parameters in μ, Λ, Λ_ξ, Φ, Ψ_δ and Ψ. The log-likelihood function $L_o(\mathbf{Y}|\theta) = \log p(\mathbf{Y}|\theta)$ based on the observed data $\mathbf{Y}$ can be written as

$$
\begin{aligned}
L_o(\mathbf{Y}|\theta) = {} & -\frac{n}{2}\{\log|\Psi| + \log|\Psi_\delta| + \log|\Phi| - 2\log|\Pi_0| + (p+q)\log(2\pi)\} \\
& + \sum_{i=1}^{n} \log \int \exp\big\{-\tfrac{1}{2}(y_i - \mu - \Lambda\xi_i)^T\Psi^{-1}(y_i - \mu - \Lambda\xi_i)\big\} \\
& \times \exp\big\{-\tfrac{1}{2}\big[\xi_{i(2)}^T\Phi^{-1}\xi_{i(2)} + (\xi_{i(1)} - \Lambda_\xi G(\xi_i))^T\Psi_\delta^{-1}(\xi_{i(1)} - \Lambda_\xi G(\xi_i))\big]\big\}d\xi_i. \quad (3)
\end{aligned}
$$

Owing to the nonlinearity of $G(\xi)$, the multiple integral involved in this log-likelihood function usually does not have an explicit form and its dimension is equal to the dimension of ξ_i. Hence, it is very difficult to obtain ML estimate by direct maximization of $L_o(\mathbf{Y}|\theta)$.

The basic idea of our approach is to consider a data augmentation scheme in which the observed data $\mathbf{Y}$ is augmented with the matrix of latent factors $\mathbf{Z}$. Treating this as a missing data problem with hypothetical missing data $\mathbf{Z}$, ML estimate is obtained by the well-known EM algorithm (Dempster et al., 1977). This common strategy has been widely applied to solve many statistical problems (see, e.g., Meng & van Dyk, 1997, and references therein); in particular, see Rubin and Thayer (1982), Liu and Rubin (1998), Shi and Lee (2000), and Lee and Tsang (1999) for applications to single-level and multi-level linear factor analysis models.

Let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ be the augmented completed-data set and $L_c(\mathbf{Y}, \mathbf{Z}|\theta) = \log p(\mathbf{X}|\theta)$ be the log-likelihood function of θ based on $\mathbf{X}$. From (1) and (2), it can be shown that $L_c(\mathbf{X}|\theta)$ is given by

$$
\begin{aligned}
-\tfrac{1}{2}\Bigg\{ & (p+q)n\log(2\pi) + n\log|\Psi| + n\log|\Psi_\delta| + n\log|\Phi| - 2n\log|\Pi_0| \\
& + \sum_{i=1}^{n}\xi_{i(2)}^T\Phi^{-1}\xi_{i(2)} + \sum_{i=1}^{n}(y_i - \mu - \Lambda\xi_i)^T\Psi^{-1}(y_i - \mu - \Lambda\xi_i) \\
& + \sum_{i=1}^{n}(\xi_{i(1)} - \Lambda_\xi G(\xi_i))^T\Psi_\delta^{-1}(\xi_{i(1)} - \Lambda_\xi G(\xi_i))\Bigg\}. \quad (4)
\end{aligned}
$$

The ξ_i in (4) are not observed random variables, and they are considered as missing data. Comparing (3) and (4), it can be seen that $L_o(\mathbf{Y}|\theta)$ is much more complicated than $L_c(\mathbf{X}|\theta)$. The r-th iteration of a standard EM algorithm with a current value $\theta^{(r)}$ is to evaluate $Q(\theta|\theta^{(r)}) = \mathrm{E}\{L_c(\mathbf{X}|\theta)|\mathbf{Y}, \theta^{(r)}\}$, where the expectation is taken with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ and $\theta^{(r)}$, and then to determine $\theta^{(r+1)}$ by maximizing $Q(\theta|\theta^{(r)})$. Evaluation of the E step is rather complicated because the conditional expectation involves complex multiple integrals that are intractable. A procedure on the basis of the MH algorithm is proposed here to execute this step. Since the M-step also does not have a closed form solution, $\theta^{(r+1)}$ will be obtained via a sequence of conditional maximization steps (see Meng & Rubin, 1993). Thus the proposed EM algorithm can be regarded as a MCECM algorithm (Wei & Tanner, 1990). As we will see, computational burden of the MCECM algorithm in solving our problem is not heavy.

2.3. *Implementing the E-step via the MH Algorithm*

It can be seen from (4) that to evaluate $Q(\theta|\theta^{(r)})$ in the E-step, we need to compute the conditional expectations of the following complete-data sufficient statistics $\{\xi_i, \xi_i\xi_i^T, G(\xi_i)G(\xi_i)^T, \xi_{i(1)}G(\xi_i)^T; i = 1, \ldots, n\}$. Owing to the generality and complexity of $G(\xi_i)$, these quantities cannot be obtained in closed form. They are approximated via a sufficiently large number of ξ_i simulated from $p(\xi_i|y_i, \theta)$, see Wei and Tanner (1990). Based on the definition of the model and assumptions, $p(\xi_i|y_i, \theta)$ is proportional to

$$
\begin{aligned}
\exp\Big\{ & -\tfrac{1}{2}\xi_{i(2)}^T\Phi^{-1}\xi_{i(2)} - \tfrac{1}{2}(y_i - \mu - \Lambda\xi_i)^T\Psi_\epsilon^{-1}(y_i - \mu - \Lambda\xi_i) \\
& -\tfrac{1}{2}(\xi_{i(1)} - \Lambda_\xi G(\xi_i))^T\Psi_\delta^{-1}(\xi_{i(1)} - \Lambda_\xi G(\xi_i))\Big\}. \quad (5)
\end{aligned}
$$

The Metropolis-Hastings (MH) algorithm (Geman & Geman, 1984; Hastings, 1970; and Metropolis et al., 1953) is a well-known MCMC method that has been widely used to simulate observations from a target density via the help of a proposal distribution from which it is easy to sample. Here $p(\xi_i|y_i, \theta)$ is taken as the target density. Based on the reasoning in Roberts (1996), it is natural to choose $N[\cdot, \sigma^2\Omega]$ as the proposal distribution, where σ^2 is a chosen value, and $\Omega^{-1} = \Sigma_\xi^{-1} + \Lambda^T\Psi^{-1}\Lambda$, with

$$\Sigma_\xi^{-1} = \begin{bmatrix} \Pi_0^T\Psi_\delta^{-1}\Pi_0, & -\Pi_0^T\Psi_\delta^{-1}\Gamma\Delta \\ -\Delta^T\Gamma^T\Psi_\delta^{-1}\Pi_0, & \Phi_\xi^{-1} + \Delta^T\Gamma^T\Psi_\delta^{-1}\Gamma\Delta \end{bmatrix}, \tag{6}$$

and $\Delta = \partial H(\xi_{i(2)})/\partial\xi_{i(2)}^T|_{\xi_{i(2)}=0}$. The MH algorithm (see Liu, Liang, & Wong, 2000) is implemented as follows: At the m-th iteration with a current value $\xi_i^{(m)}$, k_0 new candidates $x_1, \ldots, x_{k_0}$ are generated from $N[\xi_i^{(m)}, \sigma^2\Omega]$, and select ξ_i among the $x's$ with probability proportional to the target density $p(x_j|y_i, \theta)$ as given in (5), $j = 1, \ldots, k_0$. Then, draw $x_1^*, \ldots, x_{k_0-1}^*$ from $N[\xi_i, \sigma^2\Omega]$ and the new candidate ξ_i is accepted as a new current value with probability

$$\min\left\{1, \frac{p(x_1|y_i,\theta) + \cdots + p(x_{k_0}|y_i,\theta)}{p(\xi_i^{(m)}|y_i,\theta) + \Sigma_{j=1}^{k_0-1}p(x_j^*|y_i,\theta)}\right\}, \tag{7}$$

where $p(x|y, \theta)$ can be obtained via (5). The quantity σ^2 can be chosen such that the average acceptance rate is approximately 0.25 or more, see Gelman, Roberts and Gilks (1995). According to our empirical experience, the MH algorithm is very efficient for our problem. Other alternatives such as the "Independence sampler" or "Langevin-Hastings" algorithms can also be considered.

Let $\{\xi_i^{(m)}, m = 1, \ldots, M; i = 1, \ldots, n\}$ be the random observations generated by our proposed MH algorithm from the conditional distribution $[\xi_i|y_i, \theta]$. Conditional expectations of the complete-data sufficient statistics required to evaluate the E-step can be approximated via these random observations as follows:

$$\begin{aligned}
\mathrm{E}(\xi_i|y_i,\theta) &= M^{-1}\sum_{m=1}^{M}\xi_i^{(m)}, \\
\mathrm{E}(\xi_i\xi_i^T|y_i,\theta) &= M^{-1}\sum_{m=1}^{M}\xi_i^{(m)}\xi_i^{(m)T}, \\
\mathrm{E}(G(\xi_i)G(\xi_i)^T|y_i,\theta) &= M^{-1}\sum_{m=1}^{M}G(\xi_i^{(m)})G(\xi_i^{(m)})^T, \\
\mathrm{E}(\xi_{i(1)}G(\xi_i)^T|y_i,\theta) &= M^{-1}\sum_{m=1}^{M}\xi_{i(1)}^{(m)}G(\xi_i^{(m)})^T, \\
\mathrm{E}(\xi_{i(2)}\xi_{i(2)}^T|y_i,\theta) &= M^{-1}\sum_{m=1}^{M}\xi_{i(2)}^{(m)}\xi_{i(2)}^{(m)T}.
\end{aligned} \tag{8}$$

2.4. Maximization Step

At the M-step, we need to maximize $Q(\theta|\theta^{(r)})$ with respect to θ. This is equivalent to solve the following system of equations:

$$\frac{\partial Q(\theta|\theta^{(r)})}{\partial\theta} = \mathrm{E}\left\{\frac{\partial}{\partial\theta}L_c(\mathbf{X}|\theta)\middle|\mathbf{Y}, \theta^{(r)}\right\} = 0. \tag{9}$$

For $k = 1, \ldots, p$; $j = 1, \ldots, q_1$, let Λ_k and $\Lambda_{\xi j}$ be the k-th and the j-th rows of Λ and Λ_ξ respectively. It can be shown that

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial \mu} = \Psi^{-1} \sum_{i=1}^{n} [y_i - \mu - \Lambda \xi_i], \; \frac{\partial L_c(\mathbf{X}|\theta)}{\partial \Phi} = \frac{1}{2} \Phi^{-1} \sum_{i=1}^{n} [\xi_{i(2)} \xi_{i(2)}^T - \Phi] \Phi^{-1},$$

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial \Lambda_k} = \psi_k^{-1} \sum_{i=1}^{n} [y_{ki} - \mu_k - \Lambda_k \xi_i] \xi_i^T, \; \frac{\partial L_c(\mathbf{X}|\theta)}{\partial \Lambda_{\xi j}} = \psi_{\delta j}^{-1} \sum_{i=1}^{n} [\xi_{ji(1)} - \Lambda_{\xi j} G(\xi_i)] G(\xi_i)^T,$$

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial \mathrm{diag}(\Psi)} = \frac{1}{2} \mathrm{diag} \left\{ \Psi^{-1} \sum_{i=1}^{n} \left[(y_i - \mu - \Lambda \xi_i)(y_i - \mu - \Lambda \xi_i)^T - \Psi \right] \Psi^{-1} \right\},$$

and

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial \mathrm{diag}(\Psi_\delta)} = \frac{1}{2} \mathrm{diag} \left\{ \Psi_\delta^{-1} \sum_{i=1}^{n} \left[(\xi_{i(1)} - \Lambda_\xi G(\xi_i))(\xi_{i(1)} - \Lambda_\xi G(\xi_i))^T - \Psi_\delta \right] \Psi_\delta^{-1} \right\}. \tag{10}$$

These simultaneous equations cannot be solved in closed form. Based on the idea given in Meng and Rubin (1993), the solution required in the M-step can be obtained by several computationally simpler conditional maximizations. Conditional on other parameters, the solution of each individual equation given in (10) can be obtained. Solution for the M-step is given by the following results and (8): For $k = 1, \ldots, p$,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \{y_i - \hat{\Lambda} \mathrm{E}[\xi_i | \mathbf{Y}, \theta]\}, \quad \hat{\Phi} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[\xi_{i(2)} \xi_{i(2)}^T | \mathbf{Y}, \theta],$$

$$\hat{\Lambda}_k^T = \left(\sum_{i=1}^{n} \mathrm{E}[\xi_i \xi_i^T | \mathbf{Y}, \theta] \right)^{-1} \sum_{i=1}^{n} \mathrm{E}[\xi_i | \mathbf{Y}, \theta] (y_{ki} - \hat{\mu}_k),$$

$$\hat{\Lambda}_{\xi j}^T = \left(\sum_{i=1}^{n} \mathrm{E}[G(\xi_i) G(\xi_i)^T | \mathbf{Y}, \theta] \right)^{-1} \sum_{i=1}^{n} \mathrm{E}[G(\xi_i) \xi_{ji(1)} | \mathbf{Y}, \theta],$$

$$\hat{\psi}_{\delta j} = \frac{1}{n} \sum_{i=1}^{n} E[(\xi_{ji(1)} - \hat{\Lambda}_{\xi j} G(\xi_i))^2 | \mathbf{Y}, \theta],$$

$$\hat{\psi}_{kk} = \frac{1}{n} \sum_{i=1}^{n} \{(y_{ki} - \hat{\mu}_k)^2 - 2(y_{ki} - \hat{\mu}_k) \hat{\Lambda}_k \mathrm{E}[\xi_i | \mathbf{Y}, \theta] + \hat{\Lambda}_k \mathrm{E}[\xi_i \xi_i^T | \mathbf{Y}, \theta] \hat{\Lambda}_k^T\}. \tag{11}$$

The above results are valid for situations where all elements of Λ and Λ_ξ are free parameters. To identify the model, some appropriate elements in Λ and Λ_ξ are fixed at known values. To deal with this situation in general, consider the following linear transformations $\Lambda_k^T = A_k \Lambda_k^{*T} + a_k$, for $k = 1, \ldots, p$, and $\Lambda_{\xi j}^T = B_j \Lambda_{\xi j}^{*T} + \mathbf{b}_j$, for $j = 1, \ldots, q_1$, where $a_k(q \times 1)$ and $b_j(q_1 \times 1)$ are constant column vectors, $A_k(q \times r_k)$ and $B_j(q_1 \times q_{1j})$ are full column rank selection matrices, and $\Lambda_k^*(1 \times r_k)$ and $\Lambda_{\xi j}^*(1 \times q_{1j})$ are reduced parameter vectors of Λ_k and Λ_ξ, respectively. Then,

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial \Lambda_k^{*T}} = A_k^T \psi_k^{-1} \sum_{i=1}^{n} \xi_i [y_{ki} - \mu_k - \xi_i^T \Lambda_k^T],$$

which yields

$$\hat{\Lambda}_k^* = \left(A_k^T \sum_{i=1}^n \mathrm{E}[\xi_i \xi_i^T | \mathbf{Y}, \theta] A_k \right)^{-1}$$

$$\times A_k^T \sum_{i=1}^n \{\mathrm{E}[\xi_i | \mathbf{Y}, \theta](y_{ki} - \hat{\mu}_k) - \mathrm{E}[\xi_i \xi_i^T | \mathbf{Y}, \theta] a_k\}, \tag{12}$$

and $\hat{\Lambda}_k^T = A_k \hat{\Lambda}_k^{*T} + a_k$. Similarly, we have

$$\hat{\Lambda}_{\xi j}^* = \left(B_j^T \sum_{i=1}^n \mathrm{E}[G(\xi_i) G(\xi_i)^T | \mathbf{Y}, \theta] B_j \right)^{-1}$$

$$\times B_j^T \sum_{i=1}^n \{\mathrm{E}[G(\xi_i) \xi_{ji(1)} | \mathbf{Y}, \theta] - \mathrm{E}[G(\xi_i) G(\xi_i)^T | \mathbf{Y}, \theta] \mathbf{b}_j\}, \tag{13}$$

and $\hat{\Lambda}_{\xi j}^T = B_j \hat{\Lambda}_{\xi j}^{*T} + \mathbf{b}_j$.

An estimate of $\mathrm{E}(\xi_i | y_i, \hat{\theta})$, which can be used as an estimate for ξ_i, can be obtained easily via corresponding sample means of the generated observations in $\mathbf{Z}$ from the MH algorithm at the last EM iteration:

$$\hat{E}(\xi_i | y_i, \hat{\theta}) = \hat{\xi}_i = M^{-1} \sum_{m=1}^M \xi_i^{(m)}. \tag{14}$$

2.5. Monitoring Convergence ofMCECM

Owing to the simulation variability introduced at its E steps, the sequence of θ produced by an MCECM algorithm typically exhibits random fluctuation around a stationary point, even at convergence. Hence, it is generally difficult to claim convergence of an MCECM algorithm according to the standard criterion that differences between consecutive iterates are within a desired level. Wei and Tanner (1990) suggested plotting $\theta^{(r)}$ against r. If the number of unknown parameters is large, some functions of $\theta^{(r)}$ can be used. One function is the likelihood function value or the difference of consecutive likelihood values. For our problem, the actual likelihood based on the observed data cannot be evaluated analytically, see (3). However, as pointed out by Meng and Schilling (1996), in monitoring convergence of a likelihood only changes in likelihood values are of interest. Hence, the bridge sampling method (Meng and Wong, 1996) based on the likelihood ratio is used to monitor convergence of the proposed MCECM algorithm.

The bridge sampling method (Meng & Wong, 1996) is to monitor convergence via the following likelihood ratio:

$$R(\theta^{(r+1)}, \theta^{(r)}) = \frac{p(\mathbf{Y}|\theta^{(r+1)})}{p(\mathbf{Y}|\theta^{(r)})}.$$

This ratio is estimated via the identity

$$R(\theta^{(r+1)}, \theta^{(r)}) = \frac{E_r[p(\mathbf{Y}, \mathbf{Z}|\theta^{(r+1)}) \alpha(\mathbf{Z})]}{E_{r+1}[p(\mathbf{Y}, \mathbf{Z}|\theta^{(r)}) \alpha(\mathbf{Z})]},$$

where E_r denotes the conditional expectation of $\mathbf{Z}$ given $(\mathbf{Y}, \theta^{(r)})$, and $\alpha(\mathbf{Z})$ is an appropriate 'bridge' function satisfying some mild conditions. According to the general suggestion in Meng and Wong, we take

$$\alpha(\mathbf{Z}) = \left[p(\mathbf{Y}, \mathbf{Z}|\theta^{(r)}) p(\mathbf{Y}, \mathbf{Z}|\theta^{(r+1)})\right]^{-1/2}.$$

Then, $R^{*(r)} = \log R(\theta^{(r+1)}, \theta^{(r)})$ can be approximated by

$$\hat{R}^{*(r)} = \log\left\{\sum_{m=1}^{M}\left[\frac{p(\mathbf{Y},\mathbf{Z}^{r,(m)}|\theta^{(r+1)})}{p(\mathbf{Y},\mathbf{Z}^{r,(m)}|\theta^{(r)})}\right]^{1/2}\right\} - \log\left\{\sum_{m=1}^{M}\left[\frac{p(\mathbf{Y},\mathbf{Z}^{r+1,(m)}|\theta^{(r)})}{p(\mathbf{Y},\mathbf{Z}^{r+1,(m)}|\theta^{(r+1)})}\right]^{1/2}\right\}, \tag{15}$$

where $\{\mathbf{Z}^{r,(m)}, m = 1, \ldots, M\}$ are simulated from $p(\mathbf{Z}|\mathbf{Y}, \theta^{(r)})$. In monitoring convergence of an MCECM algorithm, we plot $\hat{R}^{*(r)}$ against r. If the plot shows a curve converging to zero with a fluctuation that can be expected from the simulation sizes, then an approximate convergence, which is enough for the purpose of statistical inference (Meng & Schilling, 1996), has been achieved. See Meng and Wong for more theoretical aspects of this general method.

2.6. *Standard Error Estimates*

Standard errors estimates of the ML estimates can be obtained by inverting either the Hessian matrix or the information matrix of the log-likelihood function based on observed data $\mathbf{Y}$. However, these matrices are generally not in closed form. Hence, we use an identity of Louis (1982) and random samples generated from $p(\mathbf{Z}|\mathbf{Y}, \hat{\theta})$ in the MH algorithm to obtain standard error estimates. It follows from Louis that

$$-\frac{\partial^2 L_o(\mathbf{Y}|\theta)}{\partial\theta\partial\theta^T} = E_{\mathbf{Z}}\left\{-\frac{\partial^2 L_c(\mathbf{Y},\mathbf{Z}|\theta)}{\partial\theta\partial\theta^T}\right\} - Var_{\mathbf{Z}}\left\{\frac{\partial L_c(\mathbf{Y},\mathbf{Z}|\theta)}{\partial\theta}\right\}, \tag{16}$$

where expectations involved in (16) are taken with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ and θ, and the whole expression is evaluated at $\hat{\theta}$. These expectations are difficult to evaluate analytically, but they can be approximated respectively by the sample mean and the sample covariance matrix of the random samples $\{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{M^*}\}$ generated from $p(\mathbf{Z}|\mathbf{Y}, \hat{\theta})$ using the MH algorithm. Thus,

$$\begin{aligned}
-\left.\frac{\partial^2 L_c(\mathbf{Y}|\theta)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat{\theta}} &\approx M^{*-2}\left.\left(\sum_{m=1}^{M^*}\frac{\partial L_c(\mathbf{Y},\mathbf{Z}^{(m)}|\theta)}{\partial\theta}\right)\left(\sum_{m=1}^{M^*}\frac{\partial L_c(\mathbf{Y},\mathbf{Z}^{(m)}|\theta)}{\partial\theta}\right)^T\right|_{\theta=\hat{\theta}} \\
&+ M^{*-1}\sum_{m=1}^{M^*}\left[-\frac{\partial^2 L_c(\mathbf{Y},\mathbf{Z}^{(m)}|\theta)}{\partial\theta\partial\theta^T} - \left(\frac{\partial L_c(\mathbf{Y},\mathbf{Z}^{(m)}|\theta)}{\partial\theta}\right)\left(\frac{\partial L_c(\mathbf{Y},\mathbf{Z}^{(m)}|\theta)}{\partial\theta}\right)^T\right]_{\theta=\hat{\theta}}
\end{aligned} \tag{17}$$

Explicit formulas for the partial derivatives can be obtained via standard matrix differentiation. First partial derivatives are given in (10) and second partial derivatives can be obtained via standard matrix calculus.

3. Simulation Study and Examples

3.1. *Analysis of Simulation Data*

Results of a simulation study are presented to give some idea on the empirical performance of the proposed MCECM algorithm. Data are drawn from a population with a model whose latent variables are related by a nonlinear regression model with an interaction term and a quadratic term. More specifically, a NSEQ model defined in (1) and (2) with six manifest variables and three latent variables (ξ^1, ξ^2, ξ^3) is considered. The structure of the loading matrix Λ in (1) is given by

$$\Lambda^T = \begin{bmatrix} 1.0 & \lambda_{21} & \lambda_{31} & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & \lambda_{52} & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix},$$

with population values $\lambda_{21} = \lambda_{31} = 0.8$, and $\lambda_{52} = 0.6$. True population values of other unknown parameters are given by: $\psi_{ii} = 0.8$ for $i = 1, 2, 3$, $\psi_{44} = \psi_{55} = 0.6$, $\psi_{66} = 0.0$; $\mu_1 = \cdots = \mu_5 = 1.0$, and $\mu_6 = 0.0$. The latent variable $\xi_{(1)} = \xi^3$ is related with $\xi_{(2)} = (\xi^1, \xi^2)^T$ via the following nonlinear model

$$\xi^3 = \gamma_{11}\xi^1 + \gamma_{12}\xi^2 + \gamma_{13}\xi^1\xi^2 + \gamma_{14}\xi^1\xi^1 + \delta.$$

So, $\Pi = 0$ and $|\Pi_0| = 1$. Population values of the unknown parameters associated with this nonlinear model are given by $\Gamma = (\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}) = (1.0, -1.0, 0.8, -0.8)$, $(\phi_{11}, \phi_{12}, \phi_{22}) = (1.0, -0.5, 1.0)$ and $\psi_\delta = 0.6$. To identify the model, the one's and zero's in Λ and ψ_{66} are treated as fixed known parameters. Hence, there are a total of 22 unknown parameters. The sample sizes used were $n = 100$, 200 and 400. For each case, 100 replications were used to study the accuracy of the ML estimates and standard error estimates.

The proposed MCECM algorithm was applied to find the ML estimates. The algorithm is robust to starting values of the parameters. Hence, any reasonable ad hoc choice is acceptable. In this article, the starting values were taken as $\lambda_{ij} = 0.0$ for all i, j as specified in Λ, $\gamma_{1i} = 0.0$ for $i = 1, \ldots, 4$, $\Phi = I_2$, $\psi_{ii} = 1.0$ for $i = 1, \ldots, 5$, $\psi_\delta = 1.0$, and $\mu_i = 2.0$ for all $i = 1, \ldots, 6$. The iteration estimation procedure updates the parameters of θ in every cycle of the MCECM iteration which consists of the E-step and the M-step. In the E-step, we first simulated $\{\xi_i^{(m)};\ m = 1, \ldots, M;\ i = 1, \ldots, n\}$ using the MH algorithm as described in section 2.3. At the j-th iteration of the MH algorithm with a current $\xi^{(j)}$ and θ, we simulated $k_0 = 5$ observations $x_1, \ldots, x_5$ from $N[\xi^{(j)}, \Omega]$ (where $\sigma^2 = 1$ and Ω is given by (6)) and selected one ξ^* among these five $x's$ with probability proportional to the target density $p(x_j|y_i, \theta)$ via simple random sampling. Then, we simulated $x_1^*, \ldots, x_4^*$ from $N[\xi^*, \Omega]$, and the candidate ξ^* is accepted as a new observation with probability given in (7). This completes the j-th iteration of the MH algorithm. Here, we took $\sigma^2 = 1$ and obtained an approximate acceptance rate 0.33, which is close to 0.25 as suggested in Gelman et al. (1995). The choice of σ^2 can be increased or decreased if the rate of acceptance is far from 0.25. We discarded the first 10 MH iterations, and then took $M = 40$ draws to approximate the conditional expectations of the complete-data sufficient statistics according to (8). We observed that the proposed MH algorithm was rather efficient. This completes the E-step of the cycle. In the M-step, the parameters of θ were updated according to (11), using the approximate conditional expectations obtained in the E-step. This gives an updated parameter vector to continue with the next MCECM iteration, if necessary. The bridge sampling was used to monitor convergence of the MCECM algorithm. In this method, the log likelihood ratio at the $(r + 1)$-th iteration of the MCECM algorithm is approximated via (15), where $\{\mathbf{Z}^{r,(m)}, m = 1, \ldots, M\}$ are obtained from the observations $\{\xi_i^{(m)}, m = 1, \ldots, M; i = 1, \ldots, n\}$ simulated in the E-step of the r-th iteration. We recommend to claim convergence if the log likelihood ratios converge from above to zero, say less than a small positive value $\varepsilon^* = 0.01$. We observed from the first few replications of the simulation that the MCECM algorithm converged after about 100 iterations. To be conservative, the algorithm was stopped after 200 iterations and $\theta^{(200)}$ is regarded as the ML estimate of θ. Standard error estimates were obtained via (17) with $M^* = 2000$.

The bias of each parameter estimate, which is the difference between the true value and the mean of estimates based on 100 replications, was calculated. The following method is used to evaluate the accuracy of the standard error estimates (see Lee, Poon & Bentler, 1995). Let $SD(\theta(k))$ be the empirical standard deviation obtained from the 100 estimates of the k-th element of θ, and $SE(\theta(k))$ be the mean of the 100 standard error estimates of $\hat{\theta}(k)$ obtained via our methods based on (17); the ratio $SE(\theta(k))/SD(\theta(k))$ is used to evaluate the accuracy. For completeness, root mean squares (RMS) between the true values and the corresponding estimates obtained from the replications are also computed. The results obtained are reported in Table 1. From this table, we have the following findings: (i) From the "Bias" and "RMS" columns, it

TABLE 1.
Performance of ML estimation

	$n = 100$			$n = 200$			$n = 400$		
Para.	Bias	SE/SD	RMS	Bias	SE/SD	RMS	Bias	SE/SD	RMS
λ_{21}	0.027	1.125	0.144	0.007	1.069	0.094	−0.004	1.064	0.067
λ_{31}	0.027	1.141	0.142	−0.005	1.133	0.093	0.004	1.184	0.078
λ_{52}	0.006	1.156	0.123	0.011	1.230	0.091	0.001	0.905	0.055
γ_{11}	0.000	1.336	0.338	0.008	1.155	0.220	0.014	1.278	0.174
γ_{12}	−0.068	1.380	0.325	0.004	1.325	0.194	0.002	0.868	0.126
γ_{13}	0.027	1.204	0.352	0.006	1.208	0.242	−0.014	0.951	0.153
γ_{14}	−0.053	1.131	0.329	0.013	1.049	0.207	−0.008	1.017	0.136
μ_1	0.007	1.034	0.122	−0.001	1.001	0.088	−0.008	1.104	0.066
μ_2	0.027	1.001	0.113	0.006	1.066	0.085	−0.000	1.168	0.061
μ_3	0.015	0.990	0.110	−0.002	1.090	0.086	−0.006	1.114	0.059
μ_4	0.007	1.064	0.133	−0.007	1.003	0.100	0.003	1.038	0.065
μ_5	0.017	1.021	0.100	−0.004	0.963	0.071	0.010	0.985	0.047
μ_6	−0.021	1.058	0.224	−0.028	1.155	0.194	−0.015	1.100	0.119
ϕ_{11}	0.015	1.355	0.258	0.014	1.087	0.167	0.026	1.139	0.128
ϕ_{12}	−0.020	1.048	0.158	0.003	1.221	0.120	−0.013	0.979	0.082
ϕ_{22}	0.012	1.328	0.282	0.012	1.328	0.212	0.019	1.033	0.135
ψ_{11}	0.008	1.252	0.169	−0.014	0.881	0.098	0.002	0.927	0.072
ψ_{22}	−0.030	1.026	0.136	−0.005	0.963	0.090	0.017	0.893	0.064
ψ_{33}	0.023	0.922	0.126	0.006	1.009	0.099	0.003	1.122	0.079
ψ_{44}	0.014	1.302	0.193	−0.016	1.241	0.257	0.014	0.925	0.086
ψ_{55}	0.001	0.931	0.091	−0.001	1.005	0.075	0.006	1.007	0.056
ψ_δ	0.094	1.243	0.240	−0.013	0.757	0.139	0.006	0.815	0.102

TABLE 2.
RMS and correlations between factor scores estimates and true values

		$n = 100$		$n = 200$		$n = 400$	
Replications		RMS	CORR	RMS	CORR	RMS	CORR
20th	$\hat{\xi}_1$	0.398	0.915	0.448	0.848	0.397	0.911
	$\hat{\xi}_2$	0.445	0.897	0.521	0.880	0.490	0.876
40th	$\hat{\xi}_1$	0.399	0.944	0.428	0.902	0.410	0.922
	$\hat{\xi}_2$	0.540	0.869	0.476	0.870	0.485	0.878
60th	$\hat{\xi}_1$	0.398	0.933	0.474	0.882	0.419	0.903
	$\hat{\xi}_2$	0.570	0.876	0.469	0.879	0.479	0.889
80th	$\hat{\xi}_1$	0.451	0.902	0.400	0.907	0.372	0.925
	$\hat{\xi}_2$	0.468	0.906	0.461	0.904	0.468	0.867
100th	$\hat{\xi}_1$	0.447	0.922	0.417	0.915	0.410	0.911
	$\hat{\xi}_2$	0.471	0.852	0.483	0.867	0.450	0.889

seems that ML estimates obtained from the MCECM algorithm are quite close to the true values. (ii) Most of the ratios $SE(\theta(k))/SD(\theta(k))$ are close to 1.0, indicating standard errors estimates obtained by our method are accurate. (iii) As expected, accuracy of ML estimates increases with sample size.

To get some idea of the estimate for the latent variables, we compute the RMS and the correlation between the n estimates and their true values. Unlike estimation of the structural parameters that utilizes information from the whole sample $\{y_i, i = 1, \ldots, n\}$, the available information in estimating the true factor score is the single individual y_i with only p measurements. With this limited information, it is expected that the estimate may not be close to the true value and its standard error may be quite large. Moreover, accuracy of the estimates is not improved with large n. This phenomenon also appears in all statistical methods on estimation of factor scores in linear factor analysis model, see, for example, Shi and Lee (1998). However, the true factor scores and their estimates are highly correlated. To save space, only RMS and correlations corresponding to the 20th, ..., 100th replications are summarized in Table 2.

3.2. *Illustrative Examples*

A small portion of the Inter-university Consortium for Political and Social Research (ICPSR) data set collected in the project WORLD VALUES SURVEY 1981–1984 AND 1990–1993 (World Value Study Group, ICPSR Version) is analyzed in the illustrative examples. The whole data set was collected in 45 societies around the world on broad topics such as work, religious belief, the meaning and purpose of life, family life, contemporary social issues, etc. As an illustration of our proposed method, only the data obtained from the United Kingdom were used. The data analyzed in this paper can be obtained from the authors upon the approval of the ICPSR funding agencies. In the first example, six variables in the original data set (variables 180, 96, 62, 176, 116 and 117) that are related with respondents' job, religious belief, and homelife were taken as manifest variables in $y = (y^1, \ldots, y^6)^T$. Among them, (y^1, y^2) are related to life, (y^3, y^4) are related to religious belief, and (y^5, y^6) are related to job satisfaction. Variable 62 was measured by a five points scale, while all others were measured by a ten points scale. Since we are using this example only to illustrate the methodology, these variables are treated as continuous variables. After deleting cases with missing data, the sample size is 196.

A nonlinear structural equation model with latent factors $\xi_{(1)} = (\xi^1)$ and $\xi_{(2)} = (\xi^2, \xi^3)$ is proposed with the following specifications: $\Pi = 0$, $H(\xi_{(2)}) = (\xi^2, \xi^3, \xi^2\xi^3)$, and

$$\Gamma^T = \begin{bmatrix} \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \end{bmatrix}, \qquad \Lambda^T = \begin{bmatrix} 1.0 & \lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 & \lambda_{63} \end{bmatrix};$$

where one's and zero's in Λ were treated as fixed parameters. In this nonlinear model, there are a total of 22 structural parameters which are elements in μ and Λ, γ_{ij} in Γ, $\phi_{ij} (i \leq j)$ in Φ, diagonal elements in Ψ and ψ_δ. ML estimates of structural parameters and direct estimates of basic latent factors were obtained via the proposed MCECM algorithm. The σ in the proposal distribution of the MH algorithm was set equal to 1.0 and $k_0 = 5$, giving an approximately average acceptance rate 0.35. At the beginning of the MCECM algorithm, we use $M = 40$ observations obtained from the MH algorithm to approximate the conditional expectations in the E-step, then use $M = 500$ after the 45th iteration. It was found that the MCECM algorithm stabilized after about 45 iterations. To be conservative, the algorithm stopped after 60 iterations. Values of the log-likelihood ratios estimated by (15) are shown in Figure 1 after the 8th iteration. After convergence, a total of $M^* = 5000$ random observations were used to calculate standard errors estimates.

ML estimates of structural parameters and their standard errors estimates are reported in Table 3. From the structure of Λ, latent factors ξ^1, ξ^2 and ξ^3 can be roughly interpreted as "life", "religious belief" and "job satisfaction" factors, while $\xi^2\xi^3$ represents the interaction of

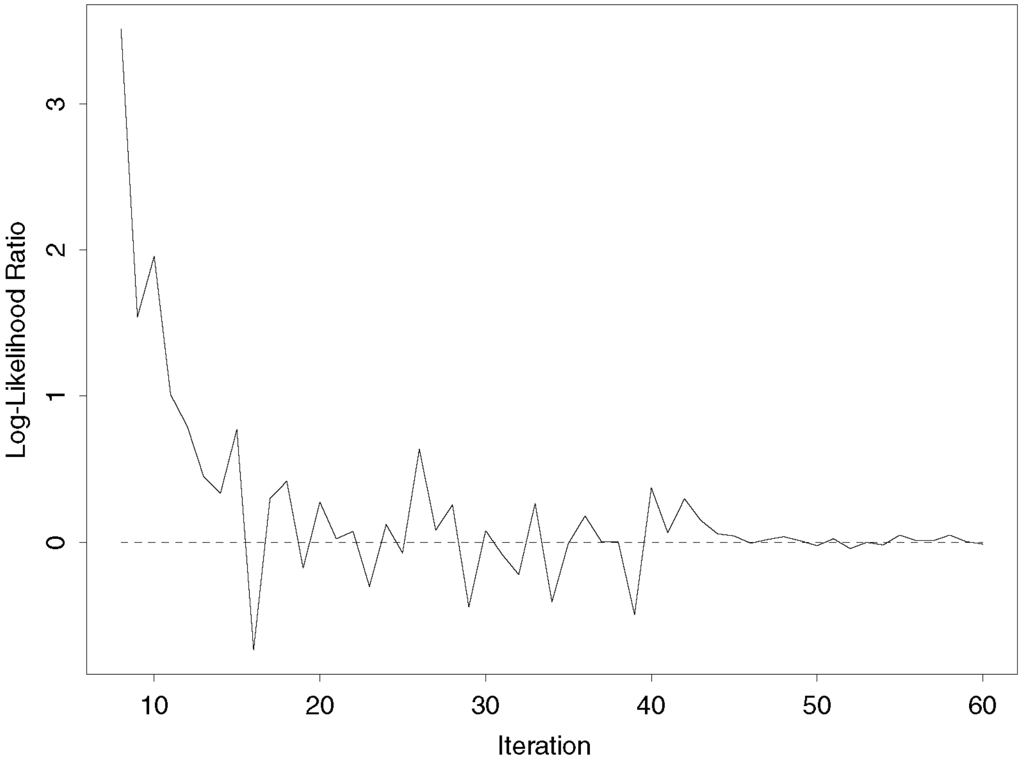

FIGURE 1.
ICPSR data with six manifest variables: Log-likelihood ratio versus EM iteration from the 8th iteration.

TABLE 3.
ML estimates and their standard errors for the ICPSR data

Parameter	EST	SD	Parameter	EST	SD
λ_{21}	0.856	0.143	ψ_{11}	0.626	0.283
λ_{42}	2.049	0.386	ψ_{22}	1.353	0.250
λ_{63}	0.795	0.139	ψ_{33}	0.674	0.307
γ_{11}	0.368	0.123	ψ_{44}	2.485	1.146
γ_{12}	0.590	0.126	ψ_{55}	1.960	0.565
γ_{13}	−0.191	0.086	ψ_{66}	4.066	0.493
μ_1	8.423	0.136	ϕ_{11}	1.780	0.449
μ_2	7.826	0.132	ϕ_{12}	−0.181	0.215
μ_3	2.350	0.147	ϕ_{22}	2.952	0.746
μ_4	5.511	0.297			
μ_5	7.551	0.169	ψ_δ	0.819	0.392
μ_6	7.357	0.180			

"religious belief" and "job satisfaction". From the estimate of γ_{13} and its standard error, we see that the corresponding t−value is −2.22. It seems that the interaction of "religious belief" and "job satisfaction" has a significant effect on "life". Based on the proposed approach, other nonlinear terms can also be analyzed similarly by choosing appropriate structures for Λ and $H(\xi)$.

According to suggestions of some reviewers, the ML estimate obtained will be compared with the Bayesian estimate. Following roughly the procedure given in Lee and Zhu (2001), a hybrid algorithm which is also a combination of the Gibbs sampler and the MH algorithm is implemented to produce the Bayesian estimate of θ in our proposed nonlinear model, using conjugate prior distributions. In the analysis of the illustrative ICPSR data set, hyper-parameters in the conjugate prior distributions were obtained via estimates with non-informative prior distributions. The hybrid algorithm in the actual estimation converged after 1800 iterations. A total of $K = 2000$ additional observations were collected after convergence, giving a sample $\{\theta^{(k)}; k = 1, \ldots, K\}$ from the appropriate posterior distribution of θ. The Bayesian estimate and its covariance matrix estimate are given respectively by

$$\tilde{\theta} = K^{-1}\sum_{k=1}^{K}\theta^{(k)}, \quad \text{and} \quad \widetilde{\text{Cov}}(\tilde{\theta}) = (K-1)^{-1}\sum_{k=1}^{K}(\theta^{(k)} - \tilde{\theta})(\theta^{(k)} - \tilde{\theta})^T.$$

Obtained results are reported in Table 4. It seems that ML estimates and Bayesian estimates are not significantly different. More discussion on the differences of these approaches will be given in the Discussion section.

A reviewer also suggested to compare the ML estimate with other estimates from the ad hoc procedures. According to findings given in Li, Hammer, Duncan, Duncan, Acock, and Boles (1998) on comparison of various ad hoc approaches, no substantial discrepancy was found in parameter estimates across different methods. Hence, to save space, we only consider the approach given in Jaccard and Wan (1996) in the comparison. Using the same data set, an ad hoc estimate and its standard error estimate were obtained with a multiple-indicator approach (Jaccard & Wan, 1996) using the LISREL 8 (Jöreskog & Sörbom, 1996) program with nonlinear constraints; see Li et al. (1998) for a set-up of the corresponding LISREL program set up. The ad hoc estimates

TABLE 4.
Bayesian and ad hoc estimates and their standard errors for the ICPSR data with six manifest variables

	Bayesian		Ad Hoc	
Parameter	EST	SD	EST	SD
λ_{21}	0.861	0.111	0.844	0.113
λ_{42}	2.232	0.270	2.030	0.350
λ_{63}	0.750	0.146	0.743	0.136
γ_{11}	0.398	0.116	0.303	0.085
γ_{12}	0.654	0.111	0.435	0.088
γ_{13}	−0.230	0.078	−0.095	0.048
ψ_{11}	0.657	0.204	0.616	0.264
ψ_{22}	1.399	0.216	1.373	0.230
ψ_{33}	0.837	0.182	0.636	0.322
ψ_{44}	2.186	0.685	2.352	1.322
ψ_{55}	2.001	0.443	1.244	0.663
ψ_{66}	4.118	0.530	4.004	0.481
ϕ_{11}	1.656	0.347	1.859	0.365
ϕ_{12}	−0.142	0.216	−0.151	0.232
ϕ_{22}	2.664	0.546	3.836	0.760
ψ_{δ}	0.661	0.231	1.293	0.309
μ_1	8.392	0.127	8.459	—
μ_2	7.826	0.138	7.857	—
μ_3	2.366	0.137	2.372	—
μ_4	5.580	0.240	5.556	—
μ_5	7.599	0.177	7.566	—
μ_6	7.378	0.185	7.388	—

are reported in Table 4. Discrepancies of estimates obtained from the ad hoc approach and the ML approach are not very significant for most parameters. However, for γ_{13} and ψ_δ that are directly related to the important interaction effects, the difference is quite substantial. And as expected, differences between the ad hoc estimate and the Bayesian estimate are more severe. Hence, interpretation of nonlinear causal effects on the basis of the ad hoc approach should be handled with care.

The main objective of the second example is for illustrating the methodology in handling a model in which $|\Pi_0|$ is a constant independent of Π but $\Pi \neq 0$. Two manifest variables y^7 and y^8 (corresponding to variables 252 and 254 in the questionnaire) measured with a ten-point scale about the job attitude were added in the first example. After removing cases with missing data, the sample size is 194. The data set was analyzed by a NSEQ model with four latent factors $\xi_{(1)} = (\xi^4, \xi^1)$, and $\xi_{(2)} = (\xi^2, \xi^3)$ with the following measurement and structural equations:

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \\ y^4 \\ y^5 \\ y^6 \\ y^7 \\ y^8 \end{pmatrix} = \mu + \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & \\ 0 & 1 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{63} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \lambda_{84} \end{pmatrix} \begin{pmatrix} \xi^1 \\ \xi^2 \\ \xi^3 \\ \xi^4 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \end{pmatrix},$$

$$\begin{pmatrix} \xi^4 \\ \xi^1 \end{pmatrix} = \begin{pmatrix} 0 & \pi \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \xi^4 \\ \xi^1 \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix} \begin{pmatrix} \xi^2 \\ \xi^3 \\ \xi^2\xi^3 \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix},$$

where one's and zero's in the coefficient matrices were treated as fixed parameters. Clearly, this is a NSEQ model with an interaction effect and $\Pi \neq 0$. There are a total of 31 parameters. ML estimates of structural parameters and direct estimates of basic latent factors were again obtained via the proposed MCECM algorithm. The σ in the proposal distribution of the MH algorithm was set equal to 1.0 and $k_0 = 5$, giving an approximately average acceptance rate 0.36. In the E-step of the r-th MCECM iteration, we simulated $M = 30 + 10r$ observations to approximate the conditional expectations. It was found that the MCECM algorithm stabilized after 35 iterations. To give some idea about the convergence, values of the log-likelihood ratios estimated by (15) are shown in Figure 2 after the 5th iteration. To be conservative, the parameters values at the 60th iteration are taken as the ML estimates. After convergence, a total of $M^* = 5000$ random observations were used to calculate standard errors estimates. ML estimates of structural parameters and their standard errors estimates are reported in Table 5. From the structure of Λ, latent factor ξ^4 can be interpreted as a "job attitude" factor, while the other latent factors ξ^1, ξ^2, and ξ^3 are interpreted as before. From the results in Tables 3 and 5, we see the expected result that the ML estimates of the original parameters in the previous simple model do not change substantially. The estimate of π is very small, indicating that the causal effect of "life" to "job attitude" is negligible. Since the main purpose of this example is just for illustrating the statistical methodology, the interpretation should not be taken too seriously.

4. Discussion

Maximum likelihood estimation of linear structural equation models is traditionally based on the analysis of covariance structure framework which utilizes asymptotic properties of the sample covariance matrix obtained from observed data. Clearly, this approach cannot be applied to the more complicated nonlinear structural equation models. In this paper, ML estimation of a general nonlinear model is developed on the basis of the following useful strategy: treat latent variables as missing data and solve the problem via the EM algorithm, realizing that the

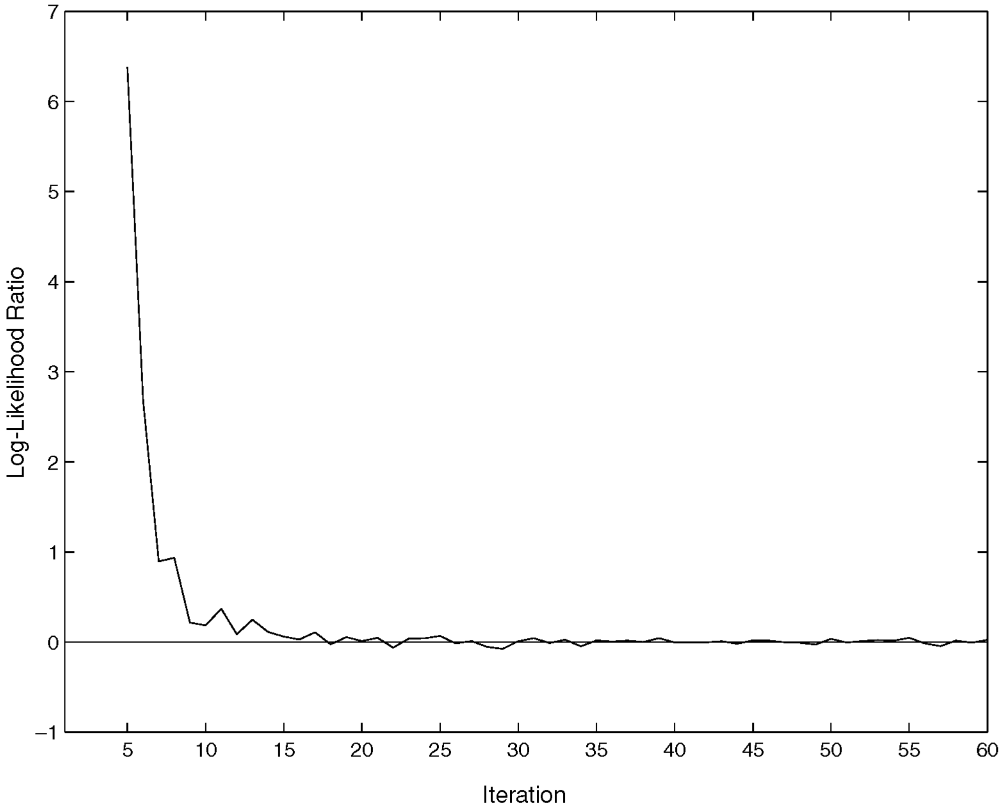

FIGURE 2.
ICPSR data with eight manifest variables: Log-likelihood ratio versus EM iteration from the 5th iteration.

TABLE 5.
ML estimates and their standard errors for the ICPSR data with eight manifest variables

Parameter	EST	SD	Parameter	EST	SD
λ_{21}	0.829	0.105	ψ_{11}	0.573	0.234
λ_{42}	2.057	0.263	ψ_{22}	1.404	0.245
λ_{63}	0.728	0.138	ψ_{33}	0.685	0.192
λ_{84}	0.975	0.155	ψ_{44}	2.495	0.872
			ψ_{55}	1.674	0.669
π	−0.008	0.187	ψ_{66}	4.188	0.512
			ψ_{77}	3.695	0.778
γ_{11}	0.010	0.152	ψ_{88}	2.405	0.949
γ_{12}	−0.295	0.148			
γ_{21}	0.366	0.084	ϕ_{11}	1.762	0.322
γ_{22}	0.523	0.095	ϕ_{12}	−0.200	0.205
γ_{23}	−0.178	0.055	ϕ_{22}	3.333	0.746
μ_1	8.419	0.126	$\psi_{\delta 1}$	3.468	0.355
μ_2	7.817	0.126	$\psi_{\delta 2}$	0.998	0.326
μ_3	2.351	0.118			
μ_4	5.526	0.241			
μ_5	7.545	0.176			
μ_6	7.368	0.180			
μ_7	5.044	0.270			
μ_8	3.683	0.251			

complete-data likelihood is less difficult to handle. The computational difficulty in the E-step is solved by the MH algorithm, while the M-step is completed by conditional maximization. Convergence of the algorithm is monitored by bridge sampling. Apparently, the proposed MCECM algorithm works pretty well for our problem.

In general, EM type algorithms, including MCECM and the Gibbs sampler, have tremendous impact on solving complicated statistical problems. These advanced computing tools have been the focus of much attention in recent literature, see for example Rubin (1991), Meng and van Dyk (1997), Wei and Tanner (1990) and references therein. In this paper, we give fundamental aspects in applying the MCECM algorithm of Wei and Tanner. The proposed algorithm works pretty well for our model; and yet its efficiency may be improved in certain directions. In the analysis of the real examples, we use a moderate M at the initial iterations and use a larger M near convergence of the algorithm. This choice of M is ad hoc and not greatly affect the speed of the algorithm since the number of iterations to achieve convergence is small. For large models with more complicated nonlinear causal effects, the number of iterations for convergence may be large. In this situation, a more subtle scheme in selecting M may improve the efficiency of the algorithm. See suggestions in Booth and Hobert (1999) in applying the EM type algorithm to generalized linear mixed models. The M-step of the algorithm is completed by several conditional maximizations, on the basis of the idea given in Meng and Rubin (1993). Since the solution can be achieved in closed form, the required computational burden is not heavy.

Another direction for improvement may be on monitoring convergence of the algorithm. Because the simulation variability introduced at its E step, monitoring convergence is not a trivial task. In this paper, we demonstrate the application of bridge sampling (Meng & Wong, 1996) with likelihood ratios. An interesting alternative is based on the final likelihood value that evaluated via path sampling (see Gelman & Meng, 1998). Comparisons on the feasibility and efficiency of bridge sampling and path sampling in analyzing complex structural equation models may be an interesting topic for further research.

Both ML estimate and Bayesian estimate are presented in the illustrative example. As we state in the Introduction, it is not our intention to claim that the proposed ML approach is better. However, it may be interesting to note the following differences between our ML procedure with the MCECM algorithm and the Bayesian procedure with the Gibbs sampler-MH algorithm in Arminger and Muthén (1998), and Lee and Zhu (2000):

1. The Bayesian approach treats structural parameters as random and requires to specify prior distributions of the parameters. This may be considered both good and bad. Good, because it allows to include useful prior information in the analysis. Bad, because prior distributions are hard to specify when there is no such information.
2. Computationally, in implementing the MCECM algorithm, latent random vectors are simulated in the E-step from the conditional distribution $p(\xi|y,\theta)$ via the MH algorithm, then structural parameters are updated in the M-step via conditional maximization. ML estimates are obtained at convergence of the MCECM algorithm. The Gibbs sampler-MH algorithm in the Bayesian procedure requires to simulate ξ from $p(\xi|y,\theta)$ via a similar MH algorithm; it also requires to simulate observations of other structural parameters from $p(\theta|y,\xi)$. After the convergence of the Gibbs sampler-MH algorithm, a sufficiently large sample of θ and ξ is further generated and Bayesian estimates are taken as the respective sampler means from the generated sample.
3. Convergence of the Gibbs sampler in the Bayesian approach is usually monitored via the 'estimated potential scale reduction' values of Gelman and Rubin (1992), which is quite different from using the bridge sampling.
4. An estimate of the covariance matrix of the Bayesian estimator can be obtained via the sampling covariance matrix from the generated sample of θ. Hence, it is easier to get standard errors than using the Louis (1982) formulae as in the MCECM algorithm.

Therefore, theoretically and computationally, ML procedure and Bayesian procedure are quite different. It is well-known that the two estimates are asymptotically equivalent, so both enjoy almost the same asymptotic properties. Generally, Bayesian estimates with good prior information are more accurate than ML estimates. In our opinion, it is more convenient to work with ML estimates in some statistical inference, such as computing the BIC in model selection (Lee & Song, 2001; Raftery, 1993). As for computational efficiency, according to our experience, the MCECM algorithm takes less computing time to produce parameters estimates. After taking the computation of the standard errors estimates into account, the Gibbs sampler-MH algorithm may require less computing time. Of course, this comparison is rough since the computing time depends on other factors such as the programming technique, the requirement of precision, etc. Both the MCECM algorithm and the Gibbs sampler-MH algorithm are not difficult to program.

In the analysis of the real example, there is discrepancy between the ad hoc approach and the proposed ML approach in estimating some important parameters. However, taking into account the fact that they are simple to implement with existing software, most ad hoc approaches are useful. Justification of the ad hoc estimates using more detailed simulation comparison with the Bayesian or ML estimate is valuable.

The distribution of the manifest random vector y is not necessarily normal; however, distributions of the latent random vector $\xi_{(2)}$ and error measurements ϵ and δ are assumed to be normal. Since the MCECM algorithm produces a sample of $\xi_{(2)}$, the normality assumption of $\xi_{(2)}$ can be assessed by investigation of this generated sample via standard method in data analysis. The normality assumption of ϵ and δ can be assessed via the following estimates of residuals:

$$\hat{\epsilon}_i = y_i - \hat{\mu} - \hat{\Lambda}\hat{\xi}_i, \quad i = 1, \ldots, n$$
$$\hat{\delta}_i = \hat{\xi}_{i(1)} - \Pi\hat{\xi}_{i(1)} - \Gamma H(\hat{\xi}_{i(2)}), \quad i = 1, \ldots, n$$

where $\hat{\xi}_i = (\hat{\xi}_{i(1)}, \hat{\xi}_{i(2)})$ is the estimate of ξ_i obtained via (14). Investigation of the robustness to normality assumptions in the light of Anderson (1989) and Browne (1987) represents an important but difficult open problem for further research. At present, great caution should be taken in interpreting results in the absence of the normality assumptions.

Another assumption of the proposed model is that $|\Pi_0|$ is a nonzero constant independent of Π. This assumption is taken for the simplicity and efficiency in completing the M-step of the MCECM algorithm. We do not claim that the class of models with a non-constant $|\Pi_0|$ is only slightly more general than our proposed model; however, we wish to emphasize that all NSEQ models in the field so far are developed with the endogenous latent variables in $\xi_{(1)}$ only regress on the exogenous latent variables in $\xi_{(2)}$ and hence $\Pi = 0$ and $|\Pi_0|$ is a constant. Therefore, almost all NSEQ models in the literature are special cases of our proposed model.

Now, let us give more detailed discussion on the ML analysis of the proposed NSEQ under the situation that $|\Pi_0|$ is a function of Π. The complete-data log-likelihood $L_c(\mathbf{X}|\theta)$ is exactly the same as in (4). Since the term $-2n \log |\Pi_0|$ involves no missing random data, the computation of the conditional expectations with respect to $\mathbf{Z}$ given $\mathbf{Y}$ and $\theta^{(r)}$ is not affected. As a result, the E-step of the MCECM algorithm is exactly the same as before. The M-step requires the solution of equation (9). All components in the first derivative $\partial L_c(\mathbf{X}|\theta)/\partial\theta$, except $\partial L_c(\mathbf{X}|\theta)/\partial\Lambda_{\xi j}$, given in (10) are not changed. Because $\Lambda_\xi = (\Pi, \Gamma)$, there are two parts in $\partial L_c(\mathbf{X}|\theta)/\partial\Lambda_{\xi j}$; they are $\partial L_c(\mathbf{X}|\theta)/\partial\Pi_j$ and $\partial L_c(\mathbf{X}|\theta)/\partial\Gamma_j$, where Π_j and Γ_j are the j-th rows of Π and Γ, respectively. The second part is

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial\Gamma_j} = \psi_{\delta j}^{-1} \sum_{i=1}^{n} [\xi_{ji(1)} - \Pi_j\xi_{i(1)} - \Gamma_j H(\xi_{i(2)})] H(\xi_{i(2)})^T,$$

which remains the same as before. However, since $|\Pi_0|$ is a function of Π, $\partial(-2n \log |\Pi_0|)/\partial\Pi$ is nonzero and the first part is not the same as before. Let $(I-\Pi)_j^{-T}$ be the j-th row of $(I-\Pi)^{-T}$, it can be shown that

$$\frac{\partial L_c(\mathbf{X}|\theta)}{\partial \Pi_j} = \psi_{\delta j}^{-1} \sum_{i=1}^{n} [\xi_{ji(1)} - \Pi_j \xi_{i(1)} - \Gamma_j H(\xi_{i(2)})] \xi_{i(1)}^T - n(I - \Pi)_j^{-T}, \tag{18}$$

where the first term on the right-hand side of (18) is the same as before and the second term is coming from $\partial(-2n \log |\Pi_0|)/\partial \Pi$. Although the other components are the same as before, owing to the additional term $-n(I - \Pi)_j^{-T}$ in $\partial L_c(\mathbf{X}|\theta)/\partial \Pi_j$, the method of conditional maximization (Meng & Rubin, 1993) cannot produce a closed form solution for the system of equations (9).

To complete the M-step under this situation, we may use some iterative procedures such as the Newton-Raphson algorithm or the scoring type algorithm to find the maximum of $Q(\theta|\theta^{(r)}) = E\{L_c(\mathbf{X}|\theta)|\mathbf{Y}, \theta^{(r)}\}$. Useful suggestions in Jamshidian and Jennrich (1993), and Lange (1995) on accelerating the EM algorithm can be adopted for improving the convergent rate of the algorithm. For instance, based on the argument in Lange (1995) that a single Newton-Raphson (or scoring) iteration at each M-step would be adequate to ensure convergence of an approximate EM algorithm, we just run a single iteration instead of finding the true maximum of $Q(\theta|\theta^{(r)})$. This method has been found to be effective in constrained ML estimation of two-level models, see Lee and Tsang (1999).

Another method to complete the M-step is the stochastic EM algorithm (Celeux & Diebolt, 1985). The key idea is to treat the model defined by (1) and (2) as simultaneous regression model given the estimates of $\xi_i's$ obtained at the E-step; that is

$$y_i = \mu + \Lambda \hat{\xi}_i + \epsilon_i, \qquad \hat{\xi}_{i(1)} = \Pi \hat{\xi}_{i(1)} + \Gamma H(\hat{\xi}_{i(2)}) + \delta_i,$$

with $\hat{\xi}_i's$ obtained at the E-step. Then the estimates of the unknown intercepts, regression coefficients, and covariance matrices of ξ_i, ϵ_i and δ_i are obtained by standard methods in simultaneous regression analysis. However, as pointed out by Marschner (2001) from a general theoretical point of view, this stochastic EM algorithm is equivalent to solving a biased estimating equation and may perform less favorably than the EM approach. Empirically, since the available data information for estimating ξ_i is essentially y_i, the estimate $\hat{\xi}_i$ is not accurate (see discussion and simulation results presented at the end of section 3. 1 and Table 2). Hence, when apply this algorithm to our model, the accuracy of the other parameters estimates is also affected. However, this stochastic EM algorithm has some appeal because of its simplicity.

Another approach is by reformulating the model via the following reduced form of the nonlinear structural equation defined by (2):

$$\xi_{(1)} = (I - \Pi)^{-1} \Gamma H(\xi_{(2)}) + (I - \Pi)^{-1} \delta. \tag{19}$$

Substituting (19) into (1), and using the notations $\Lambda = (\Lambda_1^T, \Lambda_2^T)^T$, $y_{(j)}$, $\mu_{(j)}$, and $\epsilon_{(j)}$, $j = 1, 2$ as in section 2.1, $\xi_{(2)}^* = ((I - \Pi)\Gamma H(\xi_{(2)})^T, \xi_{(2)}^T)^T$, and $\epsilon_{(1)}^* = \Lambda_1 (I - \Pi)^{-1} \delta + \epsilon_{(1)}$, we have

$$y_{(1)} = \mu_{(1)} + \Lambda_1 \xi_{(2)}^* + \epsilon_{(1)}^*,$$
$$y_{(2)} = \mu_{(2)} + \Lambda_2 \xi_{(2)}^* + \epsilon_{(2)}.$$

Note that $\epsilon_{(1)}^*$ and $\epsilon_{(2)}$ are uncorrelated and the covariance matrix of $\epsilon_{(1)}^*$, $\Psi_{\epsilon 1}^*$, is equal to $\Lambda_1 (I - \Pi)^{-1} \Psi_\delta (I - \Pi)^{-T} \Lambda_1^T + \Psi_{\epsilon 1}$, where $\Psi_{\epsilon j}$ is the covariance matrix of $\epsilon_{(j)}$, $j = 1, 2$. The joint density of $(y_i, \xi_{i(2)})$ is equal to

$$p(y_i, \xi_{i(2)}|\theta) = p(y_i|\xi_{i(2)}, \theta) p(\xi_{i(2)}|\theta) = (2\pi)^{-(p+q_2)/2} |\Psi_{\epsilon 1}^*|^{-n/2} |\Psi_{\epsilon 2}|^{-n/2} |\Phi|^{-n/2}$$
$$\times \exp\left[-\frac{1}{2}\{\xi_{i(2)}^T \Phi^{-1} \xi_{i(2)}\right.$$

$$+(y_{i(1)} - \mu_{(1)} - \Lambda_1\xi^*_{i(2)})^T \Psi^{*^{-1}}_{\epsilon 1} (y_{i(1)} - \mu_{(1)} - \Lambda_1\xi^*_{i(2)})$$
$$+(y_{i(2)} - \mu_{(2)} - \Lambda_2\xi^*_{i(2)})^T \Psi^{-1}_{\epsilon 2} (y_{i(2)} - \mu_{(2)} - \Lambda_2\xi^*_{i(2)})\}\Big]. \qquad (20)$$

The observed-data likelihood is equal to

$$L^*_o(\mathbf{Y}|\theta) = \prod_{i=1}^{n} \int p(y_i, \xi_{i(2)}|\theta)\, d\xi_{i(2)}, \qquad (21)$$

where the dimension of the integral is only q_2. Let $\mathbf{Z}_2 = (\xi_{1(2)}, \ldots, \xi_{n(2)})$ be the matrix of latent vectors, the complete-data log-likelihood of θ on the basis of $(\mathbf{Y}, \mathbf{Z}_2)$ is

$$L^*_c(\mathbf{Y}, \mathbf{Z}_2|\theta) = \log \prod_{i=1}^{n} p(y_i, \xi_{i(2)}|\theta) = -\frac{n}{2}\{(p + q_2)\log(2\pi)$$
$$+ \log|\Lambda_1(I - \Pi)^{-1}\Psi_\delta(I - \Pi)^{-T}\Lambda_1^T + \Psi_{\epsilon 1}| + \log|\Psi_{\epsilon 2}| + \log|\Phi|\}$$
$$-\frac{1}{2}\sum_{i=1}^{n}\big[\xi^T_{i(2)}\Phi^{-1}\xi_{i(2)} + (y_{i(2)} - \mu_{(2)} - \Lambda_2\xi^*_{i(2)})^T\Psi^{-1}_{\epsilon 2}(y_{i(2)} - \mu_{(2)} - \Lambda_2\xi^*_{i(2)})$$
$$+ (y_{i(1)} - \mu_{(1)} - \Lambda_1\xi^*_{i(2)})^T\{\Lambda_1(I - \Pi)^{-1}\Psi_\delta(I - \Pi)^{-T}\Lambda_1^T + \Psi_{\epsilon 1}\}^{-1}$$
$$\times (y_{i(1)} - \mu_{(1)} - \Lambda_1\xi^*_{i(2)})\big]. \qquad (22)$$

Let $Q^*(\theta|\theta^{(r)}) = E\{L^*_c(\mathbf{Y}, \mathbf{Z}_2|\theta)|\mathbf{Y}, \theta^{(r)}\}$. The E-step can be completed by a similar MH algorithm which simulates $\xi_{i(2)}$ from $p(\xi_{i(2)}|y_i, \theta)$. Since the dimension of $\xi_{i(2)}$ is only q_2, simulating $\xi_{i(2)}$ may require less computing time than simulating the whole ξ_i whose dimension is $q_1 + q_2$. Hence, the E-step can be completed efficiently. In the M-step, we require to solve

$$\frac{\partial Q^*(\theta|\theta^{(r)})}{\partial\theta} = E\{\frac{\partial}{\partial\theta}L^*_c(\mathbf{Y}, \mathbf{Z}_2|\theta)|\mathbf{Y}, \theta^{(r)}\} = 0. \qquad (23)$$

Since $L^*_c(\mathbf{Y}, \mathbf{Z}_2|\theta)$ involves the term $-2^{-1}n\log|\Lambda_1(I - \Pi)^{-1}\Psi_\delta(I - \Pi)^{-T}\Lambda_1^T + \Psi_{\epsilon 1}|$, its derivatives with respect to Λ_1, Π, Ψ_δ, and $\Psi_{\epsilon 1}$ are more complicated than the derivative of $L_c(\mathbf{X}|\theta)$ with respect to Π. Hence, the system of equation in (23) is more difficult to solve and the M-step cannot be completed by the conditional maximization in closed form. The Newton-Raphson algorithm, or the stochastic EM algorithm has to be implemented for completing the M-step.

References

Anderson, T.W. (1989). Linear latent variable models and covariance structures. *Journal of Econometrics, 41*, 91–119.

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.

Bagozzi, R.P., Baumgartner, H., & Yi, Y. (1992). State versus action orientation and the theory of reasoned action: An application to coupon usage. *Journal of Consumer Research, 18*, 505–517.

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.

Bentler, P.M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47*, 541–570.

Berger, J.O., & Perrichi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association, 91*, 109–122.

Bollen, K.A., & Paxton, P. (1998). Two-stage least squares estimation of interaction effects. In R.E. Schumacker & G.A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation models* (pp. 125–151). Mahwah, NJ: Lawrence Erlbaum Associates.

Booth, J.G., & Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B, 61*, 265–285.

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

Browne, M.W. (1987). Robustness of statistical inference in factor analysis and related models. *Biometrika, 74*, 375–384.

Busemeyer, J.R., & Jones, L.E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93*, 549–562.

Celeux, G., & Diebolt, J. (1989). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly, 2*, 73–82.

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.

Etezadi-Amoli, J., & McDonald, R.P. (1983). A second generation nonlinear factor analysis. *Psychometrika, 48*, 315–342.

Fraser, C. (1980). *COSAN user's guide*. Toronto, Canada: The Ontario Institute for Studies in Education.

Gelman, A., & Meng, X.L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science, 13*, 163–185.

Gelman, A., Roberts, G.O., & Gilks, W.R. (1995). Efficient Metropolis humping rules. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 5* (pp. 599–607). Oxford, England: Oxford University Press.

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.

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

Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted. *Psychological Bulletin, 112*, 351–362.

Jaccard, J., & Wan, C.K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin, 117*, 348–357.

Jamshidian, M., & Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association, 88*, 221–228.

Jonsson, F.Y. (1998). Modeling interaction and non-linear effects: A step by step LISREL example. In R.E. Schumacker & G.A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation models* (pp. 17–42). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Jöreskog, K.G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G.A. Marcoulides & R.E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 57–88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kass, R.E., & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773–795.

Kenny, D.A., & Judd, C.M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*, 201–210.

Klein, A., Moosbrugger, H., Schermelleh-Engel, K., & Frandk, D (1997). A new approach to the estimation of latent interaction effects in structural equation models. In W. Bandilla & F. Fanlbaum (Eds.), *SOFTSTAT '97—Advances in statistical software* (pp. 479–488). Stuttgart, Germany: Lucius & Lucius.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Association, Series B, 57*, 425–437.

Lee, S.Y., Poon, W.Y., & Bentler, P.M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology, 48*, 339–358.

Lee, S.Y., & Song, X.Y. (2001). Hypothesis testing and model comparison in two-level structural equation models. *Multivariate Behavioral Research, 36*, 639–655.

Lee, S.Y., & Tsang, S.Y. (1999). Constrained maximum likelihood estimation of two-level covariance structure model via EM type algorithms. *Psychometrika, 64*, 435–450.

Lee, S.Y., & Zhu, H.T. (2000). Statistical analysis of nonlinear structural equation model with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 53*, 209–232.

Li, F., Harmer, P., Duncan, T.E., Duncan, S.C., Acock, A., & Boles, S. (1998). Approaches to testing interaction effects using structural equation modeling methodology. *Multivariate Behavioral Research, 33*, 1–39.

Liu, C., & Rubin, D.B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica, 8*, 729–747.

Liu, J.S., Liang, F.M., & Wong, W.H. (2000). The use of multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association, 95*, 121–134.

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

Marschner, I.C. (2001). On stochastic version of the EM algorithm. *Biometrika, 88*, 281–286.

McDonald, R.P. (1962). A general approach to nonlinear factor analysis. *Psychometrika, 27*, 123–157.

McDonald, R.P. (1967a). Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika, 32*, 77–112.

McDonald, R.P. (1967b). Factor interaction in nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology, 20*, 205–215.

Meng, X.L., & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*, 267–278.

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

Meng, X.L., & van Dyk, D. (1997). The EM algorithm—An old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B, 59,* 511–567.

Meng, X.L., & Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistic Sinica, 6,* 831–860.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics, 21,* 1087–1091.

Mooijaart, A., & Bentler, P. (1986). Random polynomial factor analysis. In E. Diday, M. Jambu, L. Lebart, J. Pages, & R. Tomassone (Eds.), *Data analysis and informatics, IV* (pp. 241–250). North-Holland: Elsevier Science Publishers.

Ping, R.A. (1996a). Interaction and quadratic effect estimation: A two step technique using structural equation analysis. *Psychological Bulletin, 119,* 166–175.

Ping, R.A. (1996b). Latent variable regression: A technique for estimating interaction and quadratic coefficients. *Multivariate Behavioral Research, 31,* 95–120.

Ping, R.A. (1996c). Estimating latent variable interactions and quadratics: The state of this art. *Journal of Management, 22,* 163–183.

Raftery, A.E. (1993). Bayesian model selection in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 163–180). Beverly Hills, CA: Sage.

Roberts, G.O. (1996). Markov Chain concepts related to sampling algorithms. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 45–57). London, England: Chapman and Hall.

Rubin, D.B. (1991). EM and beyond. *Psychometrika, 56,* 241–254.

Rubin, D.B., & Thayer, D.T. (1982). EM algorithm for ML factor analysis. *Psychometrika, 47,* 69–76.

Schumacker, R.E., & Marcoulides, G.A. (Eds.). (1998). *Interaction and nonlinear effects in structural equation models.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Wei, G.C.G., & Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor man's data augmentation algorithm. *Journal of the American Statistical Association, 85,* 699–704.

World Values Survey: 1981–1984 & 1990–1993. (1994). Ann Arbor, MI: Inter-University Consortium of Political and Social Research. (For the I.C.P.S.R. version the Institute for Social Research is the producer, and the Inter-University Consortium of Political and Social Research is the distributor.)

Manuscript received 25 JUN 1999
Final version received 15 MAY 2001