Statistica Sinica **8**(1998), 749-766

ANALYSIS OF TWO-LEVEL STRUCTURAL EQUATION MODELS VIA *EM* TYPE ALGORITHMS

Sik-Yum Lee and Wai-Yin Poon

Chinese University of Hong Kong

Abstract: In this paper, the maximum likelihood estimation of a general two-level structural equation model with an unbalanced design is formulated as a missing data problem by treating the latent random vectors at the group level as hypothetical missing data. The commonly used EM algorithm is utilized to obtain the solution. Expressions for the E-step are derived and it is shown that the complex optimization of the M-step can be completed conveniently with existing software. Some accelerated procedures such as the EM gradient algorithm and the Quasi-Newton EM algorithm are modified to improve the convergence rate of the basic EM algorithm. Results from simulation studies and analysis of examples illustrate the features and potential of the EM approach.

Key words and phrases: EM gradient algorithm, EQS, information matrix, latent random vectors, LISREL, maximum likelihood.

1. Introduction

Formulation of structural equation models is an important technique used in studying correlations and causations among observed and latent variables. In past years, based on the standard assumption that the observational data are independent, models have been applied widely in educational research. However, there are still many fundamental researches in educational testing which require the analysis of multilevel data from various types of hierarchical sampling designs. For example, consider the problem of deciding correct educational policy on programs that involve judgement of the performance of schools, programs and students; investigators are required to assess students' performance and how that performance is influenced by factors through the activities of teachers in classrooms and effectiveness of school origanization. To analyze these influences based on the non-independent multilevel data requires the statistical modeling of the causations and correlations at each of these levels. In past years, significant contributions have been developed to deal with multilevel data in educational testing (see for example Goldstein (1987) and Bock (1989)). Moreover, a number of authors have also established some theoretical results for analysis of structural equation models with multilevel data. For example, McDonald and Goldstein (1989) analyzed the two-level model with balanced sampling designs, and Lee

(1990) developed the generalized least squares and the maximum likelihood (ML) theory for more general models with unbalanced designs. In addition to the theoretical contributions, certain computational methods have also been proposed recently. Longford and Muthen (1992) derived procedures for applying the scoring algorithm in the context of the factor analysis model. It is noted that even in this simple special case, expressions for implementing the scoring algorithm are very complicated and tedious to program. Clearly, the situation is worse when the scoring algorithm is applied to the general models. Muthen (1990) showed that the solution established by McDonald and Goldstein (1989) with balanced data can be obtained using available softwares EQS (Bentler (1992)) and LISREL (Jöreskog and Sörborm (1996)). However, this method cannot be applied to models with unbalanced designs. Raudenbush (1995), considering the observed unbalanced data as "incomplete", and the "complete data" as balanced, proposed using the EM algorithm to obtain the solution of this missing-data problem. Expressions for the E-step were derived and it was suggested that the more complex computation of the M-step be performed by Muthen's (1990) method. As pointed out by Raudenbush (1995), a likely disadvantage of his approach is the slow convergence that occurs when sample sizes vary substantially. Lee and Poon (1992) showed that the "multi-sample" option of LISREL (Jöreskog and Sörbom (1996)) and EQS (Bentler (1992)) can be used to obtain a consistent estimator of a special two-level model in which the covariance structures of the individual levels are invariant across groups. The method cannot be applied to models with different group structures.

In this paper, we investigate the application of the EM algorithm to obtain the ML solution of the two-level general structural equation models by treating the latent random vectors at the group level as missing. Hence, the approach is quite different from Raudenbush (1995) but is similar to the procedure suggested by Rubin and Thayer (1982) for the ML factor analysis where the latent factor scores are treated as missing data. We show that the proposed algorithm has the following features that are better in one or more aspects than the procedures cited in the previous paragraph: (i) general structural equation models with unbalanced designs and different within group covariance structures can be analyzed, (ii) ML solution can be obtained conveniently with the standard LISREL (Jöreskog and Sörbom (1996)) or EQS (Bentler (1992)) software packages, so it is easy to apply in practice, (iii) the convergence is fast and expressions for implementation are simple; hence the computational burden of the algorithm is light.

Organization of the rest of the paper is as follows. The model and the ML estimation are discussed in Section 2. The motivation of the EM algorithm is presented in Section 3. Expressions for the E-step are derived and procedures for computing the M-step are discussed. Moreover, some modifications in accelerating the EM algorithm are presented. Section 4 includes some examples and results of simulation studies which give some evidence about the empirical performance of the proposed algorithms. Examples based on models with cross level parameters are included. Finally, the paper concludes with a discussion of the algorithm.

2. ML Estimation of Two-level Structural Equation Model

Suppose $\underset{\sim}{x}_{gi}$ is a $p \times 1$ observed random vector such that

$$\underset{\sim}{x}_{gi} = \underset{\sim}{v}_g + \underset{\sim}{v}_{gi}, \tag{1}$$

for $g = 1, \ldots, G,\ i = 1, \ldots, N_g$, where $\underset{\sim}{v}_g$ is a latent random vector varying at the group level, and $\underset{\sim}{v}_{gi}$ is a latent random vector varying at the individual level. It is assumed that the random vectors $\{\underset{\sim}{v}_g,\ g = 1, \ldots, G\}$ are i.i.d; and for a given g, $\{\underset{\sim}{v}_{gi},\ i = 1, \ldots, N_g\}$ are i.i.d; and that $\underset{\sim}{v}_g$ and $\underset{\sim}{v}_{gi}$ are also independent. Suppose $\underset{\sim}{v}_g$ is distributed as $N[\underset{\sim}{0}, \underset{\sim}{\Sigma}_B^*]$ and $\underset{\sim}{v}_{gi}$ is distributed as $N[\underset{\sim}{0}, \underset{\sim}{\Sigma}_{gW}^*]$, where the between group covariance structure $\underset{\sim}{\Sigma}_B^* = \underset{\sim}{\Sigma}_B(\underset{\sim}{\theta}^*)$ and the within group covariance structure $\underset{\sim}{\Sigma}_{gW}^* = \underset{\sim}{\Sigma}_{gW}(\underset{\sim}{\theta}^*)$ are matrix functions of an unknown $q \times 1$ parameter vector $\underset{\sim}{\theta}^*$. The N_g may be different, so we are dealing with models with unbalanced designs. Without loss of generality, it is assumed that the mean vectors are equal to zero. It should be noted that $\underset{\sim}{\Sigma}_B^*$ and $\underset{\sim}{\Sigma}_{gW}^*$ may have general structure where cross-level parameters are allowed. Examples of $\underset{\sim}{\Sigma}_B^*$ and $\underset{\sim}{\Sigma}_{gW}^*$ are the factor analysis model, the LISREL model (Jöreskog and Sörbom (1996)), and the Bentler and Week's (1980) model in EQS.

Let $\underset{\sim}{z}_g' = (\underset{\sim}{x}_{g1}', \ldots, \underset{\sim}{x}_{gN_g}')$, the distribution of $\underset{\sim}{z}_g$ is $N[\underset{\sim}{0}, (\underset{\sim}{J}_g \otimes \underset{\sim}{\Sigma}_B^*) + (\underset{\sim}{I}_g \otimes \underset{\sim}{\Sigma}_{gW}^*)]$, where $\underset{\sim}{J}_g$ is an $N_g \times N_g$ square matrix of unit elements, and $\underset{\sim}{I}_g$ is the identity matrix of order N_g. From the results in Lee (1990), it can be shown that the negative log-likelihood function based on the observed data $\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_G$ is proportional to

$$\begin{aligned} F(\underset{\sim}{\theta}^*) = \sum_{g=1}^{G} \Big\{ & \log|\underset{\sim}{\Sigma}_g^*| + N_g^{-1}\ \mathrm{tr}[\underset{\sim}{\Sigma}_g^{*^{-1}} \sum_{i,j} \underset{\sim}{x}_{gi}\underset{\sim}{x}_{gj}'] \\ & + (N_g - 1)\log|\underset{\sim}{\Sigma}_{gW}^*| + N_g^{-1}\ \mathrm{tr}[\underset{\sim}{\Sigma}_{gW}^{*^{-1}} \sum_{i \neq j} (\underset{\sim}{x}_{gi}\underset{\sim}{x}_{gi}' - \underset{\sim}{x}_{gi}\underset{\sim}{x}_{gj}')] \Big\}, \end{aligned} \tag{2}$$

where $\underset{\sim}{\Sigma}_g^* = \underset{\sim}{\Sigma}_{gW}^* + N_g\underset{\sim}{\Sigma}_B^*$. Direct minimization of (2) to obtain the ML estimate of $\underset{\sim}{\theta}^*$ is very tedious even for the case where $\underset{\sim}{\Sigma}_B^*$ and $\underset{\sim}{\Sigma}_{gW}^*$ have the simple factor analysis structures (Longford and Muthen (1992)). In the next section, it will be shown that the ML estimate can be obtained conveniently with much less effort by using the EM algorithm with either EQS (Bentler (1992)) or LISREL (Jöreskog and Sörbom (1996)).

3. Estimation of the Model Using the EM Algorithm

From the definition of the two-level model given in (1), it is clear that if $\underset{\sim}{v}_g$ is observed, the model will become rather simple and can be analyzed without much difficulty. Thus, we consider $\{(\underset{\sim}{x}_{g1}, \ldots, \underset{\sim}{x}_{gN_g}, \underset{\sim}{v}_g), g = 1, \ldots, G\}$ as the complete data set, and treat the random vectors $\underset{\sim}{v}_g$ as missing. Hence, this model can be formulated as a missing data problem and the EM algorithm (Dempster et al. (1977)) is a natural procedure to obtain the solution. The idea presented here has been proposed by Rubin (1991) and is similar to that given by Rubin and Thayer (1982), where they obtained the ML factor analysis solution via the EM algorithm by treating the latent factor scores as missing data.

Let $\underset{\sim}{X}$ and $\underset{\sim}{V}$ denote the observed data and the missing data with elements given by the $\underset{\sim}{x}_{gi}$'s, and the $\underset{\sim}{v}_g$'s, respectively. The negative log- likelihood function of the complete data set is proportional to

$$\begin{aligned} L(\underset{\sim}{X}, \underset{\sim}{V} | \underset{\sim}{\theta}^*) = & \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N_g} [\log |\underset{\sim}{\Sigma}^*_{gW}| + (\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)' \underset{\sim}{\Sigma}^{*-1}_{gW} (\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)] \\ & + \frac{1}{2} \sum_{g=1}^{G} [\log |\underset{\sim}{\Sigma}^*_B| + \underset{\sim}{v}'_g \underset{\sim}{\Sigma}^{*^{-1}}_B \underset{\sim}{v}_g], \end{aligned} \tag{3}$$

where $\underset{\sim}{\Sigma}^*_{gW}$ and $\underset{\sim}{\Sigma}^*_B$ are the within and between group covariance structures that depend on the unknown parameter vector $\underset{\sim}{\theta}^*$.

In the E-step of the EM algorithm, we need to find $E[L(\underset{\sim}{X}, \underset{\sim}{V} | \underset{\sim}{\theta}^*) | \underset{\sim}{X}, \underset{\sim}{\theta}]$, that is, the expected value of the complete-data negative log-likelihood given the observed data $\underset{\sim}{X}$ and the current value of the parameter $\underset{\sim}{\theta}$. The second step of the EM algorithm, the M-step, requires minimizing this expected negative log-likelihood with respect to $\underset{\sim}{\theta}^*$, as if it were based on the complete data. Hence, the M-step gives the next value of $\underset{\sim}{\theta}$; using the new $\underset{\sim}{\theta}$, the E-step is computed and the algorithm continues.

3.1. The E-Step

It can be seen from (3) that finding $E[L(\underset{\sim}{X}, \underset{\sim}{V} | \underset{\sim}{\theta}^*) | \underset{\sim}{X}, \underset{\sim}{\theta}]$ requires one to find the

$$E\Big[\sum_{i=1}^{N_g} (\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)' \underset{\sim}{\Sigma}^{*-1}_{gW} (\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g) | \underset{\sim}{X}, \underset{\sim}{\theta})\Big] = \mathrm{tr} \underset{\sim}{\Sigma}^{*-1}_{gW} \sum_{i=1}^{N_g} E\Big[(\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)(\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)' | \underset{\sim}{X}, \underset{\sim}{\theta})\Big], \tag{4}$$

and

$$E\Big[\underset{\sim}{v}'_g \underset{\sim}{\Sigma}^{*^{-1}}_B \underset{\sim}{v}_g | \underset{\sim}{X}, \underset{\sim}{\theta}\Big] = \mathrm{tr} \underset{\sim}{\Sigma}^{*^{-1}}_B \; E\Big[\underset{\sim}{v}_g \underset{\sim}{v}'_g | \underset{\sim}{X}, \underset{\sim}{\theta}\Big]. \tag{5}$$

Given the current parameter vector $\underset{\sim}{\theta}$, the joint distribution of $(\underset{\sim}{v}_g^{'}, \underset{\sim}{x}_{g1}', \ldots, \underset{\sim}{x}_{gN_g}')'$ is $N[\underset{\sim}{0}, \underset{\sim}{\Omega}_g]$, where

$$\underset{\sim}{\Omega}_g = \begin{bmatrix} \underset{\sim}{\Omega}_{g11} & \underset{\sim}{\Omega}_{g12} \\ \underset{\sim}{\Omega}_{g21} & \underset{\sim}{\Omega}_{g22} \end{bmatrix},$$

with $\underset{\sim}{\Omega}_{g11} = \underset{\sim}{\Sigma}_B = \underset{\sim}{\Sigma}_B(\underset{\sim}{\theta})$, $\underset{\sim}{\Omega}_{g12} = (\underset{\sim}{\Sigma}_B, \ldots, \underset{\sim}{\Sigma}_B)$, $\underset{\sim}{\Omega}_{g22} = (\underset{\sim}{J}_g \otimes \underset{\sim}{\Sigma}_B) + (\underset{\sim}{I}_g \otimes \underset{\sim}{\Sigma}_{gW})$, and $\underset{\sim}{\Sigma}_{gW} = \underset{\sim}{\Sigma}_{gW}(\underset{\sim}{\theta})$. Letting $\underset{\sim}{\Sigma}_g = \underset{\sim}{\Sigma}_{gW} + N_g \underset{\sim}{\Sigma}_B$, it can be shown that

$$\underset{\sim}{\Omega}_{g22}^{-1} = (\underset{\sim}{I}_g \otimes \underset{\sim}{\Sigma}_{gW}^{-1}) - [\underset{\sim}{J}_g \otimes N_g^{-1}(\underset{\sim}{\Sigma}_{gW}^{-1} - \underset{\sim}{\Sigma}_g^{-1})]. \tag{6}$$

It follows from (6) that $\underset{\sim}{\Omega}_{g12}\underset{\sim}{\Omega}_{g22}^{-1} = (\underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}, \ldots, \underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1})$ and $\underset{\sim}{\Omega}_{g11} - \underset{\sim}{\Omega}_{g12}\underset{\sim}{\Omega}_{g22}^{-1}\underset{\sim}{\Omega}_{g21} = \underset{\sim}{\Sigma}_B - N_g\underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{\Sigma}_B$. Hence, the conditional distribution of $\underset{\sim}{v}_g$ given $(\underset{\sim}{x}_{g1}, \ldots, \underset{\sim}{x}_{gN_g})$ is multivariate normal with covariance matrix $\underset{\sim}{\Sigma}_B - N_g\underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{\Sigma}_B$ and mean vector $\underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{t}_g$, where $\underset{\sim}{t}_g = \Sigma_{i=1}^{N_g}\ \underset{\sim}{x}_{gi}$. Based on this result, the conditional expectation of $(\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)(\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)^{'}$ and $\underset{\sim}{v}_g\underset{\sim}{v}_g^{'}$ given $(\underset{\sim}{X}, \underset{\sim}{\theta})$ can be obtained. Hence, it can be shown that

$$\sum_{i=1}^{N_g} E[(\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)(\underset{\sim}{x}_{gi} - \underset{\sim}{v}_g)^{'}|\underset{\sim}{X}, \underset{\sim}{\theta})] = N_g\underset{\sim}{C}_g(\theta), \tag{7}$$

where

$$\underset{\sim}{C}_g(\underset{\sim}{\theta}) = N_g^{-1}\Big[\sum_{i=1}^{N_g}\underset{\sim}{x}_{gi}\ \underset{\sim}{x}_{gi}' - \underset{\sim}{t}_g\underset{\sim}{t}_g'\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{\Sigma}_B - \underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{t}_g\underset{\sim}{t}_g'\Big] + \underset{\sim}{D}_g(\theta),$$

with

$$\underset{\sim}{D}_g(\theta) = E[(\underset{\sim}{v}_g\underset{\sim}{v}_g^{'})|\underset{\sim}{X}, \underset{\sim}{\theta})] = (\underset{\sim}{\Sigma}_B - N_g\underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{\Sigma}_B) + \underset{\sim}{\Sigma}_B\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{t}_g\underset{\sim}{t}_g'\underset{\sim}{\Sigma}_g^{-1}\underset{\sim}{\Sigma}_B.$$

It should be noted that with given $\underset{\sim}{X}$ and $\underset{\sim}{\theta}, \underset{\sim}{C}_g(\theta)$ and $\underset{\sim}{D}_g(\theta)$ are known; it follows from (4) and (7) that

$$\begin{aligned} &E[L(\underset{\sim}{X}, \underset{\sim}{V}|\underset{\sim}{\theta}^*)|\underset{\sim}{X}, \underset{\sim}{\theta}] = M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}) \\ =&\sum_{g=1}^{G}\frac{N_g}{2}[\log|\underset{\sim}{\Sigma}_{gW}^*| + \text{tr}\ \underset{\sim}{\Sigma}_{gW}^{*-1}\underset{\sim}{C}_g(\underset{\sim}{\theta})] + \frac{G}{2}[\log|\underset{\sim}{\Sigma}_B^*| + \text{tr}\ \underset{\sim}{\Sigma}_B^{*-1}\underset{\sim}{C}_B(\underset{\sim}{\theta})], \end{aligned}$$

where $\underset{\sim}{C}_B(\theta) = G^{-1}[\underset{\sim}{D}_1(\underset{\sim}{\theta}) + \cdots + \underset{\sim}{D}_G(\underset{\sim}{\theta})]$, which is also a known matrix with the current vector $\underset{\sim}{\theta}$. This completes the E-step of the algorithm.

3.2. The M-Step

The function to be minimized at this step can be expressed as

$$M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}) = \sum_{k=1}^{G+1}\ 2^{-1}N_k\ [\log|\underset{\sim}{\Sigma}_k^*| + \text{tr}\underset{\sim}{\Sigma}_k^{*^{-1}}\underset{\sim}{C}_k(\underset{\sim}{\theta})], \tag{8}$$

where for $k \leq G,\ N_k = N_g,\ \underset{\sim}{\Sigma}_k^* = \underset{\sim}{\Sigma}_{gW}(\underset{\sim}{\theta}^*), \underset{\sim}{C}_k(\underset{\sim}{\theta}) = \underset{\sim}{C}_g(\underset{\sim}{\theta})$; while for $k = G + 1, N_k = G, \underset{\sim}{\Sigma}_k^* = \underset{\sim}{\Sigma}_B(\underset{\sim}{\theta}^*)$ and $\underset{\sim}{C}_k(\underset{\sim}{\theta}) = \underset{\sim}{C}_B(\underset{\sim}{\theta})$. It should be noted that there is no closed form solution for this minimization problem and some iterative procedure is required. However, $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$ is a simple function which has exactly the same form as the ML fit function in structural equation modelling for multiple groups with the known matrices $\underset{\sim}{C}_k(\underset{\sim}{\theta})$ playing the role of the sample covariance matrices. If there are no cross-level parameters and each within group structure has no common parameters, that is, $\underset{\sim}{\Sigma}_B^* = \underset{\sim}{\Sigma}_B(\underset{\sim}{\theta}_B^*),\ \underset{\sim}{\Sigma}_{gW}^* = \underset{\sim}{\Sigma}_{gW}(\underset{\sim}{\theta}_g^*),\ g = 1, \ldots, G$, and $\{\underset{\sim}{\theta}_B^*, \underset{\sim}{\theta}_1^*, \ldots, \underset{\sim}{\theta}_G^*\}$ is a set of distinct parameter vectors, then $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$ is just a sum of $G+1$ independent functions that depend on some separable parameters. Hence, the minimum of $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$ can be obtained easily by separate minimization of $G+1$ small and simple functions, each of them just involving one $\underset{\sim}{\Sigma}_B^*$ or $\underset{\sim}{\Sigma}_{gW}^*$. This minimization can be completed conveniently using the standard "multiple-sample" option of LISREL or EQS. If there are some cross-level parameters and/or common parameters in certain within-group structures, the minimum of $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$ can also be obtained easily and conveniently by the "multiple-sample" option of LISREL or EQS with appropriate equality constraints on the corresponding elements of the parameter vector $\underset{\sim}{\theta}^*$.

It should be noted that it is much more complicated to minimize the function $F(\underset{\sim}{\theta}^*)$ than $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$. From (2), it can be seen that each term of $F(\underset{\sim}{\theta}^*)$ depends on $\underset{\sim}{\Sigma}_B^*$ through $\underset{\sim}{\Sigma}_g^*$; hence, even though $\underset{\sim}{\Sigma}_{gW}^*$ and $\underset{\sim}{\Sigma}_B^*$ contain different or common parameters, the minimization of $F(\underset{\sim}{\theta}^*)$ has to be carried out with respect to all the parameters simultaneously. As a result, the computational burden in direct minimization of $F(\underset{\sim}{\theta}^*)$ via the scoring or other algorithms is much more complex than the minimization of $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$ using the EM algorithm.

3.3. Acceleration of the EM algorithm

In general, it is well-known that the convergence of the EM algorithm may be slow in certain practical applications (Dempster et al. (1977)). Recently, a number of suggestions had been proposed to accelerate the algorithm. For instance, Jamshidian and Jennrich (1993) advocated a conjugate gradient version of the EM algorithm; Lange (1995a,b) respectively recommended an EM gradient algorithm and a Quasi-Newton acceleration of the EM algorithm; and Liu and Rubin (1994) proposed an ECME algorithm. In our situation, all the above mentioned modifications can be adapted. To save space, we only briefly discuss how to utilize Lange's (1995a,b) approaches in accelerating the EM algorithm.

Based on the argument that a single Newton-Raphson iteration at each M-step would be adequate to ensure convergence of an approximate EM algorithm, the EM gradient algorithm proposed by Lange (1995a) is to update the current parameter vector $\underset{\sim}{\theta}^*$ at the ith iteration by

$$\underset{\sim}{\theta}_{i+1} = \underset{\sim}{\theta}_i - d^{20}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)^{-1}\ d^{10}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)\Big|_{\underset{\sim}{\theta}^*=\underset{\sim}{\theta}_i}, \tag{9}$$

where d^{10} and d^{20} are respectively the first and second partial derivatives with respect to $\underset{\sim}{\theta}^*$ in $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)$. Hence, this algorithm avoids the search for the exact optimum in the M-step of the EM algorithm while preserving the local convergence properties of the EM algorithm. The Quasi-Newton acceleration (Lange (1995b)) is to replace (9) by

$$\underset{\sim}{\theta}_{i+1} = \underset{\sim}{\theta}_i - [d^{20}\ M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i) + \underset{\sim}{B}_i]^{-1}\ d^{10}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)\Big|_{\underset{\sim}{\theta}^*=\underset{\sim}{\theta}_i}, \tag{10}$$

in which $\underset{\sim}{B}_i$ is the current approximation to the missing Hessian matrix; and is updated as described in Lange (1995b). By differentiating the function $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)$ in (8), it can be shown that

$$d^{10}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i) = \sum_{k=1}^{G+1}\frac{N_k}{2}\Big\{\underset{\sim}{\triangle}_k(\underset{\sim}{\Sigma}_k^* \otimes \underset{\sim}{\Sigma}_k^*)^{-1}\mathrm{vec}[\underset{\sim}{C}_k(\underset{\sim}{\theta}_i) - \underset{\sim}{\Sigma}_k^*]\Big\}, \tag{11}$$

$$\begin{aligned} d^{20}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i) = \sum_{k=1}^{G+1}\frac{N_k}{2}\Big\{&\underset{\sim}{\triangle}_k[(\underset{\sim}{\Sigma}_k^*\otimes\underset{\sim}{\Sigma}_k^*)^{-1}+2[\underset{\sim}{\Sigma}_k^{*^{-1}}\otimes\underset{\sim}{\Sigma}_k^{*^{-1}}(\underset{\sim}{C}_k(\underset{\sim}{\theta}_i)-\underset{\sim}{\Sigma}_k^*)\underset{\sim}{\Sigma}_k^{*^{-1}}]]\underset{\sim}{\triangle}_k' \\ &+\underset{\sim}{\nabla}_k[\underset{\sim}{I}_p\otimes(\underset{\sim}{\Sigma}_k^*\otimes\underset{\sim}{\Sigma}_k^*)^{-1}]\mathrm{vec}[\underset{\sim}{C}_k(\underset{\sim}{\theta}_i)-\underset{\sim}{\Sigma}_k^*]\Big\}, \end{aligned} \tag{12}$$

where $\underset{\sim}{\triangle}_k = \partial\underset{\sim}{\Sigma}_k(\underset{\sim}{\theta}^*)/\partial\underset{\sim}{\theta}^*, \underset{\sim}{\nabla}_k = \partial^2\underset{\sim}{\Sigma}_k(\underset{\sim}{\theta}^*)/\partial\underset{\sim}{\theta}^*\partial\underset{\sim}{\theta}^*$, $\underset{\sim}{I}_p$ is a $p \times p$ identity matrix and $\mathrm{vec}(\underset{\sim}{A})$ is a vector that stores elements of $\underset{\sim}{A}'$ columnwise sequentially. In the above EM gradient algorithm and the Quasi-Newton acceleration, it is important that $d^{20}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)$ be a positive definite matrix. However, it can be seen from (12) that $d^{20}M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta}_i)$ may not be positive definite in general. Because $E(\underset{\sim}{t}_g\underset{\sim}{t}_g') = N_g(\underset{\sim}{\Sigma}_{gW}^* + N_g\underset{\sim}{\Sigma}_B^*) = N_g\underset{\sim}{\Sigma}_g^*$, we have that $E[\underset{\sim}{D}_g(\underset{\sim}{\theta})|_{\underset{\sim}{\theta}^*=\underset{\sim}{\theta}}] = \underset{\sim}{\Sigma}_B$, and $E[\underset{\sim}{C}_B(\underset{\sim}{\theta})|_{\underset{\sim}{\theta}^*=\underset{\sim}{\theta}}] = \underset{\sim}{\Sigma}_B$, then $E[\underset{\sim}{C}_k(\underset{\sim}{\theta}_i) - \underset{\sim}{\Sigma}_k(\underset{\sim}{\theta}_i)] = 0$ for all $k = 1, \ldots, G+1$; hence

$$\underset{\sim}{I}(\underset{\sim}{\theta}_i) = E[d^{20}M(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_i)] = \sum_{k=1}^{G+1}\frac{N_k}{2}\Big(\underset{\sim}{\triangle}_k(\underset{\sim}{\Sigma}_k \otimes \underset{\sim}{\Sigma}_k)^{-1}\ \underset{\sim}{\triangle}_k'\Big),$$

where $\underset{\sim}{\Sigma}_k = \underset{\sim}{\Sigma}_k^*(\underset{\sim}{\theta}_i)$. It should be noted that under the mild regularity conditions that $\underset{\sim}{\triangle}_k$ is of full rank; the "information matrix" $\underset{\sim}{I}(\underset{\sim}{\theta}_i)$ is always positive definite. With this as motivation, we propose the following modified EM gradient algorithm :

$$\underset{\sim}{\theta}_{i+1} = \underset{\sim}{\theta}_i - \underset{\sim}{I}(\underset{\sim}{\theta}_i)^{-1}\ d^{10}\ M(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_i), \tag{13}$$

in which the Hessian matrix $d^{20}M(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_i)$ is replaced by the "information matrix" $\underset{\sim}{I}(\underset{\sim}{\theta}_i)$. In structural equation modelling, this is a common practice (Lee and Jennrich (1979)). The procedure defined by (13) is also similar to the algorithm of Titterington (1984) where $d^{20}M(\underset{\sim}{\theta}_i|\underset{\sim}{\theta}_i)$ is replaced by the information matrix of the complete data, which is clearly more difficult to evaluate than $\underset{\sim}{I}(\underset{\sim}{\theta}_i)$. Moreover, it follows from Lange (1995b) that $d^{10}M(\underset{\sim}{\theta}|\underset{\sim}{\theta}) = dF(\underset{\sim}{\theta})$ holds at $\underset{\sim}{\theta} = \underset{\sim}{\theta}_i$

whenever $\utilde{\theta}_i$ is an interior point of the parameter feasible region; thus, the modified EM gradient algorithm defined in (13) can also be viewed as a gradient method in minimization of the objective function $F(\utilde{\theta}^*)$ for the observed data. The algorithm is said to have converged if $\|\utilde{\theta}_{i+1} - \utilde{\theta}_i\|$ is sufficiently small. Since $\utilde{I}(\utilde{\theta}_i)$ is positive definite, this convergence criterion is equivalent to the condition that the norm of $d^{10}M(\utilde{\theta}_i|\utilde{\theta}_i)$ or $dF(\utilde{\theta}_i)$ is sufficiently small. Hence, the modified EM gradient algorithm will converge to a minimum of $F(\utilde{\theta}^*)$. The analogous Quasi-Newton EM algorithm (Lange (1995b)) is defined as:

$$\utilde{\theta}_{i+1} = \utilde{\theta}_i - [\utilde{I}(\utilde{\theta}_i) + \utilde{B}_i]^{-1} d^{10}M(\utilde{\theta}_i|\utilde{\theta}_i). \tag{14}$$

Due to the nature of the Quasi-Newton EM algorithm, it may not be very convenient to utilize the LISREL program or the EQS program in obtaining the solution. However, since $\utilde{I}(\utilde{\theta}_i)$ and $d^{10}M(\utilde{\theta}_i|\utilde{\theta}_i)$ only involve the first derivatives of the basic covariance structures with respect to $\utilde{\theta} : \partial\utilde{\Sigma}_{gW}(\utilde{\theta})/\partial\utilde{\theta}$ and $\partial\utilde{\Sigma}_B(\utilde{\theta})/\partial\utilde{\theta}$, they can be computed and implemented without much difficulty. For the modified EM gradient algorithm, the EQS software (Bentler (1992)) which is essentially based on the scoring type algorithm in the minimization procedure, can be used to obtain the M-step iterations.

3.4. Asymptotic properties for statistical inference

At convergence, the EM algorithm or its accelerated procedures give the maximum likelihood estimate $\hat{\utilde{\theta}}^*$ of $\utilde{\theta}^*$ based on the observed data. Hence, it follows from the asymptotic results developed by Lee (1990) that the asymptotic distribution of $T^{1/2}(\hat{\utilde{\theta}}^* - \utilde{\theta}^*)$ is $N[\utilde{0}, 2\utilde{H}(\utilde{\theta}^*)^{-1}]$, where $T = N_1 + \cdots + N_G$, and

$$\utilde{H}(\utilde{\theta}^*) = G^{-1}\sum_{g=1}^{G}\Big(\utilde{\triangle}_{gW}[\utilde{\Sigma}^*_{gW} \otimes \utilde{\Sigma}^*_{gW}]^{-1}\utilde{\triangle}'_{gw} + N_g^{-1}\utilde{\triangle}_g[\utilde{\Sigma}^*_g \otimes \utilde{\Sigma}^*_g]^{-1}\utilde{\triangle}'_g\Big),$$

with $\utilde{\triangle}_{gW} = \partial\utilde{\Sigma}_{gW}(\utilde{\theta}^*)/\partial\utilde{\theta}^*$, and $\utilde{\triangle}_g = \partial\utilde{\Sigma}_g(\utilde{\theta}^*)/\partial\utilde{\theta}^*$. The estimate of the asymptotic standard errors of $\hat{\utilde{\theta}}^*$ can be obtained from the diagonal elements of $2\utilde{H}(\hat{\utilde{\theta}}^*)^{-1}$.

Let H_o be the null hypothesis that $\utilde{\Sigma}^*_{gW} = \utilde{\Sigma}_{gW}(\utilde{\theta}^*)$ for $g = 1, \ldots, G$ and $\utilde{\Sigma}^*_B = \utilde{\Sigma}_B(\utilde{\theta}^*)$, and let H_1 be the general hypothesis that all covariance matrices $\utilde{\Sigma}^*_{gW}$ and $\utilde{\Sigma}^*_B$ are any positive definite matrices. Let $\utilde{\theta}_o$ be the parameter vector under H_1; clearly, $\utilde{\theta}_o$ just consists of the distinct elements in all the within-group covariance matrices and the between-group covariance matrix. The ML estimate $\hat{\utilde{\theta}}_o$ of $\utilde{\theta}_o$ can be obtained by using either the proposed EM algorithm or by direct minimization of the function $F(\utilde{\theta}_o)$ as defined in (2) without imposing any structures in $\utilde{\Sigma}^*_B$ and $\utilde{\Sigma}^*_{gW}$. Based on the standard ML theory, the asymptotic likelihood ratio statistic to test the goodness-of-fit of the proposed covariance structures $\underset{\sim}{\Sigma}^*_{gW}$ and $\underset{\sim}{\Sigma}^*_B$ in the model is given by $\chi^2_L = -2T[F(\hat{\underset{\sim}{\theta}}_o) - F(\hat{\underset{\sim}{\theta}}^*)]$, which is asymptotically distributed as chi-square with degrees of freedom $(G + 1)p(p+1)/2 - q$. Null hypotheses on the goodness-of-fit of other "nested" models may also be tested by this type of asymptotic likelihood ratio test.

4. Examples and Simulation Studies

In this section, we will use some artificial examples to provide some idea about the empirical performance of the *EM* algorithm and the accelerated *EM* algorithms. Our main purpose is to demonstrate that these procedures work well in analyzing the two-level structural equation models; but we do not attempt to give an empirical comparison of the various *EM* methods. We will also provide results from some simulation studies to illustrate the validity of the asymptotic properties that are important for statistical inferences of the model. The related source code can be obtained from the authors upon request.

For the sake of simplicity, our examples and simulation studies will be based on the two-level confirmatory factor analysis model with the basic covariance structures defined as:

$$\underset{\sim}{\Sigma}^*_B = \underset{\sim}{\Lambda}^*_B \underset{\sim}{\Phi}^*_B \underset{\sim}{\Lambda}^{*'}_B + \underset{\sim}{\Psi}^*_B, \quad \text{and} \quad \underset{\sim}{\Sigma}^*_{gW} = \underset{\sim}{\Lambda}^*_{gW} \underset{\sim}{\Phi}^*_{gW} \underset{\sim}{\Lambda}^{*'}_{gW} + \underset{\sim}{\Psi}^*_{gW}, \tag{15}$$

where $\underset{\sim}{\Lambda}^*_B$ and $\underset{\sim}{\Lambda}^*_{gW}$ are the factor loading matrices, $\underset{\sim}{\Phi}^*_B$ and $\underset{\sim}{\Phi}^*_{gW}$ are the factors' covariance matrices and $\underset{\sim}{\Psi}^*_B$ and $\underset{\sim}{\Psi}^*_{gW}$ are the covariance matrices of the error measurements. We consider an unbalanced design situation with $G = 120$ groups in which $N_g = 4$ for $g = 1, \ldots, 40$, $N_g = 6$ for $g = 41, \ldots, 80$ and $N_g = 8$ for $g = 81, \ldots, 120$. Hence, the total sample size is 720. An artificial data set will be generated based on the following true population values:

$$\underset{\sim}{\Lambda}^{*'}_B = \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.8 & 0.8 & 0.8 & 0.8 \end{bmatrix} = \underset{\sim}{\Lambda}^{*'}_{gW}, \ g = 1, \ldots, 120$$

$$\underset{\sim}{\Phi}^*_B = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}; \underset{\sim}{\Phi}^*_{gW} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}, \ g = 1, \ldots, 120$$

and $\underset{\sim}{\Psi}^{*'}_B = \underset{\sim}{\Psi}^{*'}_{gW} = 0.36\underset{\sim}{I}_8$, $g = 1, \ldots, 120$. For illustration, the simulated data will be fit to a specific model of (15) with $\underset{\sim}{\Lambda}^*_{gW} = \underset{\sim}{\Lambda}^*_W, \underset{\sim}{\Phi}^*_{gW} = \underset{\sim}{\Phi}^*_W$, and $\underset{\sim}{\Psi}^*_{gW} = \underset{\sim}{\Psi}^*_W$. Hence, we are now considering a model with invariant covariance structure at the individual level. Of course, other special cases can be studied. To ensure identifiability, the zero's and one's in $\underset{\sim}{\Lambda}^*_B, \underset{\sim}{\Lambda}^*_W, \underset{\sim}{\Phi}^*_B, \underset{\sim}{\Phi}^*_W$ and the zero off-diagonal elements of $\underset{\sim}{\Psi}^*_B$ and $\underset{\sim}{\Psi}^*_W$ are treated as fixed known parameters that are not to be estimated. So, there are 34 unknown parameters in this model, which we call Model I.

Based on the simulated data set and the proposed covariance structures, we apply the EM algorithm developed in the previous sections to get the ML estimate of the unknown parameters. The starting values of the unknown parameters in $\{\underset{\sim}{\Lambda}^*_B, \underset{\sim}{\Lambda}^*_{gW}\}$ and $\{\underset{\sim}{\Psi}^*_B, \underset{\sim}{\Psi}^*_{gW}\}$ were taken to be 1.6 and 0.72 respectively; that is, they were taken to be twice of the true population values. The starting values of the unknown correlations in the correlation matrices of the factors were all taken to be zero. The algorithm is said to have converged to the solution if the root mean squares (RMS) of the change of the elements in the parameter vector between the EM steps is less than 0.0005. At the M-step of the basic EM algorithm, we say that the minimum has been achieved if the RMS of the change of the elements in the gradient vector is less than 0.0001. The convergence of the EM algorithm is displayed in Table 1. The second column gives the number of iterations required to complete the M-step. The EM algorithm converged to the ML solution in 6 EM iterations which contain a total of 19 iterations at the M-step. Hence, it seems that the convergence is reasonably fast.

Table 1. Convergence of the EM algorithm.

EM	M-	RMS of						
Iter	Iter	$\lvert\hat{\underset{\sim}{\theta}}^*_{i+1} - \hat{\underset{\sim}{\theta}}^*_i\rvert$	$\Lambda^*_B(1,1)$	$\Phi^*_B(2,1)$	$\Psi^*_B(1,1)$	$\Lambda^*_W(1,1)$	$\Phi^*_W(2,1)$	$\Psi^*_W(1,1)$
0			1.600	0.000	0.720	1.000	0.000	0.720
1	5	0.49042	0.957	0.255	0.394	0.936	0.282	0.442
2	4	0.10308	0.816	0.396	0.344	0.798	0.436	0.403
3	3	0.01737	0.793	0.422	0.334	0.777	0.468	0.397
4	3	0.00322	0.789	0.425	0.332	0.774	0.473	0.397
5	2	0.00084	0.788	0.425	0.331	0.774	0.474	0.397
6	2	0.00028	0.787	0.425	0.331	0.774	0.474	0.397

The modified EM gradient (EMG) algorithm and the modified Quasi- Newton EM ($QNEM$) algorithm have also been applied to the same simulated data set, based on the same starting values. The convergence of the EMG algorithm is presented in Table 2. It can be seen that the algorithm converged rapidly to the solution in 7 iterations. It should be noted that the computation of one EMG iteration is basically equal to one M-step iteration in the basic EM algorithm, so clearly the EMG algorithm is better. But since the basic EM algorithm performs quite well, the improvement is not dramatic. We found that the $QNEM$ algorithm also converged in 7 iterations and its convergence is very similar to the EMG algorithm. In fact, since the problem involved only hypothetical missing data, the contribution of the missing Hessian matrix in improving the convergence of the efficient EMG algorithm is minor. To save space, the convergence of the QNEM algorithm is not reported.

Table 2. Convergence of the EMG algorithm.

EM Iter	RMS of $d^{10}M(\underset{\sim}{\theta}\vert\underset{\sim}{\theta})$	RMS of $\vert\hat{\underset{\sim}{\theta}}^*_{i+1}-\hat{\underset{\sim}{\theta}}^*_i\vert$	$\Lambda^*_B(1,1)$	$\Phi^*_B(2,1)$	$\Psi^*_B(1,1)$	$\Lambda^*_W(1,1)$	$\Phi^*_W(2,1)$	$\Psi^*_W(1,1)$
0			1.600	0.000	0.720	1.600	0.000	0.720
1	0.05836	0.40940	1.090	0.093	0.388	1.070	0.099	0.442
2	0.03774	0.15393	0.871	0.254	0.336	0.852	0.271	0.403
3	0.00946	0.05350	0.807	0.381	0.328	0.788	0.415	0.398
4	0.00099	0.01390	0.792	0.418	0.329	0.776	0.463	0.397
5	0.00016	0.00314	0.789	0.424	0.330	0.774	0.472	0.397
6	0.00004	0.00094	0.788	0.425	0.330	0.774	0.474	0.397
7	0.00001	0.00034	0.787	0.425	0.331	0.774	0.474	0.397

To study the performance of the algorithms in analyzing models with cross level parameters, we have simulated another data set based on the same settings as above and fit the data to the above model with additional constraints that $\underset{\sim}{\Lambda}^*_B = \underset{\sim}{\Lambda}^*_W$. There are 26 unknown parameters in this model; we call this Model II. We found that the convergence of the algorithms are quite similar. Using the same starting values as before, the EM algorithm converged to the ML solution in 6 EM iterations with 19 M-step iterations, while both the EMG and $QNEM$ algorithms converged in 7 iterations. The convergence is not presented to save space.

Moreover, we consider another simulated data set with $G = 120$ groups which are divided evenly into six clusters with unequal sample sizes. The level-one sample sizes N_g for groups in these clusters are equal to $2, 8, 16, 32, 64$ and 96, respectively. Using the same starting values and convergence criterion, the EM algorithm converged in 5 EM iterations with 16 M-step iterations for Model I; while for Model II, it converged in 5 EM iterations with 20 M-step iterations. The EMG and QNEM algorithms converged in 5 iterations for both Model I and Model II. Hence, it is apparent that the proposed algorithms converged rapidly even for data sets with level-one sample sizes substantially varied.

Based on the above results, we used the EMG algorithm to obtain the ML estimates in the simulation study concerning the asymptotic behavior. Models I and II as described above were considered. For each model, the following sample size designs of G and N_g were used:

A: $G = 60$, 20 groups with $N_g = 4,\ 6,\ 8;$ $T = 360;$
B: $G = 60$, 20 groups with $N_g = 8,\ 12,\ 16;$ $T = 720;$
C: $G = 120$, 40 groups with $N_g = 4,\ 6,\ 8;$ $T = 720;$
D: $G = 120$, 40 groups with $N_g = 8,\ 12,\ 16;$ $T = 1440.$

For each of these cases, 100 replications were completed; and for each replication, the ML estimates, the standard error estimates and the goodness-of-fit statistic χ^2_L were computed based on the results developed in Sections 3.3 and 3.4.

The root mean squares between the ML estimates and the true values of the parameters in Model I and Model II are presented in Tables 3 and 4, respectively. From results in these tables, we observe that the ML estimates based on these sample sizes are accurate. As expected, the accuracy increased with the sample sizes, hence the ML estimates of the parameters in the equal within-group covariance structures are generally more accurate than the estimates of the parameters in the between-group covariance structure.

Table 3. Root mean squares of the estimates and population values in Model I.

		Sample Size Design			
Parameter	True Value	A	B	C	D
$\Lambda_B^*(1,1)$	0.80	0.140	0.119	0.098	0.092
$\Lambda_B^*(2,1)$	0.80	0.151	0.124	0.111	0.089
$\Lambda_B^*(3,1)$	0.80	0.136	0.117	0.100	0.094
$\Lambda_B^*(4,1)$	0.80	0.115	0.120	0.092	0.090
$\Lambda_B^*(5,2)$	0.80	0.138	0.121	0.102	0.093
$\Lambda_B^*(6,2)$	0.80	0.141	0.110	0.085	0.099
$\Lambda_B^*(7,2)$	0.80	0.117	0.118	0.090	0.087
$\Lambda_B^*(8,2)$	0.80	0.127	0.113	0.090	0.076
$\Phi_B^*(2,1)$	0.30	0.162	0.139	0.106	0.099
$\Psi_B^*(1,1)$	0.36	0.095	0.087	0.072	0.070
$\Psi_B^*(2,2)$	0.36	0.113	0.091	0.068	0.068
$\Psi_B^*(3,3)$	0.36	0.102	0.106	0.071	0.080
$\Psi_B^*(4,4)$	0.36	0.112	0.101	0.073	0.072
$\Psi_B^*(5,5)$	0.36	0.118	0.116	0.073	0.069
$\Psi_B^*(6,6)$	0.36	0.093	0.093	0.076	0.065
$\Psi_B^*(7,7)$	0.36	0.114	0.099	0.073	0.069
$\Psi_B^*(8,8)$	0.36	0.118	0.106	0.074	0.069
$\Lambda_W^*(1,1)$	0.80	0.053	0.030	0.036	0.024
$\Lambda_W^*(2,1)$	0.80	0.053	0.032	0.031	0.025
$\Lambda_W^*(3,1)$	0.80	0.047	0.033	0.033	0.024
$\Lambda_W^*(4,1)$	0.80	0.051	0.032	0.032	0.025
$\Lambda_W^*(5,2)$	0.80	0.047	0.033	0.033	0.026
$\Lambda_W^*(6,2)$	0.80	0.049	0.029	0.036	0.024
$\Lambda_W^*(7,2)$	0.80	0.054	0.033	0.035	0.021
$\Lambda_W^*(8,2)$	0.80	0.052	0.031	0.030	0.024
$\Phi_W^*(2,1)$	0.50	0.054	0.036	0.037	0.022
$\Psi_W^*(1,1)$	0.36	0.039	0.029	0.024	0.017
$\Psi_W^*(2,2)$	0.36	0.037	0.026	0.029	0.017
$\Psi_W^*(3,3)$	0.36	0.037	0.026	0.027	0.018
$\Psi_W^*(4,4)$	0.36	0.040	0.029	0.028	0.020
$\Psi_W^*(5,5)$	0.36	0.036	0.024	0.025	0.017
$\Psi_W^*(6,6)$	0.36	0.035	0.025	0.027	0.019
$\Psi_W^*(7,7)$	0.36	0.038	0.026	0.028	0.017
$\Psi_W^*(8,8)$	0.36	0.039	0.027	0.025	0.018

Table 4. Root mean squares of estimates and population values in Model II.

		Sample Size Design			
Parameter	True Value	A	B	C	D
$\Lambda_B^*(1,1) = \Lambda_W^*(1,1)$	0.80	0.046	0.037	0.029	0.022
$\Lambda_B^*(2,1) = \Lambda_W^*(2,1)$	0.80	0.053	0.035	0.034	0.023
$\Lambda_B^*(3,1) = \Lambda_W^*(3,1)$	0.80	0.050	0.033	0.035	0.020
$\Lambda_B^*(4,1) = \Lambda_W^*(4,1)$	0.80	0.044	0.035	0.035	0.020
$\Lambda_B^*(5,2) = \Lambda_W^*(5,2)$	0.80	0.047	0.029	0.030	0.022
$\Lambda_B^*(6,2) = \Lambda_W^*(6,2)$	0.80	0.051	0.032	0.034	0.022
$\Lambda_B^*(7,2) = \Lambda_W^*(7,2)$	0.80	0.046	0.027	0.031	0.021
$\Lambda_B^*(8,2) = \Lambda_W^*(8,2)$	0.80	0.043	0.031	0.029	0.026
$\Phi_B^*(2,1)$	0.30	0.147	0.158	0.106	0.097
$\Psi_B^*(1,1)$	0.36	0.099	0.083	0.073	0.071
$\Psi_B^*(2,2)$	0.36	0.104	0.088	0.073	0.065
$\Psi_B^*(3,3)$	0.36	0.104	0.093	0.086	0.072
$\Psi_B^*(4,4)$	0.36	0.116	0.095	0.074	0.073
$\Psi_B^*(5,5)$	0.36	0.097	0.093	0.072	0.058
$\Psi_B^*(6,6)$	0.36	0.078	0.086	0.071	0.066
$\Psi_B^*(7,7)$	0.36	0.095	0.097	0.074	0.073
$\Psi_B^*(8,8)$	0.36	0.097	0.093	0.064	0.072
$\Phi_W^*(2,1)$	0.50	0.052	0.032	0.038	0.023
$\Psi_W^*(1,1)$	0.36	0.034	0.023	0.024	0.020
$\Psi_W^*(2,2)$	0.36	0.040	0.027	0.030	0.019
$\Psi_W^*(3,3)$	0.36	0.039	0.025	0.031	0.017
$\Psi_W^*(4,4)$	0.36	0.042	0.024	0.025	0.018
$\Psi_W^*(5,5)$	0.36	0.040	0.024	0.026	0.022
$\Psi_W^*(6,6)$	0.36	0.035	0.025	0.026	0.018
$\Psi_W^*(7,7)$	0.36	0.039	0.026	0.030	0.017
$\Psi_W^*(8,8)$	0.36	0.037	0.024	0.027	0.017

Let $SD(\hat{\theta}^*_{(i)})$ be the empirical standard deviation obtained from the 100 estimates of the ith element of $\underset{\sim}{\theta}^*$, $\hat{\theta}^*_{(i)}$; and $\overline{SE}(\hat{\theta}^*_{(i)})$ be the mean of the 100 standard error estimates of $\hat{\theta}^*_{(i)}$. The ratios $SD(\hat{\theta}^*_{(i)})/\overline{SE}(\hat{\theta}^*_{(i)})$ corresponding to the parameters in Models I and II are reported in Tables 5 and 6, respectively. It can be seen that these ratios are close to 1.0, indicating the proposed method of getting the standard error estimates via $\underset{\sim}{H}(\hat{\underset{\sim}{\theta}}^*)$ is acceptable.

Table 5. $SD(\hat{\underset{\sim}{\theta}}^*_{(i)})/\overline{SE}(\hat{\underset{\sim}{\theta}}^*_{(i)})$ of the parameters estimates in Model I.

	Sample Size Design			
Parameter	A	B	C	D
$\Lambda_B^*(1,1)$	1.087	0.996	1.083	1.117
$\Lambda_B^*(2,1)$	1.167	1.042	1.219	1.079
$\Lambda_B^*(3,1)$	1.070	0.996	1.109	1.102
$\Lambda_B^*(4,1)$	0.906	1.026	1.012	1.083
$\Lambda_B^*(5,2)$	1.067	1.016	1.116	1.112
$\Lambda_B^*(6,2)$	1.113	0.933	0.935	1.197
$\Lambda_B^*(7,2)$	0.923	1.000	0.986	1.045
$\Lambda_B^*(8,2)$	1.011	0.960	0.989	0.918
$\Phi_B^*(2,1)$	1.092	1.010	0.987	0.996
$\Psi_B^*(1,1)$	0.950	0.940	1.026	1.069
$\Psi_B^*(2,2)$	1.127	0.953	0.953	1.047
$\Psi_B^*(3,3)$	1.029	1.161	1.008	1.245
$\Psi_B^*(4,4)$	1.099	1.113	1.031	1.117
$\Psi_B^*(5,5)$	1.182	1.213	0.991	1.071
$\Psi_B^*(6,6)$	0.963	1.016	1.069	1.000
$\Psi_B^*(7,7)$	1.161	1.074	1.024	1.063
$\Psi_B^*(8,8)$	1.188	1.147	1.012	1.058
$\Lambda_W^*(1,1)$	1.167	0.929	1.113	1.054
$\Lambda_W^*(2,1)$	1.169	0.988	0.976	1.083
$\Lambda_W^*(3,1)$	1.047	1.011	1.030	1.059
$\Lambda_W^*(4,1)$	1.121	1.006	1.016	1.082
$\Lambda_W^*(5,2)$	1.024	1.050	1.023	1.144
$\Lambda_W^*(6,2)$	1.072	0.915	1.114	1.075
$\Lambda_W^*(7,2)$	1.206	1.048	1.096	0.922
$\Lambda_W^*(8,2)$	1.133	0.957	0.958	1.044
$\Phi_W^*(2,1)$	1.157	1.100	1.125	0.952
$\Psi_W^*(1,1)$	1.127	1.157	0.976	1.003
$\Psi_W^*(2,2)$	1.039	1.050	1.166	0.975
$\Psi_W^*(3,3)$	1.061	1.050	1.083	0.999
$\Psi_W^*(4,4)$	1.157	1.172	1.148	1.152
$\Psi_W^*(5,5)$	1.018	0.967	1.011	0.975
$\Psi_W^*(6,6)$	0.988	0.996	1.106	1.060
$\Psi_W^*(7,7)$	1.079	1.044	1.113	0.986
$\Psi_W^*(8,8)$	1.101	1.103	1.007	1.002

Table 6. $SD(\hat{\underset{\sim}{\theta}}^*_{(i)})/\overline{SE}(\hat{\underset{\sim}{\theta}}^*_{(i)})$ of the parameters estimates in Model II.

	Sample Size Design			
Parameter	A	B	C	D
$\Lambda^*_B(1,1) = \Lambda^*_W(1,1)$	1.099	1.199	1.000	1.024
$\Lambda^*_B(2,1) = \Lambda^*_W(2,1)$	1.261	1.151	1.165	1.027
$\Lambda^*_B(3,1) = \Lambda^*_W(3,1)$	1.195	1.070	1.172	0.915
$\Lambda^*_B(4,1) = \Lambda^*_W(4,1)$	1.060	1.134	1.180	0.942
$\Lambda^*_B(5,2) = \Lambda^*_W(5,2)$	1.133	0.960	1.036	1.033
$\Lambda^*_B(6,2) = \Lambda^*_W(6,2)$	1.230	1.052	1.169	1.044
$\Lambda^*_B(7,2) = \Lambda^*_W(7,2)$	1.101	0.897	1.057	0.961
$\Lambda^*_B(8,2) = \Lambda^*_W(8,2)$	1.034	1.014	0.988	1.186
$\Phi^*_B(2,1)$	0.997	1.160	1.020	1.013
$\Psi^*_B(1,1)$	0.988	0.959	1.076	1.147
$\Psi^*_B(2,2)$	1.060	1.009	1.051	1.051
$\Psi^*_B(3,3)$	1.052	1.082	1.258	1.142
$\Psi^*_B(4,4)$	1.147	1.080	1.088	1.191
$\Psi^*_B(5,5)$	0.979	1.074	1.045	0.931
$\Psi^*_B(6,6)$	0.825	0.985	1.022	1.066
$\Psi^*_B(7,7)$	1.006	1.091	1.096	1.186
$\Psi^*_B(8,8)$	1.013	1.044	0.950	1.095
$\Phi^*_W(2,1)$	1.126	0.989	1.161	1.011
$\Psi^*_W(1,1)$	0.966	0.928	0.985	1.162
$\Psi^*_W(2,2)$	1.149	1.072	1.216	1.084
$\Psi^*_W(3,3)$	1.119	1.007	1.243	0.973
$\Psi^*_W(4,4)$	1.182	0.953	1.008	1.059
$\Psi^*_W(5,5)$	1.163	0.983	1.074	1.243
$\Psi^*_W(6,6)$	0.994	1.003	1.070	1.046
$\Psi^*_W(7,7)$	1.105	1.055	1.195	0.964
$\Psi^*_W(8,8)$	1.019	0.948	1.092	0.995

Based on the definition of the model and the parameters' specification, the asymptotic goodness-of-fit statistics χ^2_L for Model I and Model II should be chi-square with 38 and 46 degrees of freedom, respectively. These were tested by the Kolomogorov-Smirnov statistic based on the χ^2_L values obtained from the 100 replications. For Model I, the p-values corresponding to the sample size designs A, B, C, D are 0.8885, 0.5173, 0.1010 and 0.6269, respectively; while for Model II, the corresponding values are 0.8952, 0.5322, 0.4612 and 0.3759, respectively.

From these results, it is reasonable to conclude that the empirical behavior of this test statistic agrees with the theoretical asymptotic result.

5. Discussion

The basic ML theory for multilevel structural equation modelling is given by Lee (1990) and some procedures for computation of the solution in some special situation have been presented by Muthen (1990), Longford and Muthen (1992), and Raudenbush (1995). However, since these procedures cannot be easily used by most practitioners, the applications of multilevel structural equation models to real life situations are still relatively limited. By treating the second level random vectors as hypothetical missing data and analyzing the model as a missing data problem, this paper investigates the application of the EM algorithm to obtain the ML solution of the general two-level model with unbalanced designs. It is shown that the EM algorithm works well and moreover its performance can also be improved with the accelerated EM gradient procedure. Based on the results obtained in previous sections, it should be evident that the EM approach has at least the following attractive features:

(i) It converged rapidly to the ML solution. In our examples with 34 and 26 unknown parameters, the EMG algorithm converged in 7 or 5 iterations.

(ii) The ML solution is obtained by minimizing a simple function $M(\underset{\sim}{\theta}^*|\underset{\sim}{\theta})$ as defined in (8). A simple program to compute $\underset{\sim}{C}_k(\underset{\sim}{\theta})$ is required and the more complex task of minimization can be completed conveniently with existing software such as LISREL (Jöreskog and Sörbom (1996)) or EQS (Bentler (1992)).

(iii) Depending on the software used in the analyses, the covariance models under consideration can be the LISREL model (Jöreskog and Sörbom (1996)) or the Bentler and Week's (Bentler and Weeks (1980)) model in EQS. It is well known that these two general models are sufficient for most real-life applications.

(iv) Since $\underset{\sim}{I}(\underset{\sim}{\theta}_i)$ is positive definite, the EM algorithm is robust to poor starting values because it is a descent algorithm that produces an acceptable step at every iteration.

The degree of improvement on convergence rate of the $QNEM$ algorithm over that of the EMG algorithm depends on the effect of the additional matrix $\underset{\sim}{B}_i$ in helping the "information" matrix $\underset{\sim}{I}(\underset{\sim}{\theta}_i)$ to approximate the Hessian matrix of $F(\underset{\sim}{\theta}^*)$ at $\underset{\sim}{\theta}_i$. It can be shown that this approximation is better with increase of N_g. Due to the extremely quick convergence of the EMG algorithm in our artificial examples, it is apparent that the approximation is very good even with quite small N_g. Thus, the performance of the EMG algorithm is apparently satisfactory. The conjugate gradient acceleration proposed by Jamshidian and

Jennrich (1993) and the ECME algorithm of Liu and Rubin (1994) are other attractive alternatives. For the sake of brevity, the present paper does not attempt to provide an empirical comparison of various EM type acceleration procedures. Such comparison would be interesting for future research.

Acknowledgements

The authors would like to thank the associate editor and referees for their helpful comments. This research was supported by a Hong Kong UGC Earmarked grant.

References

Bentler, P. M. (1992). *EQS Structural Equations Program Manual.* BMDP Statistical Software : Los Angeles.

Bentler, P. M. and Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika* **45**, 289-308.

Bock, R. D. (1989). *Multilevel Analysis of Educational Data.* Academic Press, Inc., San Diego, CA.

Dempster, A. P., Laird, N. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Goldstein, H. (1987). *Multilevel Model in Educational and Social Research.* Oxford University Press, New York.

Jamshidian, M. and Jennrich. R. I. (1993). Conjugate gradient acceleration of the EM algorithm. *J. Amer. Statist. Assoc.* **88**, 221-228.

Jöreskog, K. G. and Sörbom (1996). *LISREL 8 : User's Reference Guide.* Scientific Software International, Chicago.

Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **57**, 425-437.

Lange, K. (1995b). A Quasi-Newton acceleration of the EM algorithm. *Statist. Sinica* **5**, 1-18.

Lee, S. Y. (1990). Multilevel analysis of Structural Equation Models. *Biometrika* **77**, 763-772.

Lee, S. Y. and Jennrich, R. I. (1979). A Study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika* **44**, 99-113.

Lee, S. Y. and Poon, W. Y. (1992). Two level analysis of covariance structures for unbalanced designs with small level-one samples. *British J. Math. Statist. Psych.* **45**, 109-123.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633-648.

Longford, N. T. and Muthén, B. (1992). Factor analysis for clustered observations. *Psychometrika* **57**, 581-597.

McDonald, R. P. and Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British J. Math. Statist. Psych.* **42**, 215-232.

Muthen, B. (1990). Mean and covariance structure analysis of hierarchical data. UCLA Statistics Series, Number 62.

Raudenbush, S. W. (1995). Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British J. Math. Statist. Psych.* **48**, 359-370.

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

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47**, 69-76.

Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. *J. Roy. Statist. Soc. Ser. B* **46**, 257-267.

Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong.
E-mail: b003784@idea.csc.cuhk.hk
E-mail: wypoon@hp735.sta.cuhk.edu.hk

(Received March 1996; accepted April 1997)