Statistica Sinica **8**(1998), 729-747

MAXIMUM LIKELIHOOD ESTIMATION OF FACTOR ANALYSIS USING THE ECME ALGORITHM WITH COMPLETE AND INCOMPLETE DATA

Chuanhai Liu and Donald B. Rubin

Bell Labs and Harvard University

Abstract: Factor analysis is a standard tool in educational testing contexts, which can be fit using the EM algorithm (Dempster, Laird and Rubin (1977)). An extension of EM, called the ECME algorithm (Liu and Rubin (1994)), can be used to obtain ML estimates more efficiently in factor analysis models. ECME has an E-step, identical to the E-step of EM, but instead of EM's M-step, it has a sequence of CM (conditional maximization) steps, each of which maximizes either the constrained expected complete-data log-likelihood, as with the ECM algorithm (Meng and Rubin (1993)), or the constrained actual log-likelihood. For factor analysis, we use two CM steps: the first maximizes the expected complete-data log-likelihood over the factor loadings given fixed uniquenesses, and the second maximizes the actual likelihood over the uniquenesses given fixed factor loadings. We also describe EM and ECME for ML estimation of factor analysis from incomplete data, which arise in applications of factor analysis in educational testing contexts. ECME shares with EM its monotone increase in likelihood and stable convergence to an ML estimate, but converges more quickly than EM. This more rapid convergence not only can shorten CPU time, but at least as important, it allows for a substantially easier assessment of convergence, as shown by examples. We believe that the application of ECME to factor analysis illustrates the role that extended EM-type algorithms, such as the even more general AECM algorithm (Meng and van Dyk (1997)) and the PX-EM algorithm (Liu, Rubin and Wu (1997)), can play in fitting complex models that can arise in educational testing contexts.

Key words and phrases: EM, ECM, incomplete data, missing data.

1. Introduction

Factor analysis has been a standard tool in psychology, psychometrics, and educational testing contexts for the better part of this century (for recent references, see Longford and Muthén (1992) and Meredith (1993)). For many years, maximum likelihood (ML) estimation has been popular for fitting factor analysis models, especially those having restrictions on the parameters, the "confirmatory case." A variety of iterative computational methods can be used to perform ML estimation (e.g., LISREL-7, Jöreskog and Sörbom (1988)), but probably the easiest to implement and one of the most stable in the sense of monotonely increasing the likelihood, is the EM algorithm (Dempster, Laird and Rubin (1977),

henceforth, DLR). Introduced by DLR, EM for ML factor analysis was described in detail in Rubin and Thayer (1982). Despite its reliable monotone convergence, the rate of convergence of EM can be painfully slow in factor analysis models, so slow, in fact, that it can be difficult to assess convergence, as noted, for example, by Bentler and Tanaka (1983) and further discussed by Rubin and Thayer (1983).

A new algorithm, the ECME algorithm of Liu and Rubin (1994), can be applied to factor analysis and related models, and holds much promise as it shares advantages with both EM and Newton-stepping algorithms. ECME is an extension of the ECM algorithm of Meng and Rubin (1993), itself an extension of EM, but ECME's rate of convergence, at least judged by the number of iterations, is substantially faster than either EM or ECM, yet it retains EM's stable monotone convergence to an ML estimate, and is only modestly more difficult to implement. This increased rate of monotone convergence makes it easier to judge convergence, and total computation time can be less than with EM, especially in difficult cases.

Briefly, the ECM (Expectation, Conditional Maximization) algorithm modifies the EM (Expectation, Maximization) algorithm by replacing its M step, which maximizes the current expected complete-data log-likelihood over the entire vector parameter θ, by a sequence of conditional maximization steps (indexed by $s = 1, \ldots, S$), each of which maximizes the expected complete-data log-likelihood but over a function of θ, say θ_s, subject to the rest of θ, say $\overline{\theta_s}$, being fixed at previously estimated values. If the $(\theta_1, \ldots, \theta_S)$ span the parameter space of θ, the ECM algorithm will converge in the same way as EM to an ML estimate. ECME (Expectation, Conditional Maximization of Either) replaces each of one or more of ECM's final CM steps with a step that conditionally maximizes the actual likelihood function over θ_s rather than the expected complete-data log-likelihood as with ECM. Typically, the conditional maximization of the actual likelihood over θ_s is more difficult than the conditional maximization of the expected complete-data log-likelihood over θ_s. Thus, ECME is typically more tedious to implement than ECM, and some of its steps are computationally more expensive. The reward, however, is an increased rate of convergence, with an attendant increased ability to assess convergence and decreased total computer time, both obtained without losing the monotone increase in likelihood and simple implementation, both advantages relative to potentially faster converging Newton-stepping methods. ECME itself can be embedded in the even more general AECM algorithm (Meng and van Dyk (1997)), which is closely related to multicycle ECM (MCECM, Meng and Rubin (1993)). Another advantage of using these EM-type algorithms is that large sample standard errors (if desired

(see warnings in Rubin and Thayer (1983), concerning their inferential propriety)) can be obtained numerically using only the code for EM (Meng and Rubin (1991)) or ECM (van Dyk, Meng and Rubin (1995)).

In Section 2 we briefly review the factor analysis model, and in Section 3 we present ML estimation using EM and ECME from complete data. Our version of ECME for factor analysis with complete data uses only two CM steps, the first conditionally maximizing the expected complete-data log-likelihood over the factor loadings given the uniquenesses using closed form expressions, and the second conditionally maximizing the actual likelihood over the uniquenesses using simple low-dimensional Newton-Raphson, which in this case is reliable. In section 5 we apply EM and ECME to the complete-data numerical example used in Rubin and Thayer (1982), which shows that in this example, ECME relative to EM takes 1/5 the number of iterations and 80% of the CPU time on a SPARC station 2.

We do not attempt a complete comparison with the array of competing algorithms for ML factor analysis (e.g., Jöreskog (1967), Jennrich and Robinson (1969), Clarke (1970), Jamshidian and Jennrich (1993)). Our emphasis is to show that ECME, like EM and ECM, is easily implemented and has stable monotone convergence, but can be more effective in practice when the rate of convergence of EM is very slow. Moreover, EM, ECM, and ECME can, with essentially no modification, handle missing data in the variables as presented in Section 4, which most other methods can not (also see Little and Rubin (1987), p.149, for EM; and Liu (1996), for ECME). The version of ECME that we implement most closely parallels the ECME algorithm with complete data and has an E step and three CM steps: the first CM-step maximizes the expected log-likelihood over the factor loading matrix, the second CM-step maximizes the constrained actual likelihood over the population mean of the variables, and the third CM-step updates the uniquenesses by maximizing the constrained actual likelihood. Our example in Section 5, analyzing three models for an educational testing data set with missing values, shows that when some of uniquenesses are close to or equal to zero, the basic EM algorithm is hopelessly slow whereas ECME converges satisfactorily. ECME, or a combination of early iterations of EM followed by ECME, appears much preferable to straight EM.

2. The Model

Factor analysis can be viewed as a normal linear regression analysis of an observed p-dimensional variable Y on an unobservable variable Z consisting of $q < p$ factors that are themselves normal; the key assumption allowing estimation despite all Z being missing is that the components of Y are conditionally

independent given Z. For n independent observations of Y, we then have

$$Y_i|(Z_i,\beta,R,\sigma^2) \overset{\text{ind}}{\sim} \mathrm{N}_p\left(\alpha + Z_i\beta, \mathrm{Diag}(\sigma_1^2,\ldots,\sigma_p^2)\right) \quad \text{for } i=1,\ldots,n,$$

where Y_i is the $(1\times p)$ vector of the ith observation, α is the $(1\times p)$ mean vector, Z_i is the $(1\times q)$ vector of the q factors that follows $Z_i \overset{\text{iid}}{\sim} N_q(0,R)$ with $R>0$, β is the $(q\times p)$ regression coefficient matrix, and σ_j^2 is non-negative scalar for $j=1,\ldots,p$. In the terminology of factor analysis, β is the factor-loading matrix, and $\sigma^2=(\sigma_1^2,\ldots,\sigma_p^2)$ is the vector of uniquenesses. Commonly, R is assumed to be the identity matrix I. The factor loading matrix, β, may contain *a priori* zeros, in which case the model is called a confirmatory factor analysis model. In general, $\theta=(\beta,R,\sigma^2)$ is to be estimated along with α.

For identifiability conditions, especially for exploratory factor analysis, see Anderson (1984) and Basilevsky (1994). In the situation without fully identifiable parameters, EM-type algorithms converge to points that are equivalent with respect to the observed likelihood function.

3. The EM and ECME Algorithms for Factor Analysis from Complete Data

Given n fully observed observations $Y_1,\ldots,Y_n$, the ML estimate of α is $\bar{Y}=(1/n)\sum_{i=1}^n Y_i$, and the complete-data sufficient statistics for θ are $C_{yy}=(1/n)\sum_{i=1}^n(Y_i\text{–}\bar{Y})'(Y_i\text{–}\bar{Y})$, $C_{yz}=(1/n)\sum_{i=1}^n(Y_i\text{–}\bar{Y})'Z_i$, and $C_{zz}=(1/n)\sum_{i=1}^n Z_i'Z_i$.

The EM algorithm is straightforward (also see Rubin and Thayer (1982)).

EM

E-step: given the current estimate of θ, calculate the expected complete-data sufficient statistics $E(C_{yy}|Y,\theta)=C_{yy}$, $\hat{C}_{yz}=E(C_{yz}|Y,\theta)=C_{yy}\gamma$, and $\hat{C}_{zz}=E(C_{zz}|Y,\theta)=\gamma' C_{yy}\gamma+\Delta$, where γ and Δ are the regression coefficient matrix and residual covariance matrix of Z on Y given θ, respectively, which can be obtained using the Gaussian sweep operator (e.g., see Little and Rubin (1987), Section 6.5) as follows:

$$\begin{bmatrix} \beta' R\beta + \mathrm{Diag}(\sigma^2) & \beta' R \\ R\beta & R \end{bmatrix} \underset{\Longrightarrow}{\mathrm{SWP}_{[1,2.,,,.p]}} \begin{bmatrix} -(\beta' R\beta + \mathrm{Diag}(\sigma^2))^{-1} & \gamma \\ \gamma' & \Delta \end{bmatrix},$$

where $\mathrm{SWP}[\mathcal{S}]$ means application of the sweep operator to the matrix on the left hand side with respect to the diagonal pivotal elements indicated by the index set $\mathcal{S}$.

M-step: replace the complete-data sufficient statistics C_{yy}, C_{yz}, and C_{zz} with their expected values, and then find the complete-data maximum likelihood estimates of (β,σ^2) and R (if it is unknown). Letting $\mathcal{S}_i$ be the set of the indexes of

all the nonzero factors for the ith component y_i of the outcome variable Y, then

$$\begin{bmatrix} C_{yy} & C_{yy}\gamma \\ \gamma' C_{yy} & \gamma' C_{yy}\gamma + \Delta \end{bmatrix} \overset{\text{SWP}[\mathcal{S}_i+p]}{\Longrightarrow} \begin{bmatrix} \tilde{\Sigma}_{(p\times p)} & * \\ \tilde{\beta}_{(q\times p)} & * \end{bmatrix},$$

where $\mathcal{S}_i + p$ is the set consisting of all the elements in $\mathcal{S}_i$ increased by p, the (i,i)th element of the $(p \times p)$ matrix $\tilde{\Sigma}$ is $\hat{\sigma}_i^2$, and $\hat{\beta}_{ji} = 0$ if, *a priori*, β_{ji} is zero; otherwise $\hat{\beta}_{ji}$ is the (j,i)th element, $\tilde{\beta}_{ji}$, of the $(q \times p)$ matrix $\tilde{\beta}$. When some outcome variables have the same nonzero loading factors, the corresponding factor-loading coefficients are obtained simultaneously. If R contains unknowns, $\hat{R}$ can be obtained by maximizing $\ln(|R|) - \text{tr}(\hat{C}_{zz}R^{-1})$, which may need some special techniques according to the specification of the unknowns of R.

As has been noticed in practice, because of the large fraction of missing information contained in the missing factor scores Z, EM for ML factor analysis can have a very slow convergence rate. With ECME, we partition $\theta = (\beta, R, \sigma^2)$ into $\theta_1 = (\beta, R)$ and $\theta_2 = \sigma^2$, where CM step 1 maximizes the expected complete-data log-likelihood over θ_1, and CM step 2 maximizes the actual likelihood over θ_2; α is still estimated by $\bar{Y}$. The reason for this choice of CM steps is that the actual likelihood is simply a p-dimensional normal with restrictions on the covariance matrix, and numerical maximization over p-dimensional σ^2 with (β, R) fixed is easier than numerical maximization over $(p \times q)$-dimensional β and (possibly) R with σ^2 fixed. Thus, each iteration of this version of ECME consists of an E-step and two CM-steps. The ECM algorithm corresponding to this partition of θ is the same as EM because, with observed factors, Z, the maximizations over (β, R) and over σ^2 involve distinct factors in the likelihood.

ECME

E-step: the same as the E-step of EM.

CM-step 1: the same as the M-step of EM for parameters β and R.

CM-step 2: find $\hat{\sigma}^2$ to maximize the actual constrained likelihood given β, which can be done, for example, using Newton-Raphson iterations. More specifically, $\hat{\sigma}^2$ maximizes the function

$$f(\sigma^2) = -\ln\left|\beta' R\beta + \text{Diag}(\sigma^2)\right| - \text{tr}\Big(C_{yy}(\beta' R\beta + \text{Diag}(\sigma^2))^{-1}\Big) \qquad (1)$$

over σ^2 for fixed β and R.

The details of the Newton-Raphson method for finding σ^2 to maximize the function $f(\sigma^2)$ in equation (1) are simple, and the method is quite reliable relative to globally maximizing the actual likelihood function of the parameters $\theta = (\beta, R, \sigma^2)$ over both σ^2 and (β, R). The detailed computation of the gradient and the Hessian for the Newton-Raphson method are special cases of those

given in Section 4. The convergence criterion for the Newton-Raphson method is not critical as long as Newton-Raphson increases the likelihood function. In practice, only one or two steps of Newton-Raphson are needed when ECME is close to convergence, but because Newton-Raphson does not guarantee (monotone) convergence, it can be important to check that Newton-Raphson actually increases the constrained likelihood.

4. The EM and ECME Algorithms for Factor Analysis from Incomplete Data

As can be seen from the previous section or Rubin and Thayer (1982), the EM algorithm actually finds the ML estimates of the parameters (β, R, σ^2) by iteratively maximizing the expected complete-data log-likelihood given $\hat{\alpha} = \sum_{i=1}^{n} Y_i/n$, which is obtained by maximizing the actual likelihood function. This version of EM can be regarded as a trivial version of ECME in which there is a CM step maximizing the constrained actual likelihood over α given (β, R, σ^2). This CM step gives $\hat{\alpha} = \sum_{i=1}^{n} Y_i/n$ at all iterations, because the maximization does not involve the current estimates of the other parameters. This clarification helps to understand the modifications needed to extend EM-type algorithms for factor analysis from incomplete data.

We define the complete data to be $\{Y_{i,\text{obs}}, Y_{i,\text{mis}}, Z_i : i = 1, \ldots, n\}$, where $Y_{i,\text{obs}}$ and $Y_{i,\text{mis}}$ are, respectively, the observed and missing components of Y_i. Denote by Y_{obs} the observed data $\{Y_{i,\text{obs}} : i = 1, \ldots, n\}$, and now denote by θ all the parameters in the model, including α, i.e., $\theta = \{\alpha, \beta, R, \sigma^2\}$. The complete-data log-likelihood function is

$$
\begin{aligned}
&L_{\text{com}}(\theta | Y_{i,\text{obs}}, Y_{i,\text{mis}}, Z_i : i = 1, \ldots, n) \\
&= -\frac{n}{2}\sum_{i=1}^{p} \ln \sigma_i^2 - \frac{1}{2}\text{trace}\Big[\text{Diag}^{-1}(\sigma^2)\sum_{i=1}^{n}(Y_i - \alpha - Z_i\beta)'(Y_i - \alpha - Z_i\beta)\Big] \\
&\quad -\frac{n}{2}\ln|R| - \frac{1}{2}\text{trace}\Big[R^{-1}\sum_{i=1}^{n} Z_i'Z_i\Big] \\
&= -\frac{n}{2}\sum_{i=1}^{p} \ln \sigma_i^2 - \frac{1}{2}\text{trace}\Big[\text{Diag}^{-1}(\sigma^2)\sum_{i=1}^{n}(Y_i - (1, Z_i)\begin{pmatrix}\alpha \\ \beta\end{pmatrix})'(Y_i - (1, Z_i)\begin{pmatrix}\alpha \\ \beta\end{pmatrix})\Big] \\
&\quad -\frac{n}{2}\ln|R| - \frac{1}{2}\text{trace}\Big[R^{-1}\sum_{i=1}^{n} Z_i'Z_i\Big], \qquad (2)
\end{aligned}
$$

which gives a set of sufficient statistics for θ as follows:

$$
S_{yy} = \sum_{i=1}^{n} Y_i'Y_i, \quad S_{z^*y} = \sum_{i=1}^{n}\begin{pmatrix}1 \\ Z_i'\end{pmatrix} Y_i, \quad \text{and} \quad S_{z^*z^*} = \sum_{i=1}^{n}\begin{pmatrix}1 \\ Z_i'\end{pmatrix}(1, Z_i).
$$

EM

E-step: Compute the expected values of the sufficient statistics $\hat{S}_{yy} = \mathrm{E}(S_{yy}|Y_{\mathrm{obs}}, \hat{\theta})$, $\hat{S}_{z^*y} = \mathrm{E}(S_{z^*y}|Y_{\mathrm{obs}}, \hat{\theta})$, and $\hat{S}_{z^*z^*} = \mathrm{E}(S_{z^*z^*}|Y_{\mathrm{obs}}, \hat{\theta})$ from the joint distribution of (Y_i, Z_i) given $\theta = \hat{\theta}$:

$$(Y_i, Z_i)|\theta \sim \mathrm{N}_{p+q}\Big((\alpha, 0), \begin{bmatrix} \beta' R\beta + \mathrm{Diag}(\sigma^2) & \beta' R \\ R\beta & R \end{bmatrix}\Big)$$

for $i = 1, \ldots, n$. This can be accomplished using, for example, the Gaussian sweep operator to compute the conditional expectation and covariance matrix,

$$(\hat{Y}_i, \hat{Z}_i) \quad \text{and} \quad \hat{V} = \begin{bmatrix} \hat{V}_{Y_i', Y_i} & \hat{V}_{Y_i', Z_i} \\ \hat{V}_{Z_i', Y_i} & \hat{V}_{Z_i', Z_i} \end{bmatrix},$$

of (Y_i, Z_i) given $Y_{i,\mathrm{obs}}$ and $(\alpha, \beta, \sigma^2)$, where $\hat{Y}_{i,\mathrm{obs}} = Y_{i,\mathrm{obs}}$ and the elements of $\hat{V}$ with at least one index corresponding to the observed components $Y_{i,\mathrm{obs}}$ are zero; then

$$\hat{S}_{yy} = \sum_i \Big(\hat{Y}_i'\hat{Y}_i + \hat{V}_{Y_i', Y_i}\Big), \qquad \hat{S}_{z^*y} = \sum_i \begin{pmatrix} \hat{Y}_i \\ \hat{Z}_i'\hat{Y}_i + \hat{V}_{Z_i', Y_i} \end{pmatrix},$$

and

$$\hat{S}_{z^*z^*} = \sum_i \begin{pmatrix} 1 & \hat{Z}_i \\ \hat{Z}_i' & \hat{Z}_i'\hat{Z}_i + \hat{V}_{Z_i', Z_i} \end{pmatrix}.$$

M-step: This is, in principle, the same as the M-step of the EM algorithm for factor analysis from complete observations when the first component of $(1, Z_i)$ in Equation (2) is, computationally, viewed as a factor, and the intercept constants α are viewed as the corresponding (unrestricted) factor loadings. To be more specific, for each $j = 1, \ldots, p$, let $\mathcal{S}_j$ be the set of the indexes of all the nonzero factors for the outcome variable y_j, including index 0 corresponding to the intercept with "factor loading" α_j. We then obtain $\hat{\alpha}_j$, $(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{q,j})$, and $(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$ by applying the sweep operator as follows:

$$\begin{bmatrix} \hat{S}_{yy} & * \\ \hat{S}_{z^*y} & \hat{S}_{z^*z^*} \end{bmatrix} \underset{\Longrightarrow}{\mathrm{SWP}[\mathcal{S}_j + 1 + p]} \begin{bmatrix} \tilde{\Sigma}_{(p\times p)} & * \\ \tilde{\alpha}_{(1\times p)} & * \\ \tilde{\beta}_{(q\times p)} & * \end{bmatrix},$$

where the (j, j)th element of the $(p \times p)$ matrix $(1/n)\tilde{\Sigma}$ is $\hat{\sigma}_j^2$, the jth element of $\tilde{\alpha}$ is $\hat{\alpha}_j$, and $\hat{\beta}_{k,j} = 0$ if, *a priori*, $\beta_{k,j}$ is zero, and otherwise $\hat{\beta}_{k,j}$ is the (k, j)th element, $\tilde{\beta}_{k,j}$, of the $(q \times p)$ matrix $\tilde{\beta}$.

There are different versions of ECME for factor analysis from incomplete data. Here we describe the version of ECME that most closely parallels the ECME algorithm already described for the complete-data case in Section 3.

ECME
E-step: the same as the E-step of the previous EM.
CM-step 1: the same as the M-step of EM for parameters β and R.
CM-step 2: find α to maximize the constrained actual likelihood given β, R, and σ^2; this has a closed-form solution as described below.
CM-step 3: find σ^2 to maximize the constrained actual likelihood given α, β, and R; this can be done using Newton-Raphson as described below.

For a fully observed vector Y, we have from the model in Section 2:

$$Y|\theta \sim N(\alpha, \beta' R\beta + \mathrm{Diag}(\sigma^2)).$$

Let $I_{i,\mathrm{obs}}$ be the set of the indexes of the observed components of Y_i for $i = 1, \ldots, n$, $\Psi = \beta' R\beta + \mathrm{Diag}(\sigma^2)$, $\alpha_{I_{i,\mathrm{obs}}}$ be the components of α corresponding to the observed components of Y_i, and $\Psi_{[I_{i,\mathrm{obs}}, I_{i,\mathrm{obs}}]}$ be the sub-matrix of Ψ whose row and column indexes correspond to the observed components of Y_i. Then we have

$$Y_{i,\mathrm{obs}} \sim N_p(\alpha_{I_{i,\mathrm{obs}}}, \Psi_{[I_{i,\mathrm{obs}}, I_{i,\mathrm{obs}}]})$$

for $i = 1, \ldots, n$, and the actual log-likelihood function is:

$$L_{\mathrm{obs}}(\theta|Y_{\mathrm{obs}}) = -\frac{1}{2}\sum_{i=1}^{n}\Big(\ln|\Psi_{[I_{i,\mathrm{obs}}, I_{i,\mathrm{obs}}]}| + \mathrm{trace}\Big[\Psi^{-1}_{[I_{i,\mathrm{obs}}, I_{i,\mathrm{obs}}]}(Y_{i,\mathrm{obs}} - \alpha_{I_{i,\mathrm{obs}}})'(Y_{i,\mathrm{obs}} - \alpha_{I_{i,\mathrm{obs}}})\Big]\Big). \tag{3}$$

Let $A^{(i)}$ be the $(p \times p)$ matrix with (j,k)th element equal to the corresponding element of $\Psi^{-1}_{[I_{i,\mathrm{obs}}, I_{i,\mathrm{obs}}]}$ if both the jth and the kth components of Y_i are observed, and zero otherwise, and let $B^{(i)} = A^{(i)}(Y_i - \alpha)'(Y_i - \alpha)A^{(i)}$. Note that $B^{(i)}$ does not depend on $Y_{i,\mathrm{mis}}$.

From (3) we obtain

$$\frac{\partial L_{\mathrm{obs}}(\theta|Y_{\mathrm{obs}})}{\partial \alpha} = \sum_{i=1}^{n} A^{(i)}(Y_i - \alpha)',$$

$$\frac{\partial L_{\mathrm{obs}}(\theta|Y_{\mathrm{obs}})}{\partial \alpha' \partial \alpha} = -\sum_{i=1}^{n} A^{(i)},$$

$$\frac{\partial L_{\mathrm{obs}}(\theta|Y_{\mathrm{obs}})}{\partial \sigma_j^2} = -\frac{1}{2}\sum_{i=1}^{n}\Big(A^{(i)}_{j,j} - B^{(i)}_{j,j}\Big),$$

and

$$\frac{\partial^2 L_{\text{obs}}(\theta|Y_{\text{obs}})}{\partial\sigma_j^2\partial\sigma_k^2} = \frac{1}{2}\sum_{i=1}^{n} A_{j,k}^{(i)}\Big(A_{j,k}^{(i)} - 2B_{j,k}^{(i)}\Big)$$

for $\sigma_j^2 > 0$ and $\sigma_k^2 > 0$ $(j,k = 1,\ldots,n)$. As a result, we have

$$\hat{\alpha} = \Big(\sum_{i=1}^{n} A^{(i)}\Big)^{-1}\Big(\sum_{i=1}^{n} A^{(i)}Y_i\Big),$$

where the missing components of Y_i can be replaced with any values because the coefficients $A^{(i)}$ for the missing components $Y_{i,\text{mis}}$ are zero. This gives CM-step 2.

For CM-step 3, the transformation $\delta_j = \ln\sigma_j^2$, that is,

$$\sigma_j^2 = \exp(\delta_j) \qquad (j = 1,\ldots,p), \tag{4}$$

is useful especially when some of the uniquenesses are close to or equal to zero. For the transformation (4) we have

$$\frac{\partial L_{\text{obs}}(\theta|Y_{\text{obs}})}{\partial\delta_j} = -\frac{\sigma_j^2}{2}\sum_{i=1}^{n}\Big(A_{j,j}^{(i)} - B_{j,j}^{(i)}\Big)$$

and

$$\frac{\partial^2 L_{\text{obs}}(\theta|Y_{\text{obs}})}{\partial\delta_j\partial\delta_k} = \begin{cases} \frac{\sigma_j^2\sigma_k^2}{2}\sum_{i=1}^{n} A_{j,k}^{(i)}\Big(A_{j,k}^{(i)} - 2B_{j,k}^{(i)}\Big), & \text{if } j \neq k, \\ \frac{\sigma_j^2\sigma_k^2}{2}\sum_{i=1}^{n} A_{j,k}^{(i)}\Big(A_{j,k}^{(i)} - 2B_{j,k}^{(i)}\Big) + \frac{\partial L_{\text{obs}}(\theta|Y_{\text{obs}})}{\partial\delta_j}, & \text{if } j = k, \end{cases}$$

which provides the gradient and the Hessian for Newton-Raphson for δ, and thus gives CM-step 3 for σ^2. Again, it is wise to check that Newton-Raphson actually increases the constrained likelihood because Newton-Raphson does not guarantee (monotone) convergence.

5. Numerical Examples

5.1. Factor analysis from complete observations

We applied both EM and ECME to the same data and model as used in Rubin and Thayer (1982) with $p = 9$,

$$C_{yy} = \begin{bmatrix} 1.0 & 0.554 & 0.227 & 0.189 & 0.461 & 0.506 & 0.408 & 0.280 & 0.241 \\ & 1.0 & 0.296 & 0.219 & 0.479 & 0.530 & 0.425 & 0.311 & 0.311 \\ & & 1.0 & 0.769 & 0.237 & 0.243 & 0.304 & 0.718 & 0.730 \\ & & & 1.0 & 0.212 & 0.226 & 0.291 & 0.681 & 0.661 \\ & & & & 1.0 & 0.520 & 0.514 & 0.313 & 0.245 \\ & & & & & 1.0 & 0.473 & 0.348 & 0.290 \\ & & & & & & 1.0 & 0.374 & 0.306 \\ & & & & & & & 1.0 & 0.672 \\ & & & & & & & & 1.0 \end{bmatrix},$$

$q = 4$, $R = I$, and, *a priori* zero factor loadings on factor-score 4 for variables 1 — 4 and zero factor loadings on factor-score 3 for variables 5 — 9. Also we used the same starting values for both EM and ECME:

$$\left(\beta^{(0)}\right)' = \begin{bmatrix} 0.5954912 & -0.4893347 & -0.3848925 & 0.0000000 \\ 0.6449102 & -0.4408213 & -0.3555598 & 0.0000000 \\ 0.7630006 & 0.5053083 & -0.0535340 & 0.0000000 \\ 0.7163828 & 0.5258722 & 0.0219100 & 0.0000000 \\ 0.6175647 & -0.4714808 & 0.0000000 & 0.1931459 \\ 0.6464100 & -0.4628659 & 0.0000000 & 0.4606456 \\ 0.6452737 & -0.3260013 & 0.0000000 & -0.3622682 \\ 0.7868222 & 0.3690580 & 0.0000000 & 0.0630371 \\ 0.7482302 & 0.4326963 & 0.0000000 & 0.0431256 \end{bmatrix},$$

which is created from the spectral decomposition of C_{yy}, and $\left(\sigma_i^2\right)^{(0)} = 10^{-8}$ for $i = 1, \ldots, 9$.

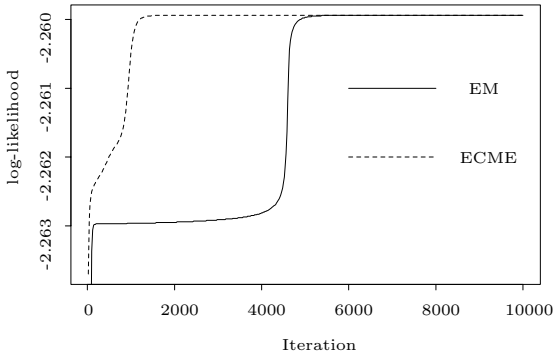

Figure 1. Convergence of EM (solid line) and ECME (dashed line) for the complete-data numerical example; displayed are increases in likelihood.

In this example, ECME converges much faster than EM, as we can see from Figures 1 and 2, where the corresponding log-likelihood function and the estimates of $\sigma^2 = (\sigma_1^2, \ldots, \sigma_9^2)$ are displayed. Specifically, here ECME converges faster than EM by a factor of five in number of iterations. Without any attempt to optimize code, on a SPARC station 2, ECME took about 25% less CPU time than EM (EM — 29 sec.; ECME — 22 sec.). More important than any savings in computer time using ECME rather than EM, at least in this example, is the easier assessment of convergence using ECME rather than EM. From Figures 1 and 2, we see that EM converges so slowly that it is difficult, at many points in the

iterative sequence, to detect changes, which can lead to stopping before actual convergence (e.g., after 1000-2000 iterations), but that this uncertain detectability of convergence does not happen with ECME. Thus when combined with its monotone convergence, ECME certainly appears to be an attractive alternative to EM for ML factor analysis with complete data.

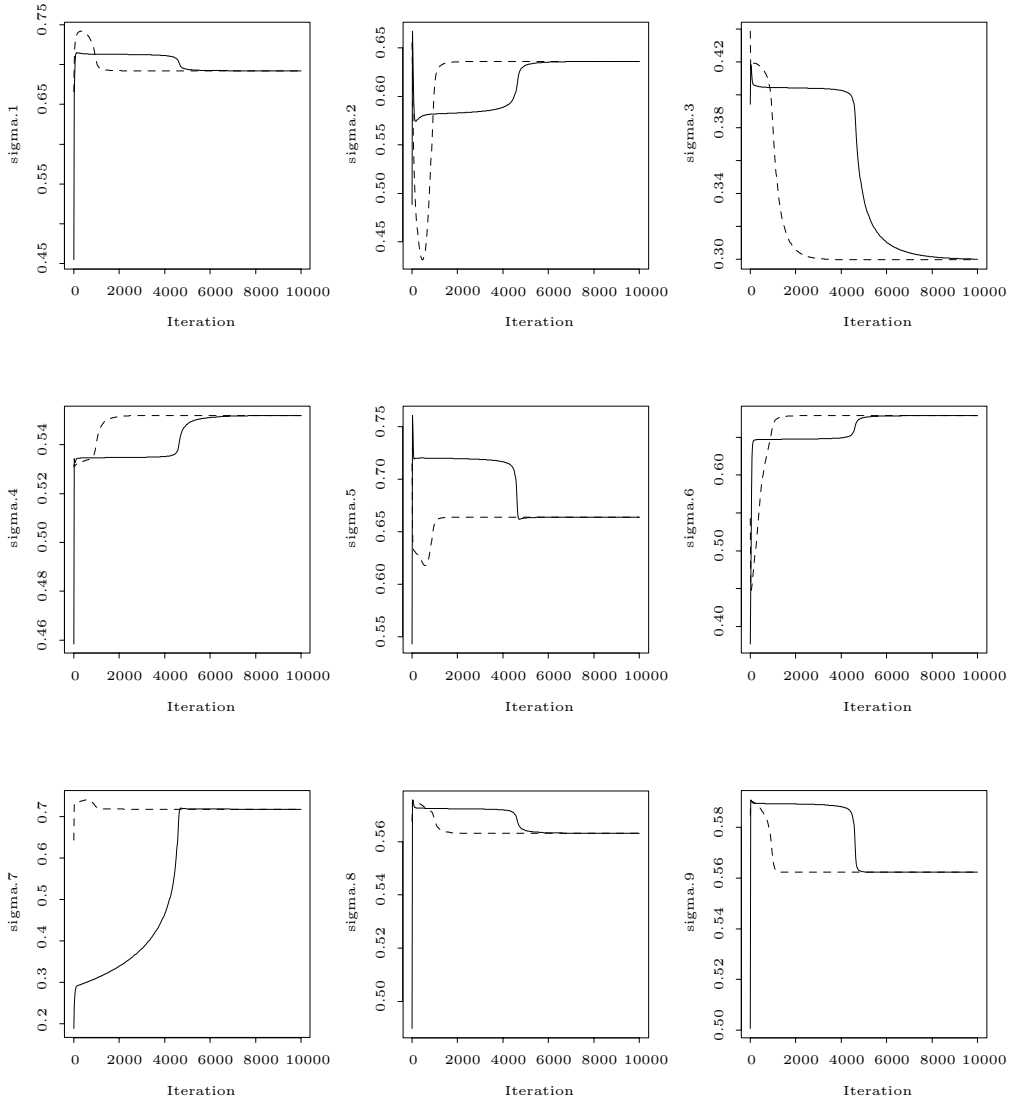

Figure 2. Convergence of EM (solid line) and ECME (dashed line) for the complete-data numerical example; displayed are the convergence of components of the uniquenesses, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_9^2)$ with sigma.1 $= \sigma_1, \ldots$, and sigma.9 $= \sigma_9$.

5.2. Factor analysis from incomplete observations

To illustrate the EM and ECME algorithms for ML estimation of factor analysis from incomplete data, we consider the example displayed in Table 1 (Efron (1994)), which is a random sample of 20 observations of five variables with missing values represented by "?", from Mardia, Kent and Bibby (1979), p. 2-5. The variables are examination grades from five courses, Mechanics, Vectors, Algebra, Analysis, and Statistics. The examinations on the first two courses used closed-book examinations and the other three used open-book.

Table 1. Marks in examinations with missing values indicated by the symbol "?" (Efron (1994))

	Closed book exams		Open book exams		
Student	Mechanics	Vectors	Algebra	Analysis	Statistics
2	53	61	72	64	73
9	30	69	50	52	45
16	17	53	57	43	51
9	30	69	50	52	45
16	17	53	57	43	51
18	48	38	41	44	33
20	30	34	43	46	18
1	?	63	65	70	63
3	51	67	65	65	?
4	?	69	53	53	53
5	?	69	61	55	45
6	?	49	62	63	62
7	44	61	52	62	?
8	49	41	61	49	?
10	?	59	51	45	51
11	?	40	56	54	?
12	42	60	54	49	?
13	?	63	53	54	?
14	?	55	59	53	?
15	?	49	45	48	?
17	39	46	46	32	?
19	46	40	47	29	?
21	?	30	32	35	21
22	?	26	15	20	?

As in Efron (1994), we assume the data in Table 1 are i.i.d. $N_5(\mu, \Psi)$ and that the missing-data mechanism is ignorable (Rubin (1976)). It has been noticed (Rubin (1994)) that ML estimation of Ψ without any restrictions beyond positive definiteness leads to a singular covariance matrix; this issue is discussed further

in Liu and Rubin (1998). Here, we assume that the covariance matrix Ψ has a pattern represented by a factor analysis model; we consider three such models:
Model I: $q = 1$ with no *a priori* zero loadings.
Model II: $q = 2$. One factor has no *a priori* zero loadings and the other has zero loadings for the two closed-book examinations.
Model III: $q = 2$. One factor has no *a priori* zero loadings and the other has zero loadings for the three open-book examinations.

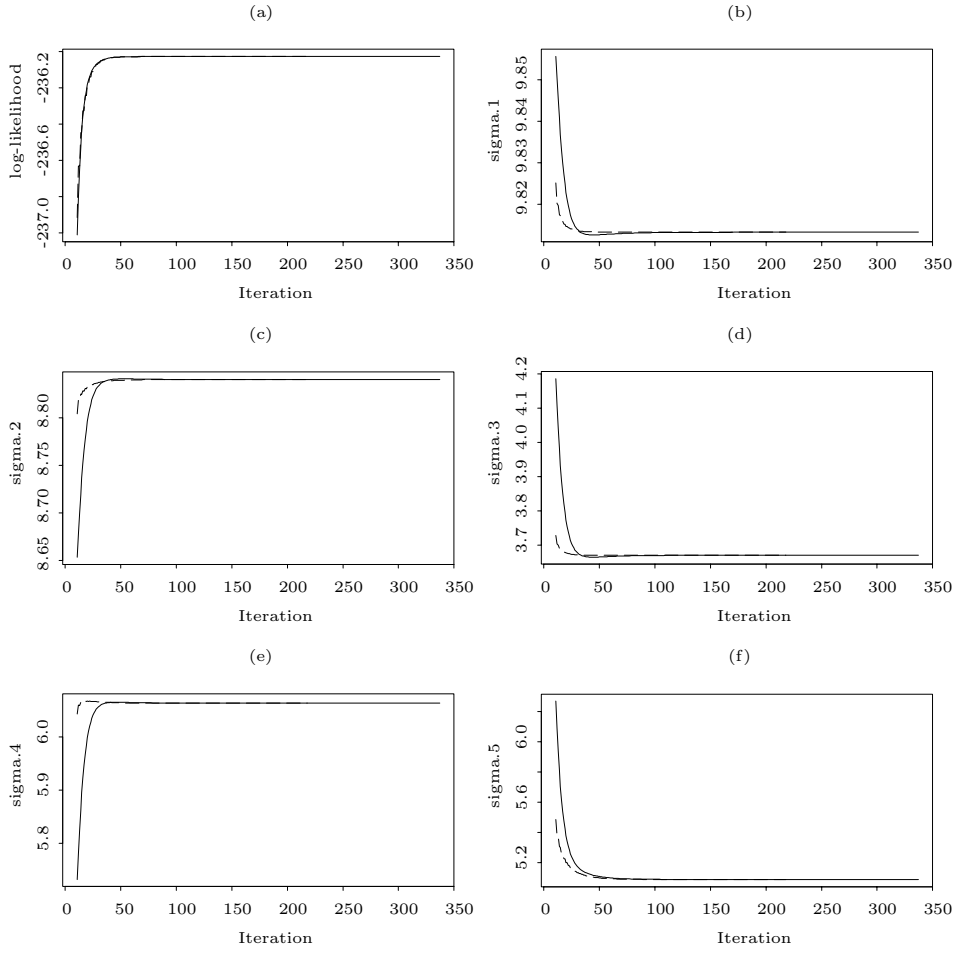

Figure 3. Convergence of EM (solid line) and ECME (dashed line) for the incomplete-data numerical example with Model I; displayed are the convergence of likelihood and components of the uniquenesses, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_5^2)$ with sigma.1 $= \sigma_1$, ..., and sigma.5 $= \sigma_5$.

For Model I, with starting values for α equal to the observed averages for each variable $(40.82, 51.91, 51.82, 49.32, 46.82)$, $\beta = (1, \ldots, 1)$, and $\sigma^2 = (1, \ldots, 1)$,

both EM and ECME converged to the same stationary point (Table 2), with log-likelihood -236.03. The numbers of iterations of EM and ECME are, respectively, 337 and 220. Without optimization of the computer codes, ECME took 3.7 sec. of CPU time and EM took 1.1 sec. The convergence of the loglikelihood and the uniquenesses is displayed in Figure 3. For this relatively simple example without optimization of the computer codes, there appears to be minimal advantage to using ECME rather than EM.

Table 2. ML estimates of Model I

	Mechanics	Vectors	Algebra	Analysis	Statistics
$\hat{\alpha}$	40.51	51.91	51.82	49.32	44.36
$\hat{\beta}$	4.48	9.64	11.45	10.48	16.82
$\hat{\sigma}^2$	96.30	78.15	13.47	36.76	25.90

The upper-left 2×2 submatrix of the covariance matrix of the observations according to Model II is

$$\Psi = \begin{bmatrix} \beta_{1,1}^2 + \beta_{2,1}^2 + \sigma_1^2 & \beta_{1,1}\beta_{1,2} + \beta_{2,1}\beta_{2,2} \\ \beta_{1,1}\beta_{1,2} + \beta_{2,1}\beta_{2,2} & \beta_{1,2}^2 + \beta_{2,2}^2 + \sigma_2^2 \end{bmatrix}.$$

Given $(\beta_{1,1}, \ldots, \beta_{1,5})$, there is a problem of nonidentifiability for $(\beta_{2,1}, \beta_{2,2})$ and (σ_1^2, σ_2^2) as is evident in the the upper-left (2×2) sub-matrix of Ψ; that is, at most three parameters, $c_{1,1} = \beta_{2,1}^2 + \sigma_1^2$, $c_{1,2} = \beta_{2,1}\beta_{2,2}$, and $c_{2,2} = \beta_{2,2}^2 + \sigma_2^2$, rather than the four, can be identified. We use the ML estimates of α, β, and $(\sigma_3^2, \sigma_4^2, \sigma_5^2)$ in Model I as the starting points for the corresponding parameters in Model II, $\sigma_1^2 = 96.30/2$, $\sigma_2^2 = 78.15/2$, $\beta_{2,1} = 6.94 = \sqrt{\sigma_1^2}$, and $\beta_{2,2} = \pm 6.25 = \sqrt{\sigma_2^2}$, which are chosen in such a way that $c_{1,1}$ and $c_{2,2}$ start with the ML estimates of σ_1^2 and σ_2^2 from Model I and $c_{1,2} = \pm 0.5\sqrt{c_{1,1}c_{2,2}}$. With these two sets of starting points, both EM and ECME converge to stationary points that have the same value of the actual likelihood function with subspace represented by the following stationary point (Table 3), with log-likelihood -235.36. The numbers of iterations of EM and ECME are, respectively, 299 (241) and 176 (148), where the values in parentheses correspond to the starting points with negative $\beta_{2,2}$. ECME took 3.3 sec. of CPU time and EM took 1.2 sec. The convergence of the loglikelihood and the diagonal elements of Ψ is displayed in Figure 4. With this example as well, there appears to be little advantage to using ECME rather than EM.

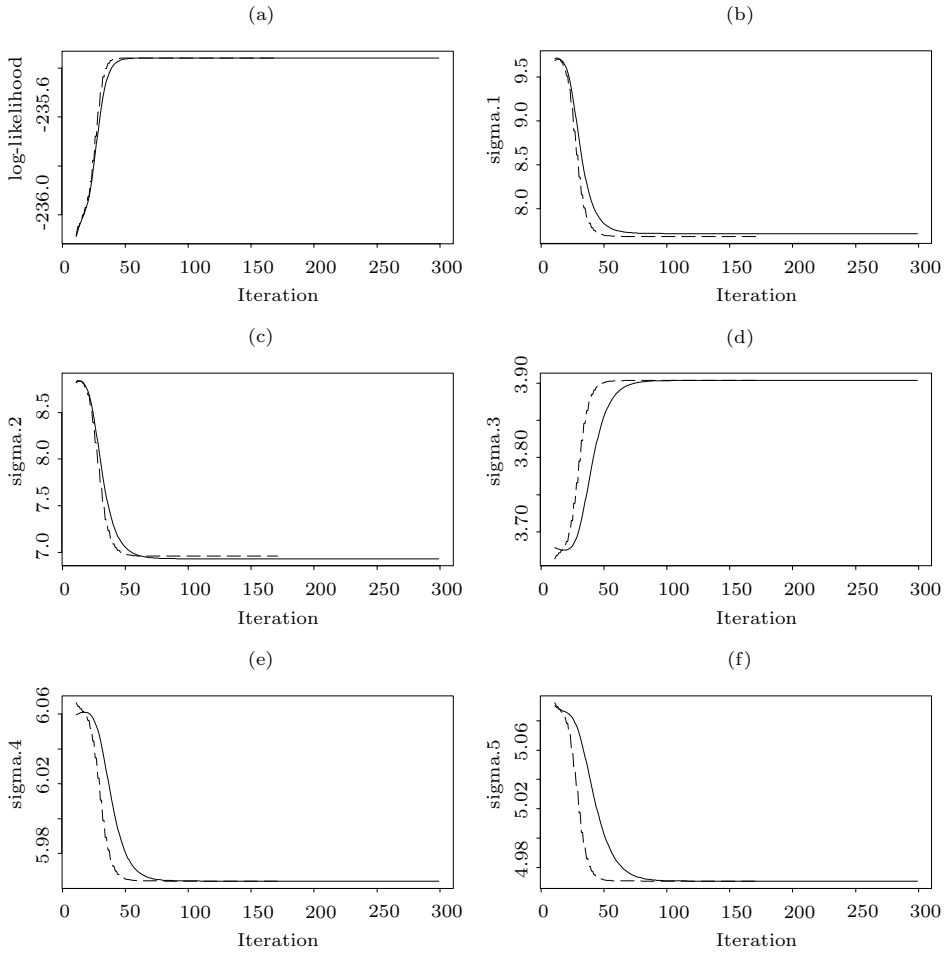

Figure 4. Convergence of EM (solid line) and ECME (dashed line) for the incomplete-data numerical example with Model II; displayed are the convergence of likelihood and components of the uniquenesses, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_5^2)$ with sigma.1 $= \sigma_1$, ..., and sigma.5 $= \sigma_5$.

Table 3. ML estimates of Model II

	Mechanics	Vectors	Algebra	Analysis	Statistics
$\hat{\alpha}$	40.20	51.91	51.82	49.32	44.48
$\hat{\beta}$	4.80	9.73	11.37	10.54	16.85
	6.07	-5.27	0.00	0.00	0.00
$\hat{\sigma}^2$	59.04	48.45	15.24	35.57	24.71

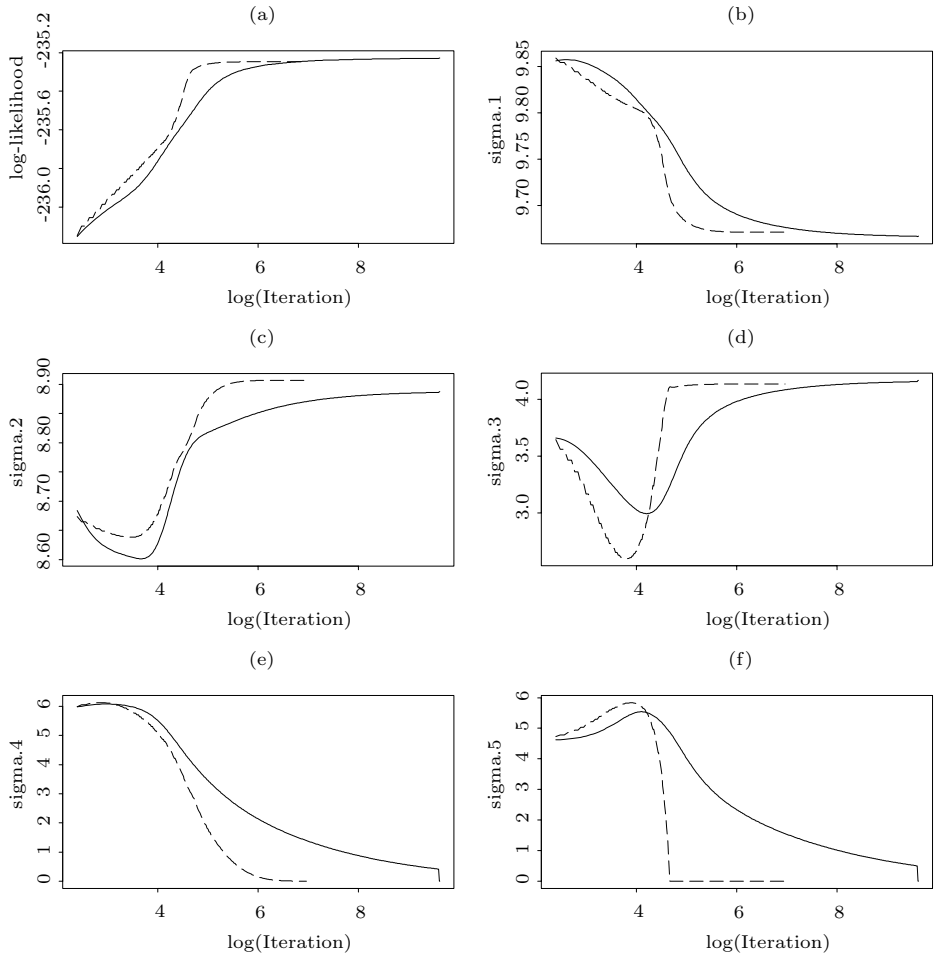

Figure 5. Convergence of 15,000 iterations of EM followed by ECME (solid line) and straight ECME (dashed line) for the incomplete-data numerical example with Model III; displayed are the convergence of likelihood and components of the uniquenesses, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_5^2)$ with sigma.1 $= \sigma_1$, ..., and sigma.5 $= \sigma_5$. Note the use of the logarithmic scale for the number of iterations.

Unlike Model II, Model III does not appear to have a problem of nonidentifiability even though Model III has one more parameter than Model II. As with Model II, for model III we use the following starting points: the ML estimates of α, β, σ_1^2, and σ_2^2 from Model I, $\sigma_3^2 = 13.47/2$, $\sigma_4^2 = 36.76/2$, $\sigma_5^2 = 25.90/2$, $\beta_{2,3} = \sqrt{\sigma_3^2} = 2.60$, $\beta_{2,4} = \pm\sqrt{\sigma_4^2} = \pm 4.29$, and $\beta_{2,5} = \pm\sqrt{\sigma_5^2} = \pm 3.60$. ECME takes about 900 iterations on average to converge with 35 sec. of CPU time, whereas EM effectively would run "forever" and never actually converge due to

the zero uniquenesses. Hence we switched to ECME after 15,000 EM iterations, and ECME then took 37 iterations on average to converge using an average total of 57 sec. of CPU time. All the EM sequences with ECME speeding lead to the same estimates with log-likelihood -235.23. The ML estimates of the parameters found are as follows:

ML estimates of Model III

	Mechanics	Vectors	Algebra	Analysis	Statistics
$\hat{\alpha}$	40.74	51.91	51.82	49.32	44.79
$\hat{\beta}$	4.79	9.59	11.17	11.33	16.34
	0.00	0.00	1.52	-4.24	5.50
$\hat{\sigma}^2$	93.46	78.98	17.36	0.00	0.00

The straight ECME sequences, however, lead to slightly different stationary points with log-likelihood about -235.26. For more discussion on multiple modes, see Rubin and Thayer (1982, 1983). In this example, there is an enormous benefit to using ECME rather than EM, at least after EM has become "stuck".

6. Practical Conclusion

The conclusion from the examples in Section 5 is clear: especially when zero uniqueness are possible, EM cannot be relied on to reach the MLE, whereas ECME or an initial run of EM followed by ECME can work in this difficult case. In practice, initial inexpensive iterations of EM followed by iterations of ECME to convergence appears to be an effective strategy. These conclusions could shift to a recommendation to use only ECME if the code for ECME were optionized.

Other practical issues that should be investigated include the fitting of factor analysis models with more complicated data structures, for example, clustered observations (Longford and Muthén (1992)) with incomplete data, and the application of very recent extensions of ECME, AECM (Meng and van Dyk (1997)) and PX-EM (Liu, Rubin and Wu (1998)). These extensions may be quite important in fitting other hidden (or latent) variables models that are common in educational testing contexts.

Acknowledgements

We thank the editors and referees for many helpful comments, and acknowledge NSF grants SES-9207456 and DMS-9404479 for partially supporting this work.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley, New York.

Basilevsky, A. (1994). *Statistical factor Analysis and Related Methods.* John Wiley, New York.

Bentler, P. M. and Tanaka, J. S. (1983). Problems with EM algorithms for ML factor analysis. *Psychometrika* **48**, 247-251.

Clarke, M. R. B. (1970). A rapidly convergent method for maximum-likelihood factor analysis. *British J. Math. Statist. Psych.* **23**, 43-52.

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

Efron, B. (1994). Missing data, imputation, and the bootstrap (with discussion). *J. Amer. Statist. Assoc.* **89**, 463-479.

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

Jennrich, R. I. and Robinson, S. M. (1969). A newton-raphson algorithm for maximum likelihood factor analysis. *Psychometrika* **34**, 111-123.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**, 443-482.

Jöreskog, K. G. and Sörbom, D. (1988). LISREL-7: A guide to the program and applications (2nd edition). Chicago: SPSS.

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

Liu, C. (1996). Maximum likelihood estimation of factor analysis with fixed and random effects from incomplete data, Technical report, Statistics Research Department, Bell Laboratories, Lucent Technologies.

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.

Liu, C. and Rubin, D. B. (1998). Estimation of ellipsoidal distributions from singular incomplete data. Technical report. Bell-Labs.

Liu, C., Rubin, B. D. and Wu, Y. (1998). Parameter expansion to accelerate EM – the PX-EM algorithm. *Biometrika* **85**, to appear.

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

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, New York.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Amer. Statist. Assoc.* **86**, 899-909.

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

Meng, X.-L. and van Dyk, D. (1997). The *EM* algorithm — an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* **58**, 525-543.

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

Rubin, D. B. (1994). Comment on "Missing data, imputation, and the bootstrap". *J. Amer. Statist. Assoc.* **89**, 475-478.

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

Rubin, D. B. and Thayer, D. T. (1983). More on EM for ML factor analysis. *Psychometrika* **48**, 253-257.

van Dyk, D. A., Meng, X.-L. and Rubin, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statist. Sinica* **5**, 55-75.

Statistics Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ 07974, U.S.A.
E-mail: liu@research.bell-labs.com

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A
E-mail: rubin@stat.harvard.edu

(Received July 1996; accepted June 1997)