On the Global and Componentwise Rates of Convergence of the EM Algorithm

Xiao-Li Meng
Department of Statistics
University of Chicago
Chicago, Illinois 60637

and

Donald B. Rubin
Department of Statistics
Harvard University
Cambridge, Massachusetts 02138

Submitted by George P. H. Styan

Dedicated to Ingram Olkin

ABSTRACT

The EM algorithm is a very general and popular iterative algorithm in statistics for finding maximum-likelihood estimates in the presence of incomplete data. In the paper that defined and popularized EM, Dempster, Laird, and Rubin (1977) showed that its global rate of convergence is governed by the largest eigenvalue of the matrix of fractions of missing information due to incomplete data. It was also mentioned that componentwise rates of convergence can differ from each other when the fractions of information loss vary across different components of a parameter vector. In this article, using the well-known diagonability theorem, we present a general description on how and when the componentwise rates differ, as well as their relationships with the global rate. We also provide an example, a standard contaminated normal model, to show that such phenomena are not necessarily pathological, but can occur in useful statistical models.

1. BRIEF INTRODUCTION TO THE EM ALGORITHM

Since it was formally introduced by Dempster, Laird, and Rubin (1977, henceforth DLR), the EM algorithm has been widely applied to many problems that can be formulated as incomplete-data problems. The popular-

655 Avenue of the Americas, New York, NY 10010 0024-3795/94/$7.00

ity of EM for finding maximum-likelihood estimates and posterior modes arises from its simplicity in implementation, stability in convergence (e.g., monotone increases in objective functions), and applicability in practice; in fact some problems now solved were considered intractable before EM. Its application can be found almost in any field that encounters statistical analysis with incomplete data; Meng and Pedlow's (1992) recent bibliographic review provides over 1000 EM related articles spanning over more than 270 journals, approximately 85% of which are nonstatistical journals.[1] It has also stimulated other powerful computational methods in statistics, as discussed in Meng and Rubin (1992, 1993).

The idea behind EM is quite simple and intuitive. It comes from a quite old ad hoc idea for handling missing data: (i) if the missing values were known, then complete-data techniques could be applied to estimate the unknown parameters of the underlying model, and (ii) if the model parameters were known, then the missing values could be imputed according to the model. An iterative procedure thus arises—iterating between (i) and (ii) until no changes occur in the parameter estimates or imputed values (see e.g., Healy and Westmacott, 1956, for the analysis of variance). The key contribution of EM, in contrast to its ad hoc predecessors, is to recognize that the correct procedure is not to impute the individual missing values, but rather to impute the complete-data sufficient statistics, since maximum-likelihood estimates depend on data only through sufficient statistics (e.g., Cox and Hinkley, 1974), and these are not necessarily linear in the data. In the more general cases with no (useful) sufficient statistics, the correct procedure is to impute the complete-data log-likelihood function itself.

The mathematical description of EM can be summarized briefly as follows. Let $f(Y \mid \theta)$ be the density of complete data Y that would occur in the absence of missing values, where $\theta = (\theta_1, \ldots, \theta_d)$ is a $1 \times d$ parameter vector with parameter space Θ. We write $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} denotes the observed values and Y_{mis} denotes the missing values. We are interested in finding $\hat{\theta}$ that maximizes

$$L(\theta \mid Y_{obs}) \equiv f(Y_{obs} \mid \theta) = \int f(Y_{obs}, Y_{mis} \mid \theta)\, dY_{mis}, \qquad (1.1)$$

that is, the maximum-likelihood estimate (MLE) for θ based on the observed data, Y_{obs}. Because of the integration in (1.1), the required maximization is typically substantially more difficult than the maximization of the complete-

[1] A preliminary EM bibliography is available upon request. This article only lists references that are directly related to it.

data likelihood, $L(\theta \mid Y) \equiv f(Y \mid \theta)$. The EM algorithm converts the difficult incomplete-data maximization into a sequence of easier complete-data maximizations.

Starting from an initial guess $\theta^{(0)}$, each iteration of EM consists of an *expectation* step and a *maximization* step. At the $(t+1)$st iteration, $t = 0, 1, \ldots$, the E-step finds the conditional expectation of the complete-data log-likelihood given the observed data and $\theta = \theta^{(t)}$:

$$Q(\theta \mid \theta^{(t)}) = \int \log L(\theta \mid Y)\, f\left(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta = \theta^{(t)}\right) dY_{\text{mis}}, \qquad (1.2)$$

where $f(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta) = f(Y \mid \theta)/f(Y_{\text{obs}} \mid \theta)$ is the conditional density of Y_{mis} given Y_{obs} and θ. The M-step then determines $\theta^{(t+1)}$ by maximizing this expected log-likelihood:

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) \qquad \text{for all} \quad \theta \in \Theta. \qquad (1.3)$$

As shown in DLR, from any starting point inside Θ, the resulting iterative sequence $\{\theta^{(t)},\ t \geq 0\}$ monotonically increases $L(\theta \mid Y_{\text{obs}})$, a feature that is generally viewed as providing stable convergence. Also, under mild conditions in practice, EM converges to an MLE; see DLR, Boyles (1983), and Wu (1983) for convergence conditions.

The iterative procedure given by (1.2) and (1.3) implicitly defines a mapping $\theta \to M(\theta)$ from the parameter space Θ to itself such that

$$\theta^{(t+1)} = M(\theta^{(t)}) \qquad \text{for} \quad t = 0, 1, \ldots . \qquad (1.4)$$

Assuming that $\theta^{(t)}$ converges to the MLE $\hat{\theta}$ and that $M(\theta)$ is differentiable at $\hat{\theta}$, a simple Taylor expansion yields

$$\theta^{(t+1)} - \hat{\theta} = (\theta^{(t)} - \hat{\theta})\, DM(\hat{\theta}) + O(\|\theta^{(t)} - \hat{\theta}\|^2), \qquad (1.5)$$

where

$$DM(\theta) = \left(\frac{\partial M_j(\theta)}{\partial \theta_i} \right) \qquad (1.6)$$

is the $d \times d$ Jacobian matrix for $M(\theta) = (M_1(\theta), \ldots, M_d(\theta))$, and $\|\cdot\|$ is the usual Euclidean norm. Thus, in a neighborhood of $\hat{\theta}$, the EM algorithm is essentially a linear iteration with iteration matrix $DM(\hat{\theta})$, since $DM(\hat{\theta})$ is typically nonzero.

2. THE RATE OF CONVERGENCE OF THE EM ALGORITHM

The performance of an iterative algorithm is commonly measured by its order and its rate of convergence. It was seen in Section 1 that the order of EM is generally linear, so we will focus on the rate of convergence for linear iterations. For EM (and for other linear iterating algorithms), the global rate of convergence is defined as

$$R = \lim_{t \to \infty} R^{(t)} \equiv \lim_{t \to \infty} \frac{\|\theta^{(t+1)} - \hat{\theta}\|}{\|\theta^{(t)} - \hat{\theta}\|}, \tag{2.1}$$

and the ith ($i = 1, \ldots, d$) componentwise rate of convergence is defined as

$$R_i = \lim_{t \to \infty} R_i^{(t)} \equiv \lim_{t \to \infty} \frac{\theta_i^{(t+1)} - \hat{\theta}_i}{\theta_i^{(t)} - \hat{\theta}_i}, \tag{2.2}$$

provided these limits exist. In the case $\theta_i^{(t)} \equiv \theta_i^{(t_0)}$ for all $t \geq t_0$ (≥ 1), we define $R_i = 0$. Such cases can happen, for example, when some components have no missing information, as in the bivariate normal example of Meng and Rubin (1991).

In view of (1.5), it is easy to see that the global rate of convergence of EM is governed by the spectral radius of $DM(\hat{\theta})$, which in this case is its largest eigenvalue (see Section 4). Under mild regularity conditions, DLR established an important identity between $DM(\hat{\theta})$ and the matrix of fractions of missing information. More specifically, after taking the second derivative of $L(\theta \mid Y)$ with respect to θ, we let

$$I_{oc}(\theta \mid Y_{obs}) = -\int \frac{\partial^2 \log L(\theta \mid Y)}{\partial\theta \cdot \partial\theta} f(Y_{mis} \mid Y_{obs}, \theta)\, dY_{mis}, \tag{2.3}$$

and similarly

$$I_{om}(\theta \mid Y_{obs}) = -\int \frac{\partial^2 \log f(Y_{mis} \mid Y_{obs}, \theta)}{\partial\theta \cdot \partial\theta} f(Y_{mis} \mid Y_{obs}, \theta)\, dY_{mis} \tag{2.4}$$

and

$$I_o(\theta \mid Y_{obs}) = -\frac{\partial^2 \log L(\theta \mid Y_{obs})}{\partial\theta \cdot \partial\theta}. \tag{2.5}$$

DLR showed that, if $Q(\theta \mid \theta^{(t)})$ is maximized by setting its first derivative equal to zero, then

$$DM(\hat{\theta}) = I_{\text{om}}(\hat{\theta} \mid Y_{\text{obs}}) I_{\text{oc}}^{-1}(\hat{\theta} \mid Y_{\text{obs}}) \equiv J(\hat{\theta} \mid Y_{\text{obs}}). \tag{2.6}$$

The right-hand side of (2.6) is the matrix of the so-called "fractions of missing information," because $I_{\text{om}}(\hat{\theta} \mid Y_{\text{obs}})$ measures the "missing information" (i.e., the loss of information due to the missing data), and $I_{\text{oc}}(\hat{\theta} \mid Y_{\text{obs}})$ measures the "complete information" (i.e., the information we would have if we had complete data). Because of the "missing-information principle" (Orchard and Woodbury, 1972; Meng and Rubin, 1991), which states that

$$I_{\text{o}}(\hat{\theta} \mid Y_{\text{obs}}) = I_{\text{oc}}(\hat{\theta} \mid Y_{\text{obs}}) - I_{\text{om}}(\hat{\theta} \mid Y_{\text{obs}}), \tag{2.7}$$

or in words,

observed information = complete information − missing information,

it follows that another expression for $J \equiv J(\hat{\theta} \mid Y_{\text{obs}})$ is

$$J = I - I_{\text{o}}(\hat{\theta} \mid Y_{\text{obs}}) I_{\text{oc}}^{-1}(\hat{\theta} \mid Y_{\text{obs}}).$$

Thus, the global rate of convergence of EM, R, is governed by the largest eigenvalue of J, which is less than 1 when $I_{\text{o}}(\hat{\theta} \mid Y_{\text{obs}}) > 0$, that is, when it is positive definite, which is a sufficient condition to guarantee that $\hat{\theta}$ is a (local) maximum-likelihood estimate. This condition will be used in the Lemma of Section 4.

It was also pointed out in DLR that the componentwise rates of convergence of EM, R_i $(i = 1, \ldots, d)$, can differ from each other because the fractions of missing information can vary across different components of θ. It is natural to ask how and when this can happen, and what the relationship is between the R_i's and R. One obvious case occurs when J is diagonal but not proportional to the identity matrix, which we show in the next section can indeed happen in practice. Then in Section 4, we give a complete answer to the general question. From a purely algebraic point of view, our results fall within the extensive literature on finding eigenvalues using the power method (e.g., Faddeev and Faddeeva, 1963).

3. AN EXAMPLE OF UNEQUAL R_i'S

Suppose $x_1, \ldots, x_n$ is a simple random sample from a univariate contaminated normal model

$$f(x \mid \mu, \sigma^2) = (1 - \pi) N(\mu, \sigma^2) + \pi N(\mu, \sigma^2/\lambda),$$

where $0 < \pi < 1$, $\lambda > 0$, and both π and λ are known. We are interested in finding the maximum-likelihood estimate of $\theta = (\mu, \sigma^2)$. The direct maximization of the likelihood is difficult because the actual density is a mixture of two densities.

As described in Chapter 10 of Little and Rubin (1987), this problem can be treated as an incomplete-data problem even though there are no missing data in the usual sense. Specifically, let

$$h(q) = \begin{cases} 1 - \pi & \text{if } q = 1, \\ \pi & \text{if } q = \lambda, \\ 0 & \text{otherwise;} \end{cases}$$

then $Y_{\text{obs}} = X = (x_1, \ldots, x_n)$ can be considered as a random sample from a population such that

$$x_l \mid \theta, q_l \overset{\text{ind}}{\sim} N(\mu, \sigma^2/q_l),$$

where the q_l's constitute an "unobserved" simple random sample from the density $h(q)$. We can thus apply EM treating $Y_{mis} = q = (q_1, \ldots, q_n)$ as missing data. Treating mixture models as missing-data problems is regarded as one of the contributions of DLR.

The implementation of EM here is quite straightforward, as described in Little and Rubin (1987, p. 210). The resultant EM mapping is given by

$$\mu^{(t+1)} = \frac{\sum_{l=1}^{n} w_l(\theta^{(t)}) x_l}{\sum_{l=1}^{n} w_l(\theta^{(t)})}, \tag{3.1}$$

and

$$(\sigma^2)^{(t+1)} = \frac{1}{n} \sum_{l=1}^{n} w_l(\theta^{(t)}) (x_l - \mu^{(t)})^2, \tag{3.2}$$

where

$$w_l(\theta) \equiv E(q_l \mid x_l, \theta) = \frac{1 - \pi + \pi\lambda^{3/2} \exp\{(1-\lambda)(x_l - \mu)^2/2\sigma^2\}}{1 - \pi + \pi\lambda^{1/2} \exp\{(1-\lambda)(x_l - \mu)^2/2\sigma^2\}}. \tag{3.3}$$

To compute the rate of convergence, we can either directly differentiate the EM mapping given by (3.1)–(3.3), or calculate the matrix of fractions of missing information using $I_{oc}(\hat{\theta} \mid X)$ and $I_{om}(\hat{\theta} \mid X)$ defined in (2.3) and (2.4), respectively. Let

$$S_m(\theta \mid X) = \frac{1}{n} \sum_{l=1}^{n} w_l(\theta) \left(\frac{x_l - \mu}{\sigma} \right)^m \quad \text{for} \quad m = 0, 1, \ldots \tag{3.4}$$

and

$$T_m(\theta \mid X) = \frac{1}{n} \sum_{l=1}^{n} v_l(\theta) \left(\frac{x_l - \mu}{\sigma} \right)^m \quad \text{for} \quad m = 0, 1, \ldots, \tag{3.5}$$

where $v_l(\theta) = \mathrm{Var}(q_l \mid x_l, \theta) = [w_l(\theta) - 1][\lambda - w_l(\theta)]$. Noting that $S_1(\hat{\theta} \mid X) = 0$ and $S_2(\hat{\theta} \mid X) = 1$, which are consequences of (3.1) and (3.2), one can verify

$$I_{oc}(\hat{\theta} \mid X) = \frac{n}{\hat{\sigma}^4} \begin{pmatrix} \hat{\sigma}^2 S_0(\hat{\theta} \mid X) & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \tag{3.6}$$

and

$$I_{om}(\hat{\theta} \mid X) = \frac{n}{\hat{\sigma}^4} \begin{pmatrix} \hat{\sigma}^2 T_2(\hat{\theta} \mid X) & T_3(\hat{\theta} \mid X)/2 \\ T_3(\hat{\theta} \mid X)/2 & T_4(\hat{\theta} \mid X)/4 \end{pmatrix}. \tag{3.7}$$

By the law of large numbers and the fact $\hat{\theta} \to \theta$ as $n \to \infty$, it is easy to show that, with probability one,

$$T_2(\hat{\theta} \mid X) \to T_2(\theta) \equiv E\left(\mathrm{Var}(q \mid x, \theta)\left(\frac{x-\mu}{\sigma}\right)^2\right),$$

$$T_3(\hat{\theta} \mid X) \to 0,$$

$$T_4(\hat{\theta} \mid X) \to T_4(\theta) \equiv E\left(\mathrm{Var}(q \mid x)\left(\frac{x-\mu}{\sigma}\right)^4\right),$$

and

$$S_0(\hat{\theta} \mid X) \to E(q) = \lambda\pi + (1-\pi).$$

It follows immediately that $J(\hat{\theta} \mid X)$, the matrix of fractions of missing information, converges to a diagonal matrix with diagonal elements

$$d_{11} = T_2(\theta)/E(q)$$

and

$$d_{22} = T_4(\theta)/2.$$

Therefore, when the ample size is large, the componentwise rate of convergence R_1 (for the mean μ) is equal to d_{11}, and R_2 (for the variance σ^2) is equal to d_{22}. Since $d_{11} \neq d_{22}$ in general (their values can be obtained via numerical integrations for a specific value of the parameter θ), the two components converge at different rates.

To illustrate these results numerically, we conducted a simulation with $n = 100$, $\pi = 0.5$, $\lambda = 0.5$, $\mu = 0$, and $\sigma^2 = 1$. The initial guess is set at an unbiased estimate of (μ, σ^2). A similar numerical example was used in Meng and Rubin (1991) to illustrate the *supplemented EM* (SEM) algorithm for computing the large-sample variance-covariance matrices associated with maximum-likelihood estimates found by EM. It is interesting to observe from Table 1 that the different-rate phenomenon can also occur with finite samples.

4. ALGEBRAIC RESULTS

We now answer the general question concerning global and componentwise rates of convergence of EM by studying the following linear iteration:

$$\varphi^{(t+1)} = \varphi^{(t)} J, \tag{4.1}$$

TABLE 1
EM ITERATIONS WITH UNEQUAL R_i'S

t	$\mu^{(t)}$	$\mu^{(t)} - \hat{\mu}$	$R_1^{(t)}$	$(\sigma^2)^{(t)}$	$(\sigma^2)^{(t)} - \hat{\sigma}^2$	$R_2^{(t)}$
0	−0.040021	0.019032	0.099236	1.014000	0.012115	0.137961
1	−0.057164	0.001889	0.100717	1.003557	0.001671	0.125666
2	−0.058863	0.000190	0.101899	1.002095	0.000210	0.126038
3	−0.059034	0.000019	0.103233	1.001911	0.000026	0.126302
4	−0.590051	0.000002	0.104778	1.001888	0.000003	0.126495
5	−0.059053	0.000000	0.106518	1.001885	0.000000	0.126619
6	−0.059053	0.000000	0.108323	1.001885	0.000000	0.126535

where $\varphi^{(t)} = \theta^{(t)} - \hat{\theta}$ and J is the matrix of fractions of missing information defined in (2.6). As a consequence of (1.5) and (2.6), this linear iteration has the same rates of convergence as EM. Recall that the largest eigenvalue of J must be less than 1 in order to guarantee the convergence of the linear iteration of (4.1). To study the linear iteration in (4.1), we use the following decomposition of J.

LEMMA. *Suppose* $I_o(\hat{\theta} \mid Y_{obs}) > 0$. *Then the* $d \times d$ *matrix* J *has the following decomposition*:

$$J = \sum_{j=1}^{k} \lambda_j u_j^\top v_j, \tag{4.2}$$

where $1 > \lambda_1 > \lambda_2 > \cdots > \lambda_k \geq 0$ *are* k $(\leq d)$ *distinct eigenvalues of* J *with the corresponding multiplicities* $m_1, \ldots, m_k$; *the* $m_j \times d$ *matrices* u_j, v_j $(j = 1, \ldots, k)$ *form the bases of the* j*th eigenvector spaces for* J *and* $J^\top$, *respectively*; *and*

$$\begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix} \left(v_1^\top, \ldots, v_k^\top \right) = I_d, \tag{4.3}$$

with I_d *the* $d \times d$ *identity matrix.*

Proof. It is well known in statistics that the matrix I_{om} of (2.4) is nonnegative definite because it is equal to

$$\int \left(\frac{\partial \log f(Y_{mis} \mid Y_{obs}, \theta)}{\partial \theta} \right) \left(\frac{\partial \log f(Y_{mis} \mid Y_{obs}, \theta)}{\partial \theta} \right)^\top f(Y_{mis} \mid Y_{obs}, \theta)\, dY_{mis},$$

which can be verified using integration by parts (e.g., Cox and Hinkley, 1974). It follows then from our assumption and (2.7) that $I_{oc}(\hat{\theta} \mid Y_{obs}) > 0$. The lemma then follows immediately from the well-known diagonability theorem and the fact that J is similar to a symmetric matrix

$$\tilde{J} = I_{oc}^{-1/2}\big(\hat{\theta} \mid Y_{obs}\big) I_{om}\big(\hat{\theta} \mid Y_{obs}\big) I_{oc}^{-1/2}\big(\hat{\theta} \mid Y_{obs}\big),$$

which is always diagonalizable (e.g., see Searle, 1982, Chapter 11).

We now summarize our main result as a proposition. To avoid trivial cases, we assume $\theta^{(0)} \neq \hat{\theta}$ so that $\varphi^{(0)} = \theta^{(0)} - \hat{\theta} \not\equiv 0$.

PROPOSITION.

(a) *For each* $i \in \mathscr{D} = \{1, 2, \ldots, d\}$ *let*

$$A_i = \big\{ j \in \mathscr{D} \mid w_{ij} \equiv \varphi^{(0)} u_j^\top v_j e_i^\top \neq 0 \big\},$$

where e_i *is the* i*th row of the identity matrix* I_d*, and let*

$$j_i = \min\{ j \in A_i, k+1 \}. \tag{4.4}$$

Then the i*th componentwise rate of convergence of EM, defined in* (2.2)*, is given by*

$$R_i = \lambda_{j_i}, \tag{4.5}$$

where $\lambda_{k+1} \equiv 0$.

(b) *Let*

$$j_0 = \min_{i \in \mathscr{D}} j_i. \tag{4.6}$$

Then the global rate of convergence of EM, defined by (2.1)*, is given by*

$$R = \lambda_{j_0} = \max_{i \in \mathscr{D}} R_i. \tag{4.7}$$

Proof. (a): Following (4.1)–(4.2), we have

$$\varphi^{(t)} = \sum_{j=1}^{k} \lambda_j^t \varphi^{(0)} u_j^\top v_j. \tag{4.8}$$

It follows immediately that, for $i \in \mathscr{D}$, if $j_i < k$ or $j_i = k$ but $\lambda_k \neq 0$, then

$$R_i = \lim_{t\to\infty} R_i^{(t)} = \lim_{t\to\infty} \frac{\varphi_i^{(t+1)}}{\varphi_i^{(t)}} = \lim_{t\to\infty} \frac{\sum_{j=j_i}^{k} \lambda_j^{t+1} w_{ij}}{\sum_{j=j_i}^{k} \lambda_j^t w_{ij}}$$

$$= \lim_{t\to\infty} \frac{\lambda_{j_i} + \sum_{j=j_i+1}^{k} (\lambda_j/\lambda_{j_i})^t \lambda_j w_{ij}/w_{ij_i}}{1 + \sum_{j=j_i+1}^{k} (\lambda_j/\lambda_{j_i})^t w_{ij}/w_{ij_i}} = \lambda_{j_i}.$$

This last step follows because $\lim_{t\to\infty}(\lambda_j/\lambda_{j_i})^t = 0$ for any $j > j_i \geq 1$. If $j_i = k$ and $\lambda_k = 0$ or $j_i = k+1$, then $\varphi_i^{(t)} \equiv 0$ for any $t \geq 1$, and hence, by definition, $R_i = 0 = \lambda_{j_i}$.

(b): By (4.6) and (4.8), we have (notice $\lambda_{j_0} > 0$ because $\varphi^{(0)} \neq 0$)

$$\frac{\|\varphi^{(t+1)}\|^2}{\|\varphi^{(t)}\|^2} = \frac{\sum_{i=1}^{d} \left[\varphi_i^{(t+1)}\right]^2}{\sum_{i=1}^{d} \left[\varphi_i^{(t)}\right]^2} = \frac{\sum_{i=1}^{d} \left[\sum_{j=j_i}^{k} (\lambda_j/\lambda_{j_0})^t \lambda_j w_{ij}\right]^2}{\sum_{i=1}^{d} \left[\sum_{j=j_i}^{k} (\lambda_j/\lambda_{j0})^t \omega_{ij}\right]^2}.$$

Since $\lim_{t\to\infty}(\lambda_j/\lambda_{j_0})^t = 0$ for any $j > j_0$, the above identity leads to

$$R^2 = \lim_{t\to\infty} \frac{\|\varphi^{(t+1)}\|^2}{\|\varphi^{(t)}\|^2} = \frac{\lambda_{j_0}^2 \sum_{i:\, j_i = j_0} w_{ij_0}^2}{\sum_{i:\, j_i = j_0} w_{ij_0}^2} = \lambda_{j_0}^2,$$

which completes our proof. ■

Part (a) of the Proposition indicates that although each R_i ($i \in \mathscr{D}$) must equal one of the eigenvalues of J and $R_i \leq \lambda_1$ (the largest eigenvalue of J), the R_i's are not necessarily equal to λ_1. Part (b) confirms our intuition that the global rate of convergence should be equal to the componentwise rate of convergence of the slowest component(s), since the whole algorithm converges if and only if all components converge.

As a consequence of the Proposition, the following corollary gives the condition under which all componentwise rates of convergence will be the same.

COROLLARY. *$R_i \equiv R$ for all $i \in \mathscr{D}$ if and only if*

$$\varphi^{(0)} u_j^\top = 0 \quad \text{for } j < j_0, \quad \text{and} \quad w_{ij_0} \neq 0 \quad \text{for all } i \in \mathscr{D}, \tag{4.9}$$

where $u_0 \equiv 0$.

Proof. Following (4.5) and (4.7), we only need to show that $j_i \equiv j_0$ for all $i \in \mathscr{D}$ if and only if (4.9) holds. It is clear that, by the definition of j_i of (4.4),

$$j_i \equiv j_0 \quad \Leftrightarrow \quad \varphi^{(0)} u_j^\top v_j = 0 \text{ for } j < j_0, \text{ and } w_{ij_0} \neq 0 \text{ for all } i \in \mathscr{D}.$$

But this is equivalent to (4.9), because $\varphi^{(0)} u_j^\top v_j = 0$ if and only if $\varphi^{(0)} u_j^\top = 0$ (since $v_j v_j^\top$ is nonsingular), completing the proof. ■

This result suggests that all R_i's are the same if and only if $\varphi^{(0)} = \theta^{(0)} - \hat{\theta}$ has "homogeneous" projections onto the eigenvector spaces. For example, $R_i \equiv \lambda_2$, the second largest eigenvalue, if and only if $\varphi^{(0)}$ is orthogonal to the first eigenvector space u_1 and there is no zero coordinate in the projection to the second space: $\varphi^{(0)} u_2^\top v_2$.

In practice, because one typically has no control over $\varphi^{(0)}$ (since $\hat{\theta}$ is unknown before running EM), it is very unlikely that there will be zero coordinates in the projection $\varphi^{(0)} u_1^\top v_1$ unless J has special structure, as in the diagonal case in Section 3. Consequently, in most practical situations, all components converge at the global rate, which equals the largest eigenvalue of the matrix of fractions of missing information. Nevertheless, as illustrated in Section 3, the phenomenon that different components converge at different rates can occur with models used in statistical practice.

Ingram Olkin has been one of statistics' most prolific contributors to multivariate models and linear algebra. We are particularly pleased, therefore, to be able to contribute to this special issue honoring his 70th birthday, especially with a topic on a method whose theoretical foundation is built upon linear algebra, and whose primary applications are in multivariate statistics.

This work was supported in part by several National Science Foundation grants awarded to Harvard University and University of Chicago, and in part by Joint Statistical Agreements between the U.S. Bureau of the Census and Harvard University. The manuscript was prepared using computer facilities supported in part by several National Science Foundation Grants awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

We thank the reviewers for helpful suggestions that enhanced the readability of the paper, especially for a nonstatistical audience.

REFERENCES

Boyles, R. A. 1983. On the convergence of the EM algorithm, *J. Roy. Statist. Soc. Ser. B* 45:47–50.

Cox, D. R. and Hinkley, D. V. 1974. *Theoretical Statistics*, Wiley, New York.

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.

Faddeev, D. K. and Faddeeva, V. N. 1963. *Computational Methods of Linear Algebra*, Freeman, San Francisco.

Healy, M. J. R. and Westmacott, M. 1956. Missing values in experiments analyzed on automatic computers, *Appl. Statist.* 5:203–206.

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

Meng, X. L. and Pedlow, S. 1992. EM: A bibliographic review with missing articles, *Proc. Statist. Comput. Sect. Amer. Statist. Assoc.*, to appear.

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. 1992. Recent extensions to the EM algorithm (with discussion), in *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Eds.), Oxford U.P., Oxford, pp. 307–320.

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

Orchard, T. and Woodbury, M. A. 1972. A missing information principle: Theory and application, in *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 697–715.

Searle, S. R. 1982. *Matrix Algebra Useful for Statistics*, Wiley, New York.

Wu, C. F. J. 1983. On the convergence properties of the EM algorithm, *Ann. Statist.* 11:95–103.

Received 6 January 1993; final manuscript accepted 22 October 1993