The Annals of Statistics
1994, Vol. 22, No. 1, 326–339

ON THE RATE OF CONVERGENCE OF THE ECM ALGORITHM[1]

BY XIAO-LI MENG

University of Chicago

The fundamental result on the rate of convergence of the EM algorithm has proven to be theoretically valuable and practically useful. Here, this result is generalized to the ECM algorithm, a more flexible and applicable iterative algorithm proposed recently by Meng and Rubin. Results on the rate of convergence of variations of ECM are also presented. An example is given to show that intuitions accurate for complete-data iterative algorithms may not be trustworthy in the presence of missing data.

1. The EM and ECM algorithms. Replacing the M-step of the EM algorithm [Dempster, Laird and Rubin (1977), hereafter DLR] by a set of conditional maximization (CM) steps, Meng and Rubin (1993) proposed a type of generalized EM algorithm—the ECM algorithm. The ECM algorithm not only maintains all the desirable properties of EM, but it can eliminate the undesirable nested iterations when the M-step of EM requires numerical iterations. In the absence of missing data, the ECM algorithm also includes several well-known complete-data iterative techniques as its special cases, such as iterative proportional fitting (IPF) for a loglinear model with contingency tables [cf. Bishop, Fienberg and Holland (1975)] and iterated conditional modes (ICM) for image reconstruction [cf. Besag (1986)].

To be specific, let $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ be the complete data with density $f(Y|\theta)$ indexed by a $1 \times d$ vector parameter $\theta \in \Theta \subseteq R^d$, where Y_{obs} and Y_{mis} are the observed and missing data, respectively. Our objective here is to find the maximum likelihood estimate (MLE) for θ given the observed data, that is, we want to find θ^* that maximizes the observed-data log-likelihood

$$L_{\text{obs}}(\theta|Y_{\text{obs}}) \propto \log \int f(Y|\theta)\, dY_{\text{mis}}. \tag{1.1}$$

Typically, the presence of missing data, or mathematically the presence of integration on the right-hand side of (1.1), makes the direct maximization of L_{obs} intractable. In contrast, in many statistical applications, the maximization of the complete-data log-likelihood $L(\theta|Y) = \log f(Y|\theta)$ is straightforward. The EM algorithm takes advantage of this simplicity by converting the difficult

Received March 1992; revised February 1993.

[1]Supported in part by NSF Grant DMS-92-04504 and by the University of Chicago/AMOCO Fund. This manuscript was prepared using computer facilities supported in part by NSF Grants DMS-89-05292, DMS-87-03942, DMS-86-01732 and DMS-84-04941, awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund.

AMS 1991 *subject classifications*. Primary 65U05; secondary 62F10.

Key words and phrases. Conditional maximization, EM algorithm, Gibbs sampler, incomplete data, missing data, SEM algorithm, speed of convergence.

problem of maximizing L_{obs} into an iterative sequence of simple maximizations of L. Given some initial value $\theta^{(0)} \in \Theta$, EM performs one E-step and one M-step at each iteration. At the $(t+1)$st, $t = 0, 1, \ldots$, iteration, the E-step finds the conditional expectation of $L(\theta|Y)$ given the observed data and the previous estimate $\theta^{(t)}$,

$$Q(\theta|\theta^{(t)}) = \int L(\theta|Y)\, f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})\, dY_{\text{mis}}; \tag{1.2}$$

then the M-step maximizes $Q(\theta|\theta^{(t)})$ as a function of θ, which uses the same method for maximizing L. It is well-known that any sequence generated by EM, $\{\theta^{(t)}, t \geq 0\}$, always increases $L_{\text{obs}}(\theta|Y_{\text{obs}})$, and it converges appropriately under some regularity conditions [see Baum, Petrie, Soules and Weiss (1970), DLR and Wu (1983)].

Although the simplicity and applicability of EM has made it one of the most popular iterative algorithms in statistics during the past 15 years or so, it has long been noticed that there exist a variety of practically important problems where the simplicity of the M-step is lost because the complete-data MLE's themselves are hard to compute and require numerical iterations. In many of these cases, however, the complete-data constrained MLE's, that is, the MLE's of θ restricted to particular subspaces of Θ, are in closed form or are relatively easy to obtain. Motivated by this observation, Meng and Rubin (1993) introduced the ECM algorithm which maintains the E-step of EM, but replaces the M-step by a set of CM-steps at the $(t+1)$st iteration: For $s = 1, \ldots, S$, find $\theta^{(t+s/S)}$ that maximizes $Q(\theta|\theta^{(t)})$ over $\theta \in \Theta$ subject to the constraint $g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$, where $G \equiv \{g_s(\theta), s = 1, \ldots, S\}$ is a set of $S(\geq 1)$ preselected (vector) functions. In other words, the sth CM-step of the $(t+1)$st iteration is to find $\theta^{(t+s/S)}$ such that

$$Q\big(\theta^{(t+s/S)}|\theta^{(t)}\big) \geq Q\big(\theta|\theta^{(t)}\big) \quad \text{for all } \theta \in \Theta_s\big(\theta^{(t+(s-1)/S)}\big), \tag{1.3}$$

where

$$\Theta_s(\theta) \equiv \{\vartheta \in \Theta\colon g_s(\vartheta) = g_s(\theta)\}. \tag{1.4}$$

The final output $\theta^{(t+S/S)}$ is taken to be the input of the next iteration, $\theta^{(t+1)}$.

An interesting and useful variation of ECM is to insert more E-steps at each iteration. For example, one may wish to perform one E-step before each CM-step. More generally, suppose one wants to insert one E-step before the s_kth CM-step, $k = 1, \ldots, K \leq S$, $s_1 \equiv 1$; then (1.3) is replaced by

$$Q\big(\theta^{(t+s/S)}|\theta^{(t+(s_k-1)/S)}\big) \geq Q\big(\theta|\theta^{(t+(s_k-1)/S)}\big) \quad \text{for all } \theta \in \Theta_s\big(\theta^{(t+(s-1)/S)}\big),\ s_k \leq s < s_{k+1}. \tag{1.5}$$

Meng and Rubin (1993) call such an extension of ECM the multicycle ECM (MCECM) algorithm, where a cycle is defined as one E-step followed by one or several CM-steps. Thus, the algorithm defined by (1.5) is a K-cycle ECM.

As shown in Meng and Rubin (1993), any ECM or MCECM sequence monotonically increases L_{obs}, a key feature of EM. Furthermore, under the assumption that $g_s(\theta)$, $s = 1, \ldots, S$, is differentiable and that the corresponding gradient $\nabla g_s(\theta)$ is of full rank at $\theta^{(t)}$, for all t, almost all of the convergence properties of EM established in DLR and in Wu (1983) hold for ECM and MCECM. The only extra condition needed for ECM and MCECM is the "space filling" condition:

$$\bigcap_{s=1}^{S} J_s\left(\theta^{(t)}\right) = \{\mathbf{0}\} \quad \text{for all } t, \tag{1.6}$$

where $J_s(\theta)$ is the column space of $\nabla g_s(\theta)$. By taking the orthogonal complement of both sides of (1.6), this condition is equivalent to saying that at any $\theta^{(t)}$, the convex hull of all feasible directions determined by the constraint spaces $\Theta_s(\theta^{(t)})$, $s = 1, \ldots, S$, where $\Theta_s(\theta)$ is defined in (1.4), is the whole Euclidean space R^d, and thus the resulting maximization by repeated conditional maximizations is over the whole parameter space Θ, not over a subspace of it. Notice that EM is a special case of ECM with $S = 1$ and $g_1(\theta) =$ constant (i.e., no constraint), whereby (1.6) is automatically satisfied because $\nabla g_1(\theta) \equiv \mathbf{0}$. Another special case of ECM is the CM algorithm (i.e., ECM without missing data), where the complete-data MLE's are found by iterating among conditional maximizations, a technique that is also known as the cyclic coordinate ascent method in the optimization literature [e.g., Zangwill (1969)]. The aforementioned IPF and ICM are examples of the CM algorithm. Meng and Rubin (1992) also discuss the similarity between ECM and the Gibbs sampler [Geman and Geman (1984)], which may lead to useful stochastic generalizations of ECM and extensions of the Gibbs sampler.

2. Measuring convergence of linear iterations. Any iterative algorithm that generates a sequence $\{\theta^{(t)}, t \geq 0\}$, such as EM and ECM, implicitly defines a mapping $\theta \to M(\theta)$ from the parameter space to itself such that $\theta^{(t+1)} = M(\theta^{(t)})$. If θ^* is a limit of $\{\theta^{(t)}, t \geq 0\}$ and $M(\theta)$ is differentiable in the neighborhood of θ^*, a Taylor series expansion then yields

$$\theta^{(t+1)} - \theta^* \approx \left(\theta^{(t)} - \theta^*\right) DM\left(\theta^*\right), \tag{2.1}$$

where

$$DM(\theta) = \left(\frac{\partial M_j(\theta)}{\partial \theta_i}\right) \tag{2.2}$$

is the $d \times d$ first derivative matrix for $M(\theta) = (M_1(\theta), \ldots, M_d(\theta))$, that is, the Jacobian matrix for the mapping M. Thus, if $DM(\theta^*)$ is nonzero, as with EM and ECM, expression (2.1) implies that the iterative algorithm determined by the mapping M is essentially a linear iteration with the rate matrix $DM(\theta^*)$. For this reason, $DM(\theta^*)$ is often referred to as the matrix rate of convergence, or simply the rate of convergence.

For multidimensional θ, a measure of the actual observed convergence rate is the global rate of convergence, which is defined as

$$r = \lim_{t \to \infty} \frac{\left\| \theta^{(t+1)} - \theta^* \right\|}{\left\| \theta^{(t)} - \theta^* \right\|}, \tag{2.3}$$

where $\| \cdot \|$ is the Euclidean norm. It is well-known that, under certain regularity conditions,

$$r = \lambda_{\max} \equiv \text{ the largest eigenvalue of } DM(\theta^*). \tag{2.4}$$

In practice, r is typically calculated by

$$r = \lim_{t \to \infty} \frac{\left\| \theta^{(t+1)} - \theta^{(t)} \right\|}{\left\| \theta^{(t)} - \theta^{(t-1)} \right\|}. \tag{2.5}$$

Expression (2.5) allows the global rate to be calculated simultaneously with the calculation of θ^* and essentially is an application of the well-known *power method* for finding the largest eigenvalue of a matrix [e.g., Faddeev and Faddeeva (1963), Dennis and Schnabel (1983)].

For multidimensional θ, we can also measure the rate of convergence component by component. The ith componentwise rate of convergence is defined as

$$r_i = \lim_{t \to \infty} \frac{\left| \theta_i^{(t+1)} - \theta_i^* \right|}{\left| \theta_i^{(t)} - \theta_i^* \right|}, \tag{2.6}$$

provided that it exists. We define $r_i \equiv 0$ if $\theta_i^{(t)} \equiv \theta_i^{(t_0)}$, for all $t \geq t_0$, t_0 fixed. For computation in practice, the alternative expression

$$r_i = \lim_{t \to \infty} \frac{\left| \theta_i^{(t+1)} - \theta_i^{(t)} \right|}{\left| \theta_i^{(t)} - \theta_i^{(t-1)} \right|}$$

is typically used in analogy to (2.5). Under broad regularity conditions, it is also easy to show that [e.g., Meng and Rubin (1994)]

$$r = \max_{1 \leq i \leq d} r_i,$$

which is consistent with the intuition that the whole algorithm converges if and only if every component does. A component whose componentwise rate equals the global rate is then called a *slowest component* for the obvious reason. A component is the slowest if it is not orthogonal to the eigenvector corresponding to $\lambda_{\max}$, and thus typically there are more than one such component.

Notice that a large value of r implies slow convergence. To be consistent with the common notion that the higher the value of the measure of convergence, the faster the algorithm converges, we may define $s = 1 - r$ as the "global speed of convergence." From (2.4), s is also the smallest eigenvalue of $\mathcal{S} = I_d - DM(\theta^*)$,

a matrix that may be called the (matrix) speed of convergence (S is often referred to as an "iteration matrix" in optimization literature).

For the EM algorithm, DLR established a fundamental identity between the (matrix) rate of convergence of EM and the matrix of *fractions of missing information*. More specifically, under very mild regularity conditions, DLR showed that for the EM mapping M^{EM},

$$DM^{\mathrm{EM}}(\theta^*) = I_{\mathrm{mis}}(\theta^*)I_{\mathrm{com}}^{-1}(\theta^*), \tag{2.7}$$

where

$$I_{\mathrm{mis}}(\theta) = \int -\frac{\partial^2 \log f(Y_{\mathrm{mis}} \mid Y_{\mathrm{obs}}, \theta)}{\partial\theta\, \partial\theta} f(Y_{\mathrm{mis}} \mid Y_{\mathrm{obs}}, \theta)\, dY_{\mathrm{mis}} \tag{2.8}$$

and

$$I_{\mathrm{com}}(\theta) = \int -\frac{\partial^2 \log f(Y \mid \theta)}{\partial\theta\, \partial\theta} f(Y_{\mathrm{mis}} \mid Y_{\mathrm{obs}}, \theta)\, dY_{\mathrm{mis}}. \tag{2.9}$$

The matrix on the right-hand side of (2.7) is called the (matrix of) *fractions of missing information*, because $I_{\mathrm{com}}(\theta^*)$ measures the complete information that one would expect to have if there were no missing data and $I_{\mathrm{mis}}(\theta^*)$ measures the loss of information due to missing data [Orchard and Woodbury (1972), Meng and Rubin (1991)]. Throughout this paper, we assume $I_{\mathrm{com}}(\theta^*)$ is positive definite. Besides its obvious theoretical value, identity (2.7) also serves as the foundation of the supplemented EM (SEM) algorithm [Meng and Rubin (1991)], which computes the $DM^{\mathrm{EM}}(\theta^*)$ of (2.7) and then uses it to inflate the complete-data asymptotic variance–covariance matrix to obtain the asymptotic variance–covariance matrix associated with the MLE when implementing EM.

In the next section, we present the matrix rates of convergence of ECM and MCECM. These results not only give us analytical tools for studying how the rates of convergence of ECM and MCECM vary with different settings of conditional maximizations (e.g., different orders for the S steps), but also provide fundamental formulas for computing the asymptotic variance–covariance matrix for the MLE's by combining ECM or MCECM with SEM [Meng and Rubin (1992)]. Section 4 provides a counterintuitive example of the relationships among the global rates of EM, ECM and MCECM.

3. The matrix rates of ECM and MCECM. Throughout the rest of the paper, we assume the same regularity conditions as in Meng and Rubin (1993). In particular, all the following calculations are performed inside Θ_0, the interior of Θ, and all required derivatives are well-defined. The following theorem gives the Jacobian matrix of the ECM mapping at the limit and thus establishes the matrix rate of covergence of ECM.

THEOREM 1. *Suppose all the outputs of an ECM,* $\theta^{(t+s/S)}$, $t \geq 1$, $s = 1, \ldots, S$, *satisfy the Lagrange multiplier equations for constrained maximization, and*

$\theta^{(t+s/S)} \to \theta^*$ as $t \to \infty$ for $s = 1, \ldots, S$. Then the rate of convergence of ECM is given by

$$DM^{\mathrm{ECM}}(\theta^*) = DM^{\mathrm{EM}}(\theta^*) + [I_d - DM^{\mathrm{EM}}(\theta^*)] \prod_{s=1}^{S} P_s, \tag{3.1}$$

where $DM^{\mathrm{EM}}(\theta^*)$ is the rate of convergence of EM given in (2.7),

$$P_s = \nabla_s [\nabla_s^T I_{\mathrm{com}}^{-1}(\theta^*)\, \nabla_s]^{-1}\, \nabla_s^{\mathrm{T}} I_{\mathrm{com}}^{-1}(\theta^*), \qquad s = 1, \ldots, S, \tag{3.2}$$

with $\nabla_s = \nabla g_s(\theta^*)$ and $\prod_{s=1}^{S} P_s \equiv P_1 \cdots P_S$.

PROOF. For any given $\xi, \eta \in \Theta$ and $1 \le s \le S$, let $G_s(\xi, \eta)$ be a maximizer of $Q(\theta|\xi)$ [defined in (1.2)] under the constraint $\theta \in \Theta_s(\eta) \equiv \{\theta: g_s(\theta) = g_s(\eta)\}$. Let $M_0(\theta) = \theta$ and

$$M_s(\theta) = G_s(\theta, M_{s-1}(\theta)) \quad \text{for all } s. \tag{3.3}$$

Then, by the construction of ECM,

$$\theta^{(t+s/S)} = M_s(\theta^{(t)}), \qquad s = 1, \ldots, S, \tag{3.4}$$

and thus $M^{\mathrm{ECM}}(\theta) \equiv M_S(\theta)$. It follows from our assumptions that

$$\theta^* = M_s(\theta^*), \qquad s = 1, \ldots, S, \tag{3.5}$$

hence

$$\theta^* = G_s(\theta^*, \theta^*) \quad \text{for all } s. \tag{3.6}$$

To obtain $DM^{\mathrm{ECM}}(\theta^*) \equiv DM_S(\theta^*)$, we first differentiate both sides of (3.3) and evaluate them at $\theta = \theta^*$ using (3.5). This yields

$$DM_s(\theta^*) = D^{10}G_s(\theta^*, \theta^*) + DM_{s-1}(\theta^*) D^{01}G_s(\theta^*, \theta^*), \qquad s = 1, \ldots, S, \tag{3.7}$$

where D^{10} denotes the partial derivative with respect to the first argument, and so on. Next we calculate $D^{10}G_s(\theta^*, \theta^*)$ and $D^{01}G_s(\theta^*, \theta^*)$ by differentiating the following two Lagrange equations with respect to ξ and η and then evaluating them at $\xi = \theta^*$ and $\eta = \theta^*$:

$$g_s(G_s(\xi, \eta)) = g_s(\eta), \tag{3.8}$$

$$D^{10}Q(G_s(\xi, \eta)|\xi) - \nabla g_s(G_s(\xi, \eta))\lambda_s(\xi, \eta) = 0, \tag{3.9}$$

where $\lambda_s(\xi, \eta)$ is the Lagrange multiplier. This yields from (3.8) and (3.6) that

$$D^{10}G_s(\theta^*, \theta^*)\, \nabla_s = 0, \tag{3.10}$$

$$D^{01}G_s(\theta^*, \theta^*)\, \nabla_s = \nabla_s. \tag{3.11}$$

Similarly, differentiating (3.9) yields

$$D^{10}G_s(\theta^*,\theta^*)I_{\text{com}}(\theta^*) - I_{\text{mis}}(\theta^*) + D^{10}\lambda_s(\theta^*,\theta^*)\ \nabla_s^T = 0 \tag{3.12}$$

and

$$D^{01}G_s(\theta^*,\theta^*)I_{\text{com}}(\theta^*) + D^{01}\lambda_s(\theta^*,\theta^*)\ \nabla_s^T = 0, \tag{3.13}$$

where $I_{\text{mis}}(\theta^*)$ and $I_{\text{com}}(\theta^*)$ are defined in (2.8) and (2.9), respectively. Notice that in deriving (3.12) and (3.13), we have used three facts besides (3.6): (i) $D^{20}Q(\theta^*|\theta^*) = -I_{\text{com}}(\theta^*)$; (ii) $D^{11}Q(\theta^*|\theta^*) = I_{\text{mis}}(\theta^*)$ (DLR); (iii) $\lambda_s(\theta^*,\theta^*) = 0$ because $D^{10}Q(\theta^*|\theta^*) = 0$ [Meng and Rubin (1993)].

From (3.11) and (3.13), we have

$$D^{01}G_s(\theta^*,\theta^*) = \nabla_s\left[\nabla_s^T I_{\text{com}}^{-1}(\theta^*)\ \nabla_s\right]^{-1}\ \nabla_s^T I_{\text{com}}^{-1}(\theta^*) \equiv P_s. \tag{3.14}$$

Combining (3.10) and (3.12) with (3.14) and (2.7) yields

$$D^{10}G_s(\theta^*,\theta^*) = DM^{\text{EM}}(\theta^*)(I_d - P_s). \tag{3.15}$$

It follows from (3.7) and (3.15) that

$$DM_s(\theta^*) - DM^{\text{EM}}(\theta^*) = \left[DM_{s-1}(\theta^*) - DM^{\text{EM}}(\theta^*)\right]P_s,$$

which implies that [notice $DM_0(\theta) \equiv I$]

$$DM_S(\theta^*) - DM^{\text{EM}}(\theta^*) = \left[I_d - DM^{\text{EM}}(\theta^*)\right]P_1 \cdots P_S,$$

and thus (3.1) follows. □

Three interesting points regarding Theorem 1 are worth mentioning. First, in the absence of missing data, $DM^{\text{EM}}(\theta^*) = 0$, and thus (3.1) implies

$$DM^{\text{CM}}(\theta^*) = \prod_{s=1}^{S} P_s, \tag{3.16}$$

where $M^{\text{CM}}(\theta)$ is the mapping determined by the CM algorithm. It follows from (3.1) and (3.16) that

$$\left[I_d - DM^{\text{ECM}}(\theta^*)\right] = \left[I_d - DM^{\text{EM}}(\theta^*)\right]\left[I_d - DM^{\text{CM}}(\theta^*)\right], \tag{3.17}$$

which, following our definition of the (matrix) speed of convergence in Section 2, has the following very appealing interpretation:

$$\text{Speed of ECM} = \text{Speed of EM} \times \text{Speed of CM}.$$

This identity is consistent with our intuition, since ECM can be viewed as a composition of two linear iterations: EM and CM.

Second, if we let

$$\rho_{s,s+1} = \left(\nabla_s^T I_{\text{com}}^{-1} \nabla_s\right)^{-1/2}\left(\nabla_s^T I_{\text{com}}^{-1} \nabla_{s+1}\right)\left(\nabla_{s+1}^T I_{\text{com}}^{-1} \nabla_{s+1}\right)^{-1/2}, \quad s = 0, \ldots, S, \tag{3.18}$$

where $\nabla_0 \equiv \nabla_{S+1} \equiv I_d$, then (3.16) can be rewritten as

$$DM^{\text{CM}}(\theta^*) = I_{\text{com}}^{1/2}\left(\prod_{s=0}^{S} \rho_{s,s+1}\right) I_{\text{com}}^{-1/2}. \tag{3.19}$$

Notice that $\rho_{s,s+1}$ of (3.18) can be viewed as the "correlation coefficient" between ∇_s and ∇_{s+1} with respect to I_{com}^{-1} or as the cosine of the angle between two subspaces of R^d. In this sense, (3.19) has a structure similar to some results on rates of convergence of the Gibbs sampler [e.g., Amit (1991)], which further demonstrates the similarity between ECM and the Gibbs sampler [Meng and Rubin (1992)].

Third, expression (3.16) and thus (3.1) have an intuitive statistical interpretation. Consider, for example, a simple complete-data problem with $\theta = (\theta_1, \theta_2)$, where we choose $g_1(\theta) = \theta_2$ and $g_2(\theta) = \theta_1$. At the $(t+1)$st iteration, the corresponding first CM-step is to maximize $L(\theta_1, \theta_2^{(t)})$ to obtain $\theta_1^{(t+1)}$, and the second CM-step is to maximize $L(\theta_1^{(t+1)}, \theta_2)$ to determine $\theta_2^{(t+1)}$, where $L(\theta)$ is the log-likelihood function. Let $\theta^* = (\theta_1^*, \theta_2^*)$ be the limit of $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$. Because $L(\theta)$ is locally quadratic around θ^*, the first CM-step is asymptotically (as $t \to \infty$) equivalent to predicting $\theta_1^{(t+1)} - \theta_1^*$ from $\theta_2^{(t)} - \theta_2^*$ by regression, that is,

$$\theta_1^{(t+1)} - \theta_1^* = \beta_{12}\left(\theta_2^{(t)} - \theta_2^*\right), \tag{3.20}$$

where $\beta_{12} = \rho\sigma_1/\sigma_2$ and ρ and σ_i, $i = 1, 2$, are from the complete-data asymptotic variance–covariance matrix

$$V_{\text{com}} \equiv I_{\text{com}}^{-1}(\theta^*) \equiv \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{3.21}$$

Similarly, the second CM-step is equivalent to

$$\begin{aligned} \theta_2^{(t+1)} - \theta_2^* &= \beta_{21}\left(\theta_1^{(t+1)} - \theta_1^*\right) \\ &= \beta_{21}\beta_{12}\left(\theta_2^{(t)} - \theta_2^*\right) \quad [\text{from } (3.20)]. \end{aligned} \tag{3.22}$$

Expressions (3.20) and (3.22) together imply

$$DM^{\text{CM}}(\theta^*) = \begin{pmatrix} 0 & 0 \\ \beta_{12} & \beta_{21}\beta_{12} \end{pmatrix} \equiv \begin{pmatrix} 0 & 0 \\ \beta_{12} & \rho^2 \end{pmatrix}, \tag{3.23}$$

which can be verified as identical to (3.16) for the current problem. Notice from (3.23) that the linear iteration determined by $DM^{\text{CM}}(\theta^*)$ converges if and only

if its largest eigenvalue, ρ^2, is less than 1. In other words, the CM algorithm is essentially alternated regressions, and it converges if and only if there is a "regression effect," that is, if $\rho^2 < 1$.

The following theorem generalizes Theorem 1 to multicycle ECM.

THEOREM 2. *Under the conditions of Theorem* 1, *the rate of convergence of the (K-cycle) MCECM given by* (1.5) *is*

$$DM^{\mathrm{MCECM}}(\theta^*) = \prod_{k=1}^{K} \left\{ DM^{\mathrm{EM}}(\theta^*) + [I_d - DM^{\mathrm{EM}}(\theta^*)] \prod_{s=s_k}^{s_{k+1}-1} P_s \right\}, \tag{3.24}$$

where $s_1 \equiv 1$ *and* $s_{K+1} \equiv S + 1$.

PROOF. Since the mapping $M^{\mathrm{MCECM}}(\theta)$ defined by a K-cycle ECM is a composition of K single-cycle ECM mappings, (3.24) follows directly from (3.1) and the chain rule. One can also derive (3.24) directly by replacing (3.3) with $M_s(\theta) = G_s(M_{s-1}(\theta), M_{s-1}(\theta))$, for $s = s_k$, $k = 1, \ldots, K$, in the proof of Theorem 1. □

4. The global rates of ECM and MCECM. The results presented in Section 3 are in terms of matrices. For multidimensional θ, as it must be with ECM, the actual observed rate of convergence is the global rate r defined in (2.3). Since the appealing relationship in (3.17) is in terms of the speed matrices,

$$\mathcal{S}^{\mathrm{ECM}} = \mathcal{S}^{\mathrm{EM}} \mathcal{S}^{\mathrm{CM}}, \tag{4.1}$$

where $\mathcal{S}^{\mathrm{ECM}} = I_d - DM^{\mathrm{ECM}}(\theta^*)$, and so on, we will focus on global speeds instead of global rates. The results, of course, are equivalent. Let s^{ECM} denote the global speed of ECM, and similarly for s^{EM} and s^{CM}. Although it would be naive to conclude from (4.1) that

$$s^{\mathrm{ECM}} = s^{\mathrm{EM}} s^{\mathrm{CM}},$$

it seems intuitive to expect from (4.1) that

$$s^{\mathrm{EM}} s^{\mathrm{CM}} \le s^{\mathrm{ECM}} \le s^{\mathrm{EM}}. \tag{4.2}$$

Since at each interaction, ECM increases the Q-function of (1.2) less than EM would do, it seems intuitive to expect that ECM converges more slowly than EM in terms of the global speed, which implies the inequality on the right-hand side of (4.2). On the other hand, our intuition may suggest that the slowest convergence of ECM occurs when EM and CM share a slowest component (Section 2), in which cases the speed of ECM is the product of the speeds of EM and CM. This leads to the inequality on the left-hand side of (4.2).

Surprisingly, neither of the inequalities in (4.2) holds in general! The following simple bivariate normal example provides counterexamples to both

inequalities. Suppose the complete data consists of $Y_i = (y_{i1}, y_{i2})^T$, $i = 1, 2$, such that

$$Y_1, Y_2 \sim_{\text{i.i.d}} N\left[\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right],$$

where ρ is known ($|\rho| < 1$) and $\theta = (\theta_1, \theta_2)$ is to be estimated. In the presence of missing data, suppose we only observe $z_1 = y_{11}$ and $z_2 = y_{22} - y_{21}$, and we are interested in finding the MLE of θ based on (z_1, z_2). Notice that, in this case, the desired MLE is in closed form:

$$\theta^* = (z_1, z_1 + z_2). \tag{4.3}$$

To compare ECM with EM, we apply both algorithms to obtain two sequences that will converge to θ^* of (4.3). The E-step is the same for both EM and ECM and in this case, since ρ is known, is equivalent to imputing y_{ij}, $i, j = 1, 2$, by their conditional expectations $y_{ij}^{(t)} = E(y_{ij}|z_1, z_2, \theta^{(t)})$, where $y_{11}^{(t)} \equiv y_{11}$. The M-step then estimates θ by the "global mode" of the complete-data likelihood using the imputed data:

$$\theta_{\text{EM}}^{(t+1)} = (\theta_{\text{EM},1}^{(t+1)}, \theta_{\text{EM},2}^{(t+1)}) = (\bar{y}_1^{(t)}, \bar{y}_2^{(t)}), \tag{4.4}$$

where $\bar{y}_j^{(t)} = (y_{1j}^{(t)} + y_{2j}^{(t)})/2$, $j = 1, 2$. In contrast, ECM in this example replaces (4.4) by two conditional maximization steps, each of which corresponds to a "conditional mode" of the complete-data likelihood based on the imputed data:

$$\begin{aligned}\theta_{\text{ECM}}^{(t+1)} &= \left(\theta_{\text{ECM},1}^{(t+1)}, \theta_{\text{ECM},2}^{(t+1)}\right)\\ &= \left(\bar{y}_1^{(t)} + \rho\left(\theta_{\text{ECM},2}^{(t)} - \bar{y}_2^{(t)}\right), \bar{y}_2^{(t)} + \rho\left(\theta_{\text{ECM},1}^{(t+1)} - \bar{y}_1^{(t)}\right)\right).\end{aligned} \tag{4.5}$$

Notice that ECM coincides with EM when $\rho = 0$, that is, when θ_1 and θ_2 are orthogonal.

To calculate the rates of convergence for EM and ECM, we first notice that, in this problem,

$$I_{\text{com}}^{-1}(\theta^*) = \frac{1}{2}\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix} \quad \text{and} \quad I_{\text{mis}}(\theta^*)I_{\text{com}}^{-1}(\theta^*) = \frac{1}{4}\begin{pmatrix}1 & 1-2\rho\\ 1 & 3\end{pmatrix}. \tag{4.6}$$

Then, applying Theorem 1 with $S = 2$, $g_1(\theta) = \theta_2$ and $g_2(\theta) = \theta_1$, we have

$$\mathcal{S}^{\text{EM}} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} - \frac{1}{4}\begin{pmatrix}1 & 1-2\rho\\ 1 & 3\end{pmatrix} = \frac{1}{4}\begin{pmatrix}3 & 2\rho-1\\ -1 & 1\end{pmatrix},$$

$$\mathcal{S}^{\text{CM}} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} - \begin{pmatrix}0 & 0\\ \rho & \rho^2\end{pmatrix} = \begin{pmatrix}1 & 0\\ -\rho & 1-\rho^2\end{pmatrix},$$

$$\mathcal{S}^{\text{ECM}} = \mathcal{S}^{\text{EM}}\mathcal{S}^{\text{CM}} = \frac{1}{4}\begin{pmatrix}3+\rho(1-2\rho) & (2\rho-1)(1-\rho^2)\\ -(1+\rho) & 1-\rho^2\end{pmatrix}.$$

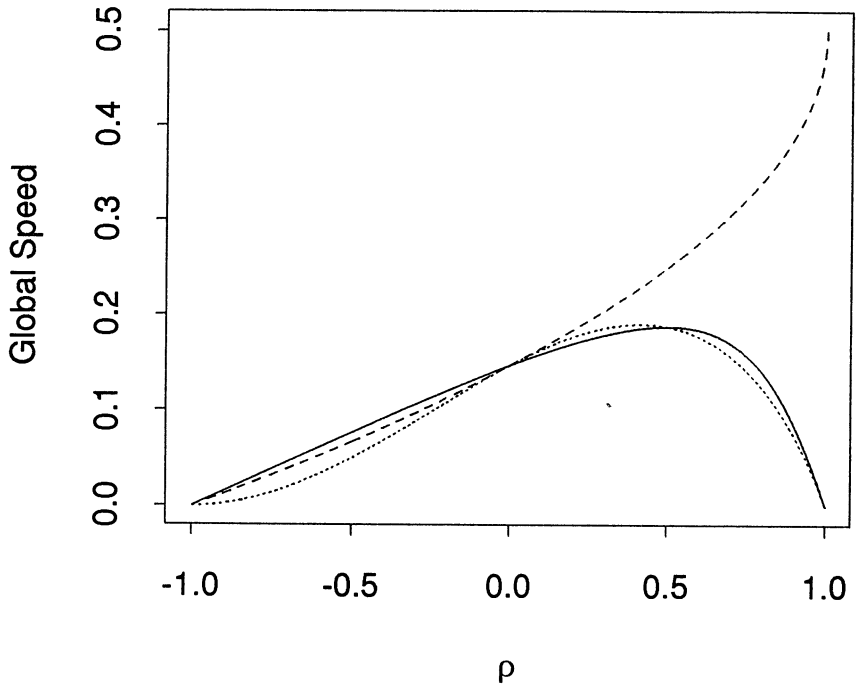

FIG. 1. *Comparisons of* s^{ECM} *(solid line),* s^{EM} *(dashed line) and* $s^{\mathrm{EM}}s^{\mathrm{CM}}$ *(dotted line).*

Calculating the smallest eigenvalues of these matrices yields, respectively,

$$s^{\mathrm{EM}} = \frac{1-\sqrt{(1-\rho)/2}}{2},$$

$$s^{\mathrm{CM}} = 1-\rho^2$$

and

$$s^{\mathrm{ECM}} = \frac{1+\rho}{8}\left(4-3\rho-\sqrt{(4-3\rho)^2-8(1-\rho)}\right).$$

To investigate (4.2), we plot s^{ECM}, s^{EM} and $s^{\mathrm{EM}}s^{\mathrm{CM}}$ in Figure 1 as functions of ρ. It is clear from the plot that $s^{\mathrm{ECM}} > s^{\mathrm{EM}}$ when $\rho < 0$, and $s^{\mathrm{ECM}} < s^{\mathrm{EM}}s^{\mathrm{CM}}$ when $0 < \rho < \frac{1}{2}$. Thus, neither of the inequalities in (4.2) holds in this example.

These comparisons are perhaps only of theoretical interest because the choice between EM and ECM in practice is typically determined by which one has closed-form maximizations, not by which one converges more rapidly. A practically more relevant comparison is between the global speeds of ECM and MCECM, which require almost the same computational effort except that the latter involves more E-steps at each iteration. Consequently, our intuition may suggest that MCECM should converge more rapidly than ECM would. Practical implementation does suggest that the use of MCECM can notably speed up the convergence of ECM [e.g., Belin, Difffendal, Rubin, Schafer and

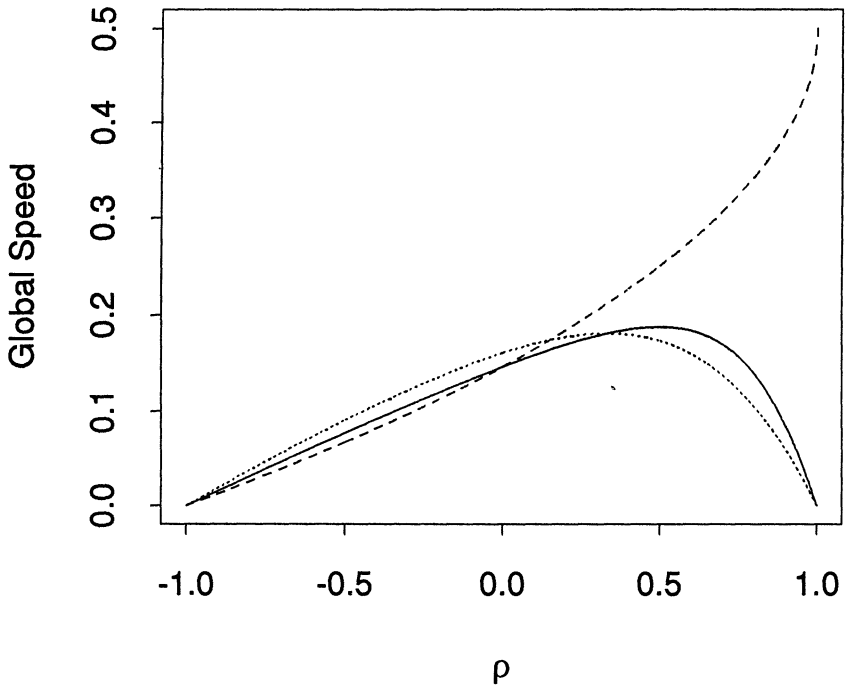

FIG. 2. *Comparisons of* s^{ECM} *(solid line),* s^{EM} *(dashed line) and* s^{MCECM} *(dotted line).*

Zaslavsky (1991)]. However, this is not true in general either! In fact, the same bivariate normal example above can serve as a counterexample, as detailed below.

Consider a two-cycle ECM in the example, that is, we add an E-step after the first CM-step. Applying Theorem 2 with $K = 2$, we obtain

$$S^{\mathrm{MCECM}} = I_2 - DM^{\mathrm{MCECM}}(\theta^*) = \frac{1+\rho}{4}\begin{pmatrix} 3-2\rho & \dfrac{\rho-2\rho^2-1}{4} \\ -1 & \dfrac{3-\rho}{4} \end{pmatrix}.$$

Thus, the global speed of the two-cycle ECM is

$$s^{\mathrm{MCECM}} = \frac{1+\rho}{8}\left[\frac{15-9\rho}{4} - \sqrt{\left(\frac{15-9\rho}{4}\right)^2 - 8(1-\rho)}\right].$$

Figure 2 plots s^{ECM}, s^{EM} and s^{MCECM} as functions of ρ. No curve completely dominates the others, and four out of the six possible dominance orderings among the three global speeds (e.g., $s^{\mathrm{EM}} < s^{\mathrm{ECM}} < s^{\mathrm{MCECM}}$) exist in the figure. In particular, $s^{\mathrm{MCECM}} < s^{\mathrm{ECM}}$ for $\rho > \frac{1}{3}$.

A final remark concerns possible explanations of these counterintuitive phenomena. Take the inequality $s^{\mathrm{ECM}} \leq s^{\mathrm{EM}}$ as an example. While the mathematical reason for violating this inequality is that $\mathcal{S}^{\mathrm{EM}}$ and $\mathcal{S}^{\mathrm{CM}}$ in (4.1) typically do not commute, its statistical interpretation is less clear. One observation we have in the bivariate normal example is that, in the case $s^{\mathrm{ECM}} > s^{\mathrm{EM}}$ (i.e., $\rho < 0$) with complete data,

$$\mathrm{Cov}_{\mathrm{com}}\left(\theta_1 - \theta_1^*, \theta_2 - \theta_2^*\right) = \mathrm{Cov}\left(\bar{y}_1, \bar{y}_2\right) = \frac{\rho}{2} < 0,$$

but with the observed data [see (4.3)],

$$\mathrm{Cov}_{\mathrm{obs}}\left(\theta_1 - \theta_1^*, \theta_2 - \theta_2^*\right) = \mathrm{Cov}\left(z_1, z_1 + z_2\right) = 1 > 0.$$

In other words, the missing data alter the sign of the correlation between the two components from negative to positive. In some sense, this could be viewed as the missing data helping to increase substantially the "dependence" between the two components. In fact, it can be shown that this "negative to positive" condition is sufficient to guarantee that ECM converges more rapidly than EM does for any two-dimensional problems. Higher-dimensional generalizations of this result are still under investigation. Perhaps the most important message here is that intuitions accurate for complete-data algorithms may not be reliable in the presence of missing data, and that more studies are needed before these counterintuitive phenomena can be better understood.

Acknowledgments. The author wishes to thank Andrew Gelman, Steven Pedlow, Donald Rubin and reviewers for very helpful suggestions that improved the presentation.

REFERENCES

AMIT, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Anal.* **37** 197–222.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

BELIN, T. R., DIFFENDAL, G. J., RUBIN, D. B., SCHAFER, J. L. and ZASLAVSKY, A. M. (1991). Documentation of handling of unresolved enumeration status in 1990 census/post-enumeration survey. STSD Decennial Census Memorandum Series V-98, U.S. Bureau of the Census.

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press.

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.

DENNIS, J. E. and SCHNABEL, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs, NJ.

FADDEEV, D. K. and FADDEEVA, V. N. (1963). *Computational Methods of Linear Algebra.* Freeman, San Francisco.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxations, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

MENG, X. L. (1990). Towards complete results for some incomplete-data problems. Ph.D. dissertation, Dept. Statistics, Harvard Univ. (Printed by U·M·I, Ann Arbor, MI.)

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* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) **4** 307–320. Oxford Univ. Press.

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 RUBIN, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra Appl.* (Special issue in honor of Ingram Olkin). To appear.

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: Theory and application. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 697–715. Univ. California Press, Berkeley.

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

ZANGWILL, W. (1969). *Nonlinear Programming—A Unified Approach.* Prentice-Hall, Englewood Cliffs, NJ.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637