Maximum Likelihood Estimation via the ECM Algorithm: A General Framework

Xiao-Li Meng; Donald B. Rubin

Biometrika, Vol. 80, No. 2 (Jun., 1993), 267-278.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28199306%2980%3A2%3C267%3AMLEVTE%3E2.0.CO%3B2-9

Biometrika is currently published by Biometrika Trust.

Biometrika (1993), **80**, 2, *pp.* 267–78
Printed in Great Britain

Maximum likelihood estimation via the ECM algorithm: A general framework

BY XIAO-LI MENG
Department of Statistics, University of Chicago, Chicago, Illinois 60637, U.S.A.
AND DONALD B. RUBIN
Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.

SUMMARY

Two major reasons for the popularity of the EM algorithm are that its maximum step involves only complete-data maximum likelihood estimation, which is often computationally simple, and that its convergence is stable, with each iteration increasing the likelihood. When the associated complete-data maximum likelihood estimation itself is complicated, EM is less attractive because the M-step is computationally unattractive. In many cases, however, complete-data maximum likelihood estimation is relatively simple when conditional on some function of the parameters being estimated. We introduce a class of generalized EM algorithms, which we call the ECM algorithm, for Expectation/Conditional Maximization (CM), that takes advantage of the simplicity of complete-data conditional maximum likelihood estimation by replacing a complicated M-step of EM with several computationally simpler CM-steps. We show that the ECM algorithm shares all the appealing convergence properties of EM, such as always increasing the likelihood, and present several illustrative examples.

Some key words: Bayesian inference; Conditional maximization; Constrained optimization; EM algorithm; Gibbs sampler; Incomplete data; Iterated conditional modes; Iterative proportional fitting; Missing data.

1. INTRODUCTION

The EM algorithm (Dempster, Laird & Rubin, 1977) is a very popular tool in modern statistics. It is an iterative method for finding maximum likelihood estimates and posterior modes in incomplete-data problems that has several appealing properties relative to other iterative algorithms such as Newton–Raphson. First, it is typically easily implemented because it relies on complete-data computations: the E-step of each iteration only involves taking expectations over complete-data conditional distributions and the M-step of each iteration only requires complete-data maximum likelihood estimation, which is often in simple closed form. Secondly, it is numerically stable: each iteration increases the likelihood or posterior density, and convergence is nearly always to a local maximum for practically important problems.

A brief review of EM establishes the required notation for our extension. Let Y denote the complete-data vector random variable with density $f(Y|\theta)$ indexed by a d-dimensional parameter $\theta \in \Theta \subseteq R^d$. If Y were observed, the objective would be to maximize the complete-data log-likelihood function of θ

$$L(\theta|Y) \propto \log f(Y|\theta),$$

or, more generally, to find the posterior mode of θ, which maximizes $L(\theta|Y)+\log p(\theta)$ for prior density $p(\theta)$ over all $\theta \in \Theta$; for Bayesian analysis, consider our log-likelihoods to be log-posteriors. In the presence of missing data, however, only a function of Y, Y_{obs}, is observed. In a convenient but imprecise notation, we write $Y=(Y_{obs}, Y_{mis})$, where Y_{mis} denotes the unobserved or missing data. For simplicity of description, we assume that the missing data are missing at random (Rubin, 1976), so that the log-likelihood for θ is

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

Because of the integration, maximizing L_{obs} can be difficult even when maximizing L is trivial.

The EM algorithm maximizes L_{obs} by iteratively maximizing L. Each iteration of EM has two steps: an E-step and an M-step. The $(t+1)$st E-step finds the conditional expectation of the complete-data log-likelihood with respect to the conditional distribution of Y_{mis} given Y_{obs} and the current estimated parameter $\theta^{(t)}$,

$$Q(\theta|\theta^{(t)}) = \int L(\theta|Y) f(Y_{mis}|Y_{obs}, \theta=\theta^{(t)})\, dY_{mis}, \tag{1·1}$$

as a function of θ for fixed Y_{obs} and fixed $\theta^{(t)}$. The $(t+1)$st M-step then finds $\theta^{(t+1)}$ to maximize $Q(\theta|\theta^{(t)})$:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)}), \quad \text{for all } \theta \in \Theta. \tag{1·2}$$

Although the general theory of EM applies to any model, it is particularly useful when the complete data Y are from an exponential family since, in such cases, the E-step reduces to finding the conditional expectation of the complete-data sufficient statistics, and the M-step is often simple. Nevertheless, even when the complete data Y are from an exponential family, there exist a variety of important applications where complete-data maximum likelihood estimation itself is complicated; for example, see Little & Rubin (1987) on selection models and log-linear models, which generally require iterative M-steps. In such cases, one way to avoid an iterative M-step within each EM iteration is to increase the Q function rather than maximize it at each M-step, resulting in a GEM algorithm (Dempster et al., 1977), which, although still increasing the log-likelihood L_{obs} at each iteration, does not in general appropriately converge without further specification on the process of increasing the Q function. The ECM algorithm is a subclass of GEM that is more broadly applicable than EM, but shares its desirable convergence properties.

More precisely, ECM replaces each M-step of EM, given by (1·2), by a sequence of S conditional maximization steps, that is CM-steps, each of which maximizes the Q function defined in (1·1) over θ but with some vector function of θ, $g_s(\theta)$ $(s=1, \ldots, S)$ fixed at its previous value. The general mathematical expressions, given in § 3, involve detailed notation, but it is easy to convey the basic idea. Suppose, as in our first example in § 2, that the parameter θ is partitioned into subvectors $\theta=(\theta_1, \ldots, \theta_S)$. In many applications it is useful to take the sth of the CM-steps to be maximization with respect to θ_s with all other parameters held fixed, whence $g_s(\theta)$ is the vector consisting of all the subvectors except θ_s. In this case, the sequence of CM-steps is equivalent to a cycle of the complete-data iterative-conditional-modes algorithm (Besag, 1986), which, if the modes are obtained by finding the roots of score functions, can also be viewed as a Gauss–Seidel iteration in an appropriate order, e.g. Thisted (1988, Ch. 4). Alternatively, it may be

useful in other applications to take the sth of the CM-steps to be simultaneous maximization over all of the subvectors except for θ_s, which is fixed, implying $g_s(\theta) = \theta_s$. Other choices for the functions g_s, perhaps corresponding to different partitions of θ at each CM-step, can also be useful, as illustrated by our second example in § 2.

Since each CM-step increases Q, it is easy to see that ECM is a GEM algorithm and therefore, like EM, monotonely increases the likelihood of θ. Furthermore, when the set of g_s is 'space-filling' in the sense of allowing unconstrained maximization over θ in its parameter space, ECM converges to a stationary point under essentially the same conditions that guarantee the convergence of EM. To establish this precisely requires formal work presented in §§ 3 and 4, but to see this intuitively, suppose that ECM has converged to θ^* and that the required derivatives of Q are all well defined; the stationarity of each ECM step implies that the corresponding directional derivatives of Q at θ^* are zero, which, under the space-filling condition on $\{g_s, s = 1, \ldots, S\}$, implies that the vector derivative of Q with respect to θ is zero at θ^*, just as with the M-step of EM. Thus, as with EM theory, if ECM converges to θ^*, θ^* must be a stationary point of L_{obs}.

Following a presentation of motivating examples in § 2, §§ 3 and 4 provide the formal treatment of the algorithm with mathematical definitions and convergence results, respectively. Section 5 then offers discussion on variations of ECM and briefly comments on ECM's relationships with other iterative techniques.

2. Motivating examples

The key idea underlying the ECM algorithm can be easily illustrated by the following three examples, which share the common feature that even with complete data, maximum likelihood estimation requires multidimensional numerical iteration, but when the parameters are restricted to particular subspaces, the resultant conditional maximizations either have analytical solutions or require lower dimensional, typically one-dimensional, iteration. Example 1 illustrates ECM in a simple but rather general model in which partitioning the parameter into a location parameter, θ_1, and a scale parameter, θ_2, leads to a straightforward ECM with two CM-steps, each involving closed-form maximization over one of the parameters while holding the other fixed, instead of an iterative M-step as with EM. The second example, a log-linear model for a 3-way contingency table, is also simple but illustrates two additional features of ECM: first, that more than two CM-steps may be useful, and secondly, that the g_s functions do not have to correspond to a simple partition of the parameter, θ. The third example illustrates that, even if some CM-steps do not have analytical solutions, ECM may still have the advantage of being computationally simpler and more stable because it involves lower-dimensional maximizations than EM.

Example 1: *A multivariate normal regression model with incomplete data.* Suppose we have n independent observations from the following k-variate normal model

$$Y_i \sim N(X_i\beta, \Sigma) \quad (i = 1, \ldots, n), \tag{2·1}$$

where X_i is a known $(k \times p)$ design matrix for the ith observation, β is a $(p \times 1)$ vector of unknown regression coefficients, and Σ is a $(k \times k)$ unknown variance-covariance matrix. By specifying particular mean structures and covariance structures, model (2·1) includes important complete-data models, such as seemingly unrelated regressions (Zellner, 1962) and general repeated measures (Jennrich & Schluchter, 1986), as special cases. It is known, however, that maximum likelihood estimation of $\theta = (\beta, \Sigma)$ is generally not

in closed form except in special cases, as when $\Sigma = \sigma^2 I$, e.g. Szatrowski (1978). This result implies that, generally, multidimensional numerical iteration is inevitable in implementing the M-step of EM if it is employed to fit model (2·1) with incomplete data, such as multivariate-normal stochastic-censoring models (Little & Rubin, 1987, Ch. 11).

For simplicity of presentation, consider the case where Σ is unstructured. Although the joint maximizing values of β and Σ are not generally in closed form, we note that if Σ were known, say $\Sigma = \Sigma^{(t)}$, then the conditional maximum likelihood estimate of β would be simply the weighted least-squares estimate:

$$\beta^{(t+1)} = \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(t)})^{-1} X_i \right\}^{-1} \left\{ \sum_{i=1}^{n} X_i^{\mathrm{T}} (\Sigma^{(t)})^{-1} Y_i \right\}. \tag{2·2}$$

On the other hand, given $\beta = \beta^{(t+1)}$, the conditional maximum likelihood estimate of Σ can be obtained directly from the cross-products of the residuals:

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta^{(t+1)})(Y_i - X_i \beta^{(t+1)})^{\mathrm{T}}. \tag{2·3}$$

Clearly, the log-likelihood function is increased by each conditional maximization (2·2) and (2·3):

$$L(\beta^{(t+1)}, \Sigma^{(t)} | Y) \geqslant L(\beta^{(t)}, \Sigma^{(t)} | Y),$$
$$L(\beta^{(t+1)}, \Sigma^{(t+1)} | Y) \geqslant L(\beta^{(t+1)}, \Sigma^{(t)} | Y).$$

These observations lead to the basic formulation of the ECM algorithm, which replaces the original M-step with the two CM-steps given by (2·2) and (2·3). More specifically, at the $(t+1)$st iteration of ECM, one first performs the same E-step as with EM, i.e. find the conditional expectation of the complete-data sufficient statistics; in this example, $E(Y_i | Y_{\mathrm{obs}}, \theta^{(t)})$ and $E(Y_i Y_i^{\mathrm{T}} | Y_{\mathrm{obs}}, \theta^{(t)})$ $(i = 1, \ldots, n)$, where $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)})$. Then one performs the first CM-step, which calculates $\beta^{(t+1)}$ using (2·2) with Y_i being replaced by $E(Y_i | Y_{\mathrm{obs}}, \theta^{(t)})$. Having obtained $\beta^{(t+1)}$, one then performs the second CM-step, which calculates $\Sigma^{(t+1)}$ using (2·3) where Y_i and $Y_i Y_i^{\mathrm{T}}$ on the right-hand side are replaced with $E(Y_i | Y_{\mathrm{obs}}, \theta^{(t)})$ and $E(Y_i Y_i^{\mathrm{T}} | Y_{\mathrm{obs}}, \theta^{(t)})$, respectively. Thus one iteration of ECM for this example consists of one E-step and two CM-steps, none of which requires numerical iteration. The ECM algorithm in this example can be viewed as an efficient generalization of iteratively reweighted least squares, e.g. Rubin (1983), in the presence of incomplete data.

Example 2: *A log-linear model for contingency tables with incomplete data.* It is well known that certain log-linear models do not have closed-form maximum likelihood estimates even with complete data, for example, the no three-way interaction model for a $2 \times 2 \times 2$ table. A well-known iterative algorithm for fitting these kinds of models is Iterative Proportional Fitting, e.g. Bishop, Fienberg & Holland (1975, Ch. 3). Let θ_{ijk} be the probability in cell ijk $(i, j, k = 1, 2)$, where the parameter space Θ is the subspace of $\{\theta_{ijk},\ i, j, k = 1, 2\}$ such that the three-way interaction is zero. In our notation, starting from the constant table (i.e. $\theta_{ijk}^{(0)} = \frac{1}{8}$), given the fully observed cell counts $Y = \{y_{ijk}\}$ and current estimated cell probabilities $\{\theta_{ijk}^{(t)}\}$, the $(t+1)$st iteration of Iterative Proportional Fitting is the final output of the following set of three steps:

$$\theta_{ijk}^{(t+1/3)} = \theta_{ij(k)}^{(t)} \cdot \frac{y_{ij+}}{N}, \tag{2·4}$$

$$\theta_{ijk}^{(t+2/3)} = \theta_{i(j)k}^{(t+1/3)} \cdot \frac{y_{i+k}}{N}, \tag{2·5}$$

$$\theta_{ijk}^{(t+3/3)} = \theta_{(i)jk}^{(t+2/3)} \cdot \frac{y_{+jk}}{N}, \tag{2·6}$$

where N is the total count, $y_{ij+} = \Sigma_k y_{ijk}$ define the two-way marginal table for the first two factors, $\theta_{ij(k)} = \theta_{ijk}/\Sigma_k \theta_{ijk}$ define the conditional probabilities of the third factor given the first two, etc. It is easy to see that (2·4) corresponds to maximizing the log-likelihood $L(\theta \mid Y)$ subject to the constraints $\theta_{ij(k)} = \theta_{ij(k)}^{(t)}$ for all i, j, k. Similarly, expressions (2·5) and (2·6) correspond to maximizing the log-likelihood $L(\theta \mid Y)$ subject to $\theta_{i(j)k} = \theta_{i(j)k}^{(t+1/3)}$ and $\theta_{(i)jk} = \theta_{(i)jk}^{(t+2/3)}$, respectively. The simplicity of Iterative Proportional Fitting comes from the facts that (a) the constraint of 'no three-way interaction' only imposes restrictions on the conditional probabilities $(\theta_{ij(k)}, \theta_{i(j)k}, \theta_{(i)jk})$, and thus, once these conditional probabilities are given, the conditional maximum likelihood estimates for the two-way marginal probabilities $(\theta_{ij+}, \theta_{i+k}, \theta_{+jk})$ are simply the sample proportions, and (b) if $\theta^{(0)} \in \Theta$, then all $\theta^{(t)} \in \Theta$, so starting from a table of constant probabilities will yield the appropriate maximum likelihood estimates.

Once we identify each iteration of Iterative Proportional Fitting as a set of conditional maximizations, we can immediately add an E-step at each iteration to fit the log-linear model to contingency tables with incomplete data. For instance, the only difference between ECM and Iterative Proportional Fitting for the above example is to replace y_{ij+} by $E(y_{ij+} \mid y_{\text{obs}}, \theta^{(t)})$, with analogous replacements for y_{i+k} and y_{+jk} at each iteration. Thus, in this case, ECM can be viewed as a natural generalization of Iterative Proportional Fitting in the presence of incomplete data.

Example 3: *A gamma model with incomplete data.* Suppose our complete data $Y = (y_1, \ldots, y_n)$ are a simple random sample from a Gamma density

$$f(y) = \frac{y^{\alpha-1} \exp(-y/\beta)}{\beta^\alpha \Gamma(\alpha)} \quad (\alpha > 0, \beta > 0),$$

and we are interested in finding the maximum likelihood estimates of α and β based on the observed data Y_{obs}, which, for example, are the result of censoring the complete data, Y.

With complete data, the log-likelihood is

$$L(\alpha, \beta \mid Y) = (\alpha - 1) \sum_{i=1}^{n} \log y_i - \frac{1}{\beta} \sum_{i=1}^{n} y_i - n\{\alpha \log \beta + \log \Gamma(\alpha)\}, \tag{2·7}$$

which does not possess closed-form maximum likelihood estimates. But it is easy to derive from (2·7) that the conditional maximum likelihood estimate for β given $\alpha = \alpha^{(t)}$ is

$$\beta^{(t+1)} = \frac{\bar{y}}{\alpha^{(t)}}, \tag{2·8}$$

where $\bar{y}$ is the sample mean of the complete data. On the other hand, given $\beta = \beta^{(t+1)}$, the conditional maximum likelihood estimate for α, $\alpha^{(t+1)}$, satisfies the following equation

$$\alpha^{(t+1)} = \psi^{-1}(\bar{g} - \log \beta^{(t+1)}), \tag{2·9}$$

where $\bar{g}$ is the sample average of $\{\log y_1, \ldots, \log y_n\}$, and ψ^{-1} is the inverse digamma function, $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$. Although (2·9) does not provide a standard analytic solution

for $\alpha^{(t+1)}$, a value can be easily obtained by a one-dimensional Newton–Raphson algorithm, e.g. Jensen, Johansen & Lauritzen (1991), or by having a general subroutine for the inverse digamma function. Iterating between (2·8) and (2·9) gives the maximum likelihood estimate of $\theta = (\alpha, \beta)$ from the observed sufficient statistics $\bar{y}$ and $\bar{g}$.

With incomplete data, implementing ECM simply means, in the E-step, replacing $\bar{y}$ in (2·8) and $\bar{g}$ in (2·9) by their corresponding conditional expectations given $\theta^{(t)} = (\alpha^{(t)}, \beta^{(t)})$ and the observed data, Y_{obs}. The direct application of EM to this problem would require, say, applying two-dimensional Newon–Raphson to (2·7), which is typically less stable than one-dimensional Newton–Raphson.

3. FORMAL DEFINITION OF THE ECM ALGORITHM

The ECM algorithm replaces the original M-step of EM with several CM-steps. In Examples 1 and 3, the number of CM-steps is $S = 2$, and, in Example 2, $S = 3$. Associated with each of these S CM-steps is a function of θ that is conditional on (or constrained) when maximizing $Q(\theta \mid \theta^{(t)})$ of (1·1). For instance, in Example 1, $\theta_1 = \beta$, $\theta_2 = \Sigma$, and, in Example 3, $\theta_1 = \beta$, $\theta_2 = \alpha$. In both examples, the first CM-step corresponds to maximizing $Q(\theta \mid \theta^{(t)})$ subject to the constraint $g_1(\theta) = g_1(\theta^{(t)})$ where $g_1(\theta) = \theta_2$. The output from the first CM-step can be denoted by $\theta^{(t+\frac{1}{2})} = (\theta_1^{(t+1)}, \theta_2^{(t)})$. Given $\theta^{(t+\frac{1}{2})}$, the second CM-step then maximizes $Q(\theta \mid \theta^{(t)})$ subject to the constraint $g_2(\theta) = g_2(\theta^{(t+\frac{1}{2})})$, where $g_2(\theta) = \theta_1$. Similarly, for Example 2, one can take $g_1(\theta) = \{\theta_{ij(k)}\}$, $g_2(\theta) = \{\theta_{i(j)k}\}$ and $g_3(\theta) = \{\theta_{(i)jk}\}$, implying maximization over $\{\theta_{ij+}\}$, $\{\theta_{i+k}\}$ and $\{\theta_{+jk}\}$, respectively; notice in this example the partition of θ changes across CM-steps.

In general, let

$$G = \{g_s(\theta);\ s = 1, \ldots, S\} \tag{3·1}$$

be a set of S pre-selected (vector) functions of θ. Starting with $\theta^{(0)} \in \Theta$, at the $(t+1)$st iteration, $t = 0, 1, \ldots$, the ECM algorithm first performs the E-step in (1·1) and then S CM-steps instead of the M-step in (1·2), where the CM-steps are defined as follows. For $s = 1, \ldots, S$, find $\theta^{(t+s/S)}$ that maximizes $Q(\theta \mid \theta^{(t)})$ over $\theta \in \Theta$ subject to the constraint $g_s(\theta) = g_s(\theta^{\{t+(s-1)/S\}})$. That is, for $s = 1, \ldots, S$, the sth CM-step in the tth iteration of ECM finds $\theta^{(t+s/S)}$ such that

$$Q(\theta^{(t+s/S)} \mid \theta^{(t)}) \geqslant Q(\theta \mid \theta^{(t)}),$$
$$\text{for all } \theta \in \Theta_s(\theta^{\{t+(s-1)/S\}}) \equiv \{\theta \in \Theta:\ g_s(\theta) = g_s(\theta^{\{t+(s-1)/S\}})\}. \tag{3·2}$$

Then the value of θ for starting the next iteration of ECM, $\theta^{(t+1)}$, is defined as the output of the final step of (3·2), that is $\theta^{(t+S/S)} \equiv \theta^{(t+1)}$.

Definition 1. An iterative algorithm is called an ECM algorithm if the $(t+1)$st iteration starts with an E-step, which finds $Q(\theta \mid \theta^{(t)})$ as a function of θ as in (1·1), and is followed by S ($\geqslant 1$) CM-steps, each of which finds $\theta^{(t+s/S)}$ as in (3·2), for $s = 1, \ldots, S$.

In order to guarantee that ECM converges appropriately just as EM would do, certain restrictions are needed on the set of constraint functions, G, so that the resulting maximum is an unconstrained maximum of L_{obs} in Θ. This can be achieved by requiring G to be 'space filling', as made precise by Definition 2 below.

Definition 2. Let $T_s(\theta)$ $(s = 1, \ldots, S)$ be the set of all feasible directions at $\theta \in \Theta$ with respect to the constraint space

$$\Theta_s(\theta) = \{\zeta \in \Theta:\ g_s(\zeta) = g_s(\theta)\}, \tag{3·3}$$

that is,

$$T_s(\theta)=\left\{\eta\in R^d:\exists\{\theta_n\}\subset\Theta_s(\theta)\text{ such that }\eta=\lim_{n\to\infty}\frac{\theta_n-\theta}{\|\theta_n-\theta\|}\right\}. \tag{3·4}$$

We say $G=\{g_s,\ s=1,\ldots,S\}$ is 'space filling' at $\theta\in\Theta$ if

$$T(\theta)\equiv\text{closure}\left\{\sum_{s=1}^{S}a_s\eta_s:\ a_s\geqslant 0,\ \eta_s\in T_s(\theta)\right\}=R^d. \tag{3·5}$$

In the optimization literature, $T_s(\theta)$ is often referred to as a 'tangent cone' since $\eta\in T_s(\theta)$ implies $\alpha\eta\in T_s(\theta)$ for any $\alpha>0$. The intuition behind (3·5) is simply that, at any point inside Θ, one must be able to search in any direction for the maximum, so that the resulting maximization is over the original parameter space Θ and not constrained to a subspace of Θ.

To avoid unnecessary complications, we assume that $g_s(\theta)$ $(s=1,\ldots,S)$ is differentiable and the corresponding gradient, $\nabla g_s(\theta)$, is of full rank at $\theta\in\Theta_0$, the interior of Θ. As illustrated by our examples, this condition is typically satisfied in practice. Under this assumption, one can show that (3·5) is equivalent to

$$J(\theta)\equiv\bigcap_{s=1}^{S}J_s(\theta)=\{0\}, \tag{3·6}$$

where $J_s(\theta)$ is the column space of the gradient of $g_s(\theta)$, that is,

$$J_s(\theta)=\{\nabla g_s(\theta)\lambda:\ \lambda\in R^{d_s}\}$$

and d_s is the dimensionality of the vector function $g_s(\theta)$. Equation (3·6) is a direct consequence of the following identity,

$$J(\theta)=\{\xi:\ \xi^{\mathrm{T}}\eta\leqslant 0\text{ for all }\eta\in T(\theta)\}, \tag{3·7}$$

which itself follows directly from the polar and bipolar theorems in the literature of constrained optimization, e.g. Fletcher (1980, Ch. 9), Lay (1982, Ch. 9). The advantage of expression (3·6) over (3·5) is that it can be verified directly in many applications. For instance, in Examples 1 and 3, $J_1(\theta)$ is orthogonal to $J_2(\theta)$ for any θ, and thus (3·6) holds for any $\theta\in\Theta$. The verification of condition (3·6) for Example 2 is also straightforward (Meng & Rubin, 1991a). When the EM algorithm itself is viewed as a special case of ECM with $S=1$ and $g_1(\theta)\equiv$ constant, condition (3·6) is automatically satisfied since $\nabla g_1(\theta)=0$ for all θ.

4. Main convergence properties of ECM

Since $\theta^{\{t+(s-1)/S\}}\in\Theta_s(\theta^{\{t+(s-1)/S\}})$, by induction, (3·2) implies

$$Q(\theta^{(t+1)}|\,\theta^{(t)})\geqslant Q(\theta^{(t)}|\,\theta^{(t)}), \tag{4·1}$$

and thus we have the following.

Theorem 1. *Any* ECM *is a* GEM.

As a result of Theorem 1, any property established by Dempster et al. (1977) and Wu (1983) for GEM holds for ECM. In particular, if the sequence $\{L_{\text{obs}}(\theta^{(t)}|\,Y_{\text{obs}}),\ t\geqslant 0\}$ is bounded above, then it converges monotonically to some value L^*, which in general is not necessarily a stationary value of L_{obs}. When G is space filling at each iteration, however, we can show that ECM converges to a stationary point just as EM does, under

the same regularity conditions that Wu (1983) used for establishing the main convergence results for EM. Specifically, we assume his conditions (6)–(10) with appropriate adjustment of notation. The following result is a direct extension of Wu's Theorem 2 (1983) to any ECM sequence. Notice that Wu's regularity condition (9) guarantees that, if the initial value $\theta^{(0)} \in \Theta$, then all iterates $\theta^{(t)} \in \Theta_0$, so all the following calculations are performed inside Θ_0.

THEOREM 2. *Suppose that all the conditional maximizations in* (3·2) *of* ECM *are unique. Then all limit points of any* ECM *sequence* $\{\theta^{(t)}, t \geq 0\}$ *belong to the set*

$$\Gamma \equiv \left\{ \theta \in \Theta: \frac{\partial L_{\text{obs}}(\theta \mid Y_{\text{obs}})}{\partial \theta} \in J(\theta) \right\}.$$

Proof. By Theorem 1 of Wu (1983), we only need to show that:

(i) the mapping defined by ECM is a closed mapping; and

(ii) if $\theta^{(t)} \notin \Gamma$, then

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) > Q(\theta^{(t)} \mid \theta^{(t)}). \tag{4·2}$$

Under the compactness condition (6) and continuous condition (10) of Wu, assertion (i) can be verified directly by using the fact that if $\theta_k \to \theta$ then, for any $\theta' \in \Theta_s(\theta)$ $(s = 1, \ldots, S)$, where $\Theta_s(\theta)$ is defined in (3·3), there exists $\theta'_k \in \Theta_s(\theta_k)$ such that $\theta'_k \to \theta'$. This fact is a consequence of the inverse mapping Theorem, e.g. Rudin (1964, Ch. 9), which is applicable here since $\nabla g_s(\theta)$ is of full rank at all $\theta \in \Theta_0$.

We now prove (ii) by contradiction. Suppose (4·2) does not hold for some $\theta^{(t)} \notin \Gamma$. Then by (4·1) and (4·2)

$$Q(\theta^{(t+s/S)} \mid \theta^{(t)}) = Q(\theta^{\{t+(s-1)/S\}} \mid \theta^{(t)}), \quad \text{for all } s = 1, \ldots, S,$$

which implies, by the assumption of the uniqueness of all conditional maximizations, that

$$\theta^{(t+1)} = \theta^{\{t+(S-1)/S\}} = \ldots = \theta^{\{t+(1/S)\}} = \theta^{(t)}. \tag{4·3}$$

In other words, for all s, $\theta^{(t)}$ is the maximizer of $Q(\theta \mid \theta^{(t)})$ under the constraint $\theta \in \Theta_s(\theta^{(t)})$, which implies that, at $\theta^{(t)}$, $Q(\theta \mid \theta^{(t)})$ decreases along any feasible direction determined by $\Theta_s(\theta^{(t)})$ for all s, and thus

$$D^{10}Q(\theta^{(t)} \mid \theta^{(t)})\eta \leq 0, \quad \text{for all } \eta \in T_s(\theta^{(t)}),\ s = 1, \ldots, S, \tag{4·4}$$

where D^{10} denotes the first order derivative with respect to the first argument of Q. By the definition of $T(\theta^{(t)})$ in (3·5), we have

$$D^{10}Q(\theta^{(t)} \mid \theta^{(t)})\eta \leq 0, \quad \text{for all } \eta \in T(\theta^{(t)}). \tag{4·5}$$

Since (Dempster et al., 1977)

$$D^{10}Q(\theta^{(t)} \mid \theta^{(t)}) = \frac{\partial L_{\text{obs}}(\theta^{(t)} \mid Y_{\text{obs}})}{\partial \theta}, \tag{4·6}$$

(3·7) and (4·5) together imply that

$$\frac{\partial L_{\text{obs}}(\theta^{(t)} \mid Y_{\text{obs}})}{\partial \theta} \in J(\theta^{(t)}),$$

which contradicts $\theta^{(t)} \notin \Gamma$. □

When G is space filling at θ, $J(\theta) = \{0\}$, and thus Theorem 2 guarantees the following result.

THEOREM 3. *Suppose that all the conditional maximizations in* (3·2) *of* ECM *are unique. Then all limit points of any* ECM *sequence* $\{\theta^{(t)}, t \geq 0\}$ *are stationary points of* $L_{\text{obs}}(\theta \mid Y_{\text{obs}})$ *if* G *is space filling at all* $\theta^{(t)}$.

The assumption of Theorems 2 and 3 that all conditional maximizations are unique is very weak in the sense that it is satisfied in many practical problems, but even this condition can be eliminated if we force $\theta^{(t+s/S)} = \theta^{\{t+(s-1)/S\}}$ wherever there is no increase in $Q(\theta \mid \theta^{(t)})$ at the sth CM-step. Alternatively, we can extend Wu's Theorem 6 (1983) to any ECM sequence by replacing the uniqueness condition with (a) the continuity of $D^{10}Q(\theta \mid \theta')$ in both θ and θ', and (b) the continuity of $\nabla g_s(\theta)$ for all s, which are also typically satisfied in practical applications. The crucial property that makes this extension possible is again the space-filling condition on G.

It is known that, in general, neither EM nor any optimization algorithm is guaranteed to converge to a global or local maximum, and ECM is not magical in this regard. Wu (1983) gave a number of conditions under which an EM sequence will converge to a local maximum. Almost all of these results can be extended to ECM with little difficulty. Among them, the following result, which is a direct consequence of Theorem 3 under the uniqueness condition or of the extension of Wu's Theorem 6 under the continuity conditions, is most useful since it covers many practical applications.

COROLLARY 1. *Suppose that* $L_{\text{obs}}(\theta \mid Y_{\text{obs}})$ *is unimodal in* $\theta \in \Theta$ *with* θ^* *being the only stationary point. Then any* ECM *sequence* $\{\theta^{(t)}\}$ *converges to the unique maximizer* θ^* *if* G *is space filling at all* $\theta^{(t)}$, *and either* (*a*) *each* CM *maximization is unique or* (*b*) $D^{10}Q(\theta \mid \theta')$ *is continuous in both* θ *and* θ' *and* $\nabla g_s(\theta)$ *is continuous in* θ *for* $s = 1, \ldots, S$.

It is clear that the general results in Theorem 2 and Theorem 3 also apply to the CM algorithm, that is ECM without missing data, in which case the E-step becomes an identity operation: $Q(\theta \mid \theta^{(t)}) \equiv L(\theta \mid Y)$. Two related issues are worth mentioning. First, if the set of constraint functions, G, is not space-filling, then, as shown in Theorem 2, CM will converge to a stationary point of the likelihood in a subspace of Θ, which may or may not be a stationary point of the likelihood in the whole parameter space. Thus, except for pathological cases, for a fixed density with parameter space Θ, one can construct a data set such that the corresponding CM sequence does not converge to a maximum of the likelihood in Θ. In this sense, the space-filling condition is not only sufficient but also necessary.

Secondly, since the space-filling condition on G does not involve data, one would expect that, if G leads to appropriate convergence of CM with complete data, it should also lead to appropriate convergence of ECM with missing data. This conjecture can be proved easily and rigorously when the complete-data density is from an exponential family, where ECM is especially useful. The advantage of the following Theorem 4 is that it enables us to conclude that ECM will converge appropriately whenever CM does so. For instance, one can immediately conclude the appropriate convergence of ECM in Example 2 without having to verify the space-filling condition, because the monotone convergence of Iterative Proportional Fitting with complete data has been established (Bishop et al., 1975, Ch. 3).

THEOREM 4. *Suppose the complete-data density is from an exponential family and the set* G *of* (3·1) *is chosen such that any corresponding* CM *sequence strictly increases the complete-data likelihood at each iteration until it reaches a stationary point. Then all the*

limit points of any corresponding ECM *sequence are stationary points of the observed-data log-likelihood,* $L_{\text{obs}}(\theta \mid Y_{\text{obs}})$.

Proof. We only need to prove (i) and (ii) with $J(\theta) = \{0\}$ in the proof of Theorem 2. The proof for (i) is unchanged because it does not involve the space-filling condition. To prove (ii) notice that, because $L(\theta \mid Y)$ is from an exponential family, we have

$$Q(\theta \mid \theta^{(t)}) = L(\theta \mid \mathcal{S}^{(t)}), \tag{4·7}$$

where $\mathcal{S}^{(t)} = E(\mathcal{S}(Y) \mid Y_{\text{obs}}, \theta^{(t)})$ with $\mathcal{S}(Y)$ being the vector of the complete-data sufficient statistics. Thus, if $\theta^{(t)}$ is not a stationary point of $L_{\text{obs}}(\theta \mid Y_{\text{obs}})$, then (4·6) and (4·7) together imply that $\theta^{(t)}$ cannot be a stationary point of $L(\theta \mid \mathcal{S}^{(t)})$. Therefore, the next iterate $\theta^{(t+1)}$ will strictly increase L by our assumption on the CM sequence, and thus (4·2) follows from (4·7). □

5. DISCUSSION

In the absence of missing data, ECM is a special case of the cyclic coordinate ascent method for function maximization in the optimization literature, e.g. Zangwill (1969, Ch. 5); also see Haberman (1974, Ch. 3) on Iterative Proportional Fitting. Although these optimization methods are well known for their simplicity and stability, because they typically converge only linearly, they have been less preferred in practice for handling complete-data problems than superlinear methods like Newton–Raphson. When used for the M-step of EM or a CM-step of ECM, however, simple and stable linear converging methods are often more suitable than superlinear converging but less stable algorithms. The reasons are first, that the advantage of superlinear convergence in each M- or CM-step does not transfer to the overall convergence of EM or ECM since EM and ECM always converge linearly regardless of the maximization method employed within the maximization step, and secondly, that the stability of the maximization method is critical for preserving the stability of EM or ECM since it is used repeatedly within each maximization step in all iterations. Finally, if one performs just one iteration of a superlinear converging algorithm within each M-step of EM, then the resulting algorithm is no longer guaranteed to increase the likelihood monotonely.

In some cases, the computation of an E-step may be much cheaper than the computation of the CM-steps, and one might wish to perform an E-step before each CM-step. Each iteration then involves S 'cycles', where a cycle is defined by one E-step followed by one CM-step. At the sth cycle of iteration $(t+1)$, the E-step finds

$$Q(\theta \mid \theta^{\{t+(s-1)/S\}}) = \int L(\theta \mid Y) f(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta = \theta^{\{t+(s-1)/S\}})\, dY_{\text{mis}}, \tag{5·1}$$

as a function of θ for fixed Y_{obs} and fixed $\theta^{\{t+(s-1)/S\}}$, and the CM-step finds $\theta^{(t+s/S)}$ to maximize this function:

$$Q(\theta^{(t+s/S)} \mid \theta^{\{t+(s-1)/S\}}) \geqslant Q(\theta \mid \theta^{\{t+(s-1)/S\}}), \quad \text{for all } \theta \in \Theta_s(\theta^{\{t+(s-1)/S\}}). \tag{5·2}$$

Alternatively, one can perform an E-step only before a few selected CM-steps, but for descriptive simplicity we focus here on the case with an E-step preceding each CM step, and we call the corresponding algorithm a 'multi-cycle ECM'. For instance, in Example 2, with multi-cycle ECM, y_{i+k} of (2·5) and y_{+jk} of (2·6) will be replaced by $E(y_{i+k} \mid y_{\text{obs}}, \theta^{\{t+(1/3)\}})$ and $E(y_{+jk} \mid y_{\text{obs}}, \theta^{\{t+(2/3)\}})$, respectively at the $(t+1)$st iteration,

instead of being, as with ECM, replaced by $E(y_{i+k}|y_{\text{obs}}, \theta^{(t)})$ and $E(y_{+jk}|y_{\text{obs}}, \theta^{(t)})$, respectively.

Since the second argument in the Q function is changing at each cycle within each iteration, a multi-cycle ECM may not be a GEM. The inequality (5·2), however, implies

$$Q(\theta^{(t+s/S)}|\theta^{\{t+(s-1)/S\}}) \geqslant Q(\theta^{\{t+(s-1)/S\}}|\theta^{\{t+(s-1)/S\}}) \quad (s=1,\ldots,S). \tag{5·3}$$

Expression (5·3) may be taken as the definition of an extended GEM, having iterations indexed by t, each of which consists of S distinct cycles indexed by s. Using the same argument for proving that GEM always increases L_{obs}, one can easily show that (5·3) implies an extended GEM algorithm increases L_{obs} at each cycle and thus increases L_{obs} at each iteration. Thus, just as with ECM, for any multi-cycle ECM sequence $\{\theta^{(t)}, t \geqslant 0\}$, $\{L_{\text{obs}}(\theta^{(t)}|Y_{\text{obs}}), t \geqslant 0\}$ converges monotonically to some L^* if the sequence itself is bounded above. All other results on ECM in § 4 apply to multi-cycle ECM.

The obvious disadvantage of using multi-cycle ECM is the extra computation at each iteration. Intuitively, as a trade-off, one might expect it to result in larger increases in L_{obs} per iteration since Q is being updated more often. Practical implementations do show this potential, but it is not true in general. That is, there exist cases where multi-cycle ECM converges more slowly than ECM. In fact, there are even cases where EM converges more slowly than ECM! Details of these examples, which are not typical in practice, and other results on the rate of convergence of these algorithms, and their uses in computing the asymptotic variance-covariance matrix via the SEM algorithm (Meng & Rubin, 1991b), appear elsewhere, e.g. Meng (1994).

A final comment concerns the relationship of ECM to the Gibbs sampler (Geman & Geman, 1984) and other methods of iterative simulation such as the Hastings/Metropolis algorithm, e.g. Hastings (1970). Typically, if such an iterative simulation method can be implemented, so can ECM but with substantially less work and more straightforward convergence properties. As discussed by Meng & Rubin (1992) and Gelman & Rubin (1992), this fact has important implications for the practical use of iterative simulation, because ECM can search out modes and thereby obtain an approximate analytical distribution, which can be used both to start the iterative simulation and to help monitor its convergence.

Acknowledgements

This work was supported in part by several National Science Foundation grants awarded to Harvard University and the University of Chicago, in part by Joint Statistical Agreements between the U.S. Bureau of the Census and Harvard University, and in part by the University of Chicago/AMOCO fund. 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 wish to thank Charles Geyer for providing the references to optimization literature, and Kenneth Lange, Alan Zaslavsky and several referees for very helpful comments and suggestions. We also thank Andrew Gelman for bringing to our attention, after our submission, an unpublished manuscript of Alvaro R. De Pierro, who presented a geometric description of a multi-cycle ECM with a partitioned parameter. Finally we thank Herbert J. A. Hoijtink for pointing out a psychometric application of a multi-cycle ECM in his Ph.D. thesis.

REFERENCES

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

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

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B **39**, 1-38.

FLETCHER, R. (1980). *Practical Methods of Optimization.* New York: Wiley.

GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6**, 721-41.

GELMAN, A. & RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457-511.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data.* University of Chicago Press.

HASTINGS, W. K. (1970). Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

JENNRICH, R. I. & SCHLUCHTER, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics* **42**, 805-20.

JENSEN, S. T., JOHANSEN, S. & LAURITZEN, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78**, 867-77.

LAY, R. S. (1982). *Convex Sets and their Applications.* New York: Wiley.

LITTLE, R. J. A. & RUBIN, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

MENG, X. L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* To appear.

MENG, X. L. & RUBIN, D. B. (1991a). IPF for contingency tables with missing data via the ECM algorithm. In *Proc. Statist. Comp. Sect.*, pp. 244-7. Washington, D.C.: American Statistical Association.

MENG, X. L. & RUBIN, D. B. (1991b). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Assoc.* **86**, 899-909.

MENG, X. L. & RUBIN, D. B. (1992). Recent extensions to the EM algorithm (with discussion). In *Bayesian Statistics 4,* Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 307-20. Oxford University Press.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-92.

RUBIN, D. B. (1983). Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences*, **4**, Ed. S. Kotz and N. L. Johnson, pp. 272-75. New York: Wiley.

RUDIN, W. (1964). *Principles of Mathematical Analysis.* New York: McGraw-Hill.

SZATROWSKI, T. H. (1978). Explicit solutions, one iteration convergence and averaging in the multivariate normal estimation problem for patterned means and covariances. *Ann. Inst. Statist. Math.* **30**, 81-8.

THISTED, R. A. (1988). *Elements of Statistical Computing.* New York: Chapman and Hall.

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.* Englewood Cliffs, New Jersey: Prentice-Hall.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Am. Statist. Assoc.* **57**, 348-68.

[*Received June* 1991. *Revised August* 1992]