Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm

XIAO-LI MENG and DONALD B. RUBIN*

The expectation maximization (EM) algorithm is a popular, and often remarkably simple, method for maximum likelihood estimation in incomplete-data problems. One criticism of EM in practice is that asymptotic variance–covariance matrices for parameters (e.g., standard errors) are not automatic byproducts, as they are when using some other methods, such as Newton–Raphson. In this article we define and illustrate a procedure that obtains numerically stable asymptotic variance–covariance matrices using only the code for computing the complete-data variance–covariance matrix, the code for EM itself, and code for standard matrix operations. The basic idea is to use the fact that the rate of convergence of EM is governed by the fractions of missing information to find the increased variability due to missing information to add to the complete-data variance–covariance matrix. We call this supplemented EM algorithm the SEM algorithm. Theory and particular examples reinforce the conclusion that the SEM algorithm can be a practically important supplement to EM in many problems. SEM is especially useful in multiparameter problems where only a subset of the parameters are affected by missing information and in parallel computing environments. SEM can also be used as a tool for monitoring whether EM has converged to a (local) maximum.

KEY WORDS: Bayesian inference; Convergence rate; EM algorithm; Incomplete data; Maximum likelihood estimation; Observed information; Parallel processors.

1. INTRODUCTION

1.1 The Theme of SEM

Over the past dozen or so years, the expectation maximization (EM) algorithm (Dempster, Laird and Rubin 1977, henceforth DLR) has become a remarkably popular tool in applied statistics and a common topic in many publications in statistics, so common in fact that articles often refer to it without citing any publication for it. A principal reason for this popularity is that it relies on flexible computing environments to find maximum likelihood estimates (MLE's) in complicated problems of missing and incomplete data primarily using complete-data tools: The M step is standard maximum likelihood estimation for complete-data problems, and the E step is usually available from standard complete-data theory of conditional distributions. This idea of capitalizing on computing power and complete-data tools to handle missing-data problems, including random parameter models, is a major theme in much of modern statistics: latent class models (Goodman 1974), missing data in ANOVA (Rubin 1976), mixture models (Titterington, Smith and Makov 1985), multiple imputation (Rubin 1987), data augmentation (Tanner and Wong, 1987), stochastic relaxation (Gelfand and Smith 1990), and so on.

Here we confine ourselves to the EM context but follow this theme of repeated computations using complete-data tools. Specifically, we supplement the maximum likelihood estimates of EM with an associated asymptotic variance–covariance matrix for them, which can be obtained using only the code for the complete-data asymptotic variance–covariance matrix, the code for the EM algorithm, and standard code for matrix operations. In particular, neither likelihoods nor any derivatives of likelihoods or log-likelihoods need to be evaluated. Previously suggested procedures for supplementing EM to obtain an asymptotic variance–covariance matrix have restrictions and limitations that our procedure does not have, as discussed in Section 1.2. We believe that our procedure, which we call the supplemented EM algorithm or the SEM algorithm, will be an important supplement to EM in many contexts, especially in modern computing environments where computer time is inexpensive relative to researcher time and where parallel processing is possible. SEM can also be applied to find the asymptotic variance–covariance matrix when the MLE's are obtained by any other method, such as by factoring the likelihood with special patterns of missing data (Little and Rubin 1987, chap. 6).

It is important to emphasize that the variance–covariance matrix obtained by SEM is based on the second derivatives of the observed-data log-likelihood and thus is guaranteed to be inferentially valid only asymptotically. Consequently, from both frequentist and Bayesian perspectives, the practical propriety of the resulting normal theory inferences is improved when the likelihood function is more nearly normal. Therefore, in practice it is generally wise to define parameterizations with attention to this fact, for example, by using log(variance) rather than variance with normal data. These parameterizations typically also improve the stability of the SEM computations. The use of such transformations of parameters is illustrated in our examples.

1.2 Other Methods

Previous methods for obtaining asymptotic variance–covariance matrices in EM contexts have limitations that

* Xiao-Li Meng is Assistant Professor, Department of Statistics, University of Chicago, Chicago, IL 60637. Donald B. Rubin is Professor and Chairman, Department of Statistics, Harvard University, Cambridge, MA 02138. Most of this research was done when the first author was a graduate student at Harvard and was supported partly by NSF Grant SES-880543 and partly by Joint Statistical Agreements 87-07, 88-02, 89-08, and 90-23 between the U.S. Bureau of the Census and Harvard University. We thank the editor, Clifford Clogg, the anonymous referees, and several colleagues (especially at Harvard University and the University of Chicago) for many helpful comments and suggestions. A preliminary and limited version, Meng and Rubin (1989), appeared in *Proceedings of the Statistical Computing Section of the American Statistical Association*. It has recently been brought to our attention that "SEM" has also been used by Celeux and Diebolt (1985) to stand for the "Stochastic EM" algorithm. Moreover, in econometrics, "SEM" is frequently used to stand for "structural equations models."

Journal of the American Statistical Association
December 1991, Vol. 86, No. 416, Applications and Case Studies

make them less automatically applicable than SEM. The technique described in Louis (1982), for instance, requires, in addition to the code for the complete-data variance–covariance matrix and the code for the E and M steps, calculation of the conditional expectation (conditional on the observed data) of the square of the complete-data score function, which is specific to each problem. Algebraic analysis is often tedious or intractable, as pointed out in Meilijson (1989). Monte Carlo evaluation can be accomplished using multiple imputation (Rubin 1987) of the missing data given the maximum likelihood estimate, but then this requires new code for drawing the imputations, and its accuracy depends on the number of imputations.

Meilijson's (1989) techniques, although like SEM in requiring no additional analysis, are unlike SEM in that they apply only to specialized cases in which the observed data are iid (independently and identically distributed) samples. Thus, as illustrated in Section 4, they cannot be automatically applied to general patterns of missing values even with iid complete data (e.g., normal or multinominal), or to many hyperparameter estimation problems. SEM can be applied to any problem to which EM has been applied, assuming one has access to the complete-data asymptotic variance–covariance matrix.

Methods such as those described by Carlin (1987), which obtain the second derivative of the observed-data log-likelihood function by numerical differentiation, not only require evaluation of this log-likelihood function (not required by SEM) but are subject to the inaccuracies and difficulties of any numerical differentiation procedure with large matrices. Although SEM involves numerical differentiation, it is not used to obtain the desired variance–covariance matrix but only the increases due to missing data that are to be added to the complete-data variance–covariance matrix. Therefore, SEM is typically more stable than pure numerical differentiation procedures because the matrix obtained by numerical differentiation is being added to an analytically obtained matrix, which is usually the dominant term. In cases with large increases in variance due to large amounts of missing information, the sequence of evaluations required by SEM appears to be very stable, with an internal check on numerical accuracy provided by the observed symmetry of the resulting variance–covariance matrix. We will discuss this further in Sections 5.2 and 5.3.

A final possibility is to obtain variance–covariance matrices by techniques such as the bootstrap or jackknife, which resample data sets from the empirical distribution and perform EM on each such data set. Such procedures might work well in large samples with iid structures, but their definition and performance is unclear in non-iid cases involving complicated patterns of missing data or models with several levels of randomness (e.g., variance components, latent structure, or empirical Bayes models). Furthermore, the use of such techniques can be viewed as unappealing in some likelihood or Bayesian contexts in which the desired measure of inferential uncertainty is the observed information matrix, which at best is only approximated by resampling estimates. SEM provides this matrix without the restrictions or limitations of these other techniques.

1.3 SEM in Single Parameter Cases

In his discussion of DLR, Smith (1977), noted the possibility of obtaining the asymptotic variance for the MLE in single parameter cases by using the rate of convergence of EM. Using a simple example from DLR (see Section 4.2 in this article), he gave the following simple relationship between V, the observed-data asymptotic variance, and V_c, the complete-data asymptotic variance:

$$V = V_c/(1 - r), \tag{1.3.1}$$

where r is the rate of convergence of EM. In other words, the observed-data asymptotic variance can be obtained by inflating the ordinary complete-data asymptotic variance by the factor $1 - r$, where r is readily available from the output of EM. Letting θ be the parameter, θ^* be the MLE of θ, and $\theta^{(t)}$ be the EM estimate of θ at the tth iteration, r can be well approximated by the ratio $(\theta^{(t+1)} - \theta^*)/(\theta^{(t)} - \theta^*)$ when t is large, or equivalently by $(\theta^{(t+1)} - \theta^{(t)})/(\theta^{(t)} - \theta^{(t-1)})$. A statistically more appealing representation of (1.3.1) is

$$V = V_c + \Delta V, \tag{1.3.2}$$

where

$$\Delta V = [r/(1 - r)]\, V_c \tag{1.3.3}$$

is the increase in variance due to missing data.

Our method, SEM, provides a general formulation of this simple procedure and extends it to the multiparameter case. Three issues arise for multiparameter cases: the correct matrix version of (1.3.2), the computation of the matrix version of V_c, and the computation of the matrix version of r. We first establish the matrix version of (1.3.2) in Section 2 after providing necessary background material. Then in Section 3 we describe the SEM algorithm, which addresses the computation of the matrix versions of V_c and r. Examples are provided in Section 4 to illustrate our algorithm, and practical issues (e.g., stopping criteria) when implementing SEM are addressed in Section 5.

2. BACKGROUND AND BASIC RESULTS

2.1 The EM Algorithm

Suppose we have a model for the complete data Y, with associated density $f(Y \mid \theta)$, where $\theta = (\theta_1, \ldots, \theta_d)$ is the unknown parameter. We write $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} represents the observed part of Y and Y_{mis} denotes the missing values. The EM algorithm finds the value of θ, θ^*, that maximizes $f(Y_{obs} \mid \theta)$, that is, the MLE for θ based on the observed data Y_{obs}.

The EM algorithm starts with an initial value $\theta^{(0)}$. Letting $\theta^{(t)}$ be the estimate of θ at the tth iteration, iteration $(t + 1)$ of EM is as follows:

E step. Find the expected complete-data log-likelihood if θ were $\theta^{(t)}$:

$$Q(\theta \mid \theta^{(t)}) = \int L(\theta \mid Y) f(Y_{mis} \mid Y_{obs}, \theta = \theta^{(t)})\, dY_{mis}, \tag{2.1.1}$$

where $L(\theta \mid Y) = \log f(Y \mid \theta)$.

M step. Determine $\theta^{(t+1)}$ by maximizing this expected log-likelihood:

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}), \qquad \text{for all } \theta.$$

The M step of EM is easy to implement in broad classes of problems, such as in exponential families, since it uses the identical computational method as ML estimation from $L(\theta \mid Y)$. The E step of EM is also very easy to implement in many problems, including many exponential family models, since it follows from standard complete-data theory for means of conditional distributions.

2.2 The Rate of Convergence of EM

The EM algorithm just described implicitly defines a mapping $\theta \to M(\theta)$ from the parameter space of θ, Θ, to itself such that

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

If $\theta^{(t)}$ converges to some point θ^* and $M(\theta)$ is continuous, then θ^* must satisfy

$$\theta^* = M(\theta^*).$$

Therefore, in the neighborhood of θ^*, by a Taylor series expansion, we have

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

where

$$DM = \left(\frac{\partial M_j(\theta)}{\partial \theta_i}\right)\Bigg|_{\theta=\theta^*} \tag{2.2.1}$$

is the $d \times d$ Jacobian matrix for $M(\theta) = (M_1(\theta), \ldots, M_d(\theta))$ evaluated at $\theta = \theta^*$. Thus, in the neighborhood of θ^*, the EM algorithm is essentially a linear iteration with rate matrix DM, since DM is typically nonzero.

2.3 The Large-Sample Variance–Covariance Matrix of $(\theta - \theta^*)$ Based on Y_{obs}

It is well known that the large-sample variance–covariance matrix of $(\theta - \theta^*)$ based on Y_{obs}, V, can be found as the inverse of the observed information matrix,

$$V = I_o^{-1}(\theta^* \mid Y_{obs}), \tag{2.3.1}$$

where $I_o(\theta \mid Y_{obs})$ is the negative second derivative of the log-likelihood of θ given Y_{obs},

$$I_o(\theta \mid Y_{obs}) = -\frac{\partial^2 \log f(Y_{obs} \mid \theta)}{\partial\theta \cdot \partial\theta}. \tag{2.3.2}$$

This function can be very difficult to evaluate directly. In contrast, the complete-data observed information matrix,

$$I_o(\theta \mid Y) = -\frac{\partial^2 \log f(Y \mid \theta)}{\partial\theta \cdot \partial\theta}, \tag{2.3.3}$$

whose inverse gives the complete-data variance–covariance matrix, is often a simple function, as is its expectation over the conditional distribution $f(Y_{mis} \mid Y_{obs}, \theta)$ evaluated at $\theta = \theta^*$:

$$I_{oc} = E[I_o(\theta \mid Y) \mid Y_{obs}, \theta]\Big|_{\theta=\theta^*} \tag{2.3.4}$$

Of particular importance, in applications of the EM algorithm it is essentially as simple to evaluate I_{oc}^{-1} as it is to evaluate the complete-data variance–covariance matrix; this fact is shown in Section 3.1.

The matrix I_{oc}^{-1} is important here because the desired observed variance–covariance matrix, V, can be written as a simple function of I_{oc}^{-1} and DM, the matrix rate of convergence of EM as defined in (2.2.1). More specifically, as we will show in the next section,

$$V = I_{oc}^{-1} + \Delta V, \tag{2.3.5}$$

where

$$\Delta V = I_{oc}^{-1}DM(I - DM)^{-1} \tag{2.3.6}$$

is the increase in variance due to missing information, and I is simply the $d \times d$ identity matrix. The numerical evaluation of DM is presented in Sections 3.3 and 3.4.

2.4 Showing That $V = I_{oc}^{-1} + \Delta V$

From the factorization

$$f(Y \mid \theta) = f(Y_{obs} \mid \theta)f(Y_{mis} \mid Y_{obs}, \theta),$$

it follows that the log-likelihood of θ given Y_{obs} is

$$L(\theta \mid Y_{obs}) = L(\theta \mid Y) - \log f(Y_{mis} \mid Y_{obs}, \theta). \tag{2.4.1}$$

Equation (2.4.1) implies, after taking second derivatives, averaging over $f(Y_{mis} \mid Y_{obs}, \theta)$, and evaluating at $\theta = \theta^*$, that

$$I_o(\theta^* \mid Y_{obs}) = I_{oc} - I_{om}, \tag{2.4.2}$$

where the matrix

$$I_{om} = E\left[-\frac{\partial^2 \log f(Y_{mis} \mid Y_{obs}, \theta)}{\partial\theta \cdot \partial\theta} \mid Y_{obs}, \theta\right]\Bigg|_{\theta=\theta^*} \tag{2.4.3}$$

can be viewed as the missing information. Thus, (2.4.2) has the following appealing interpretation:

observed information = complete information

− missing information,

which has been called the "missing information principle" by Orchard and Woodbury (1972).

Equation (2.4.2) can be written as

$$I_o(\theta^* \mid Y_{obs}) = (I - I_{om}I_{oc}^{-1})I_{oc}. \tag{2.4.4}$$

Equation (2.4.4) is useful because DLR showed that, if $Q(\theta \mid \theta^{(t)})$ [defined by (2.1.1)] is maximized in the M step by setting its first derivative equal to zero, as with standard complete-data maximum likelihood estimation, then

$$DM = I_{om}I_{oc}^{-1}. \tag{2.4.5}$$

Substituting (2.4.5) into (2.4.4) and inverting gives

$$V = I_{oc}^{-1}(I - DM)^{-1} \tag{2.4.6}$$

$$= I_{oc}^{-1} + I_{oc}^{-1}DM(I - DM)^{-1}, \tag{2.4.7}$$

which is the same as (2.3.5) and (2.3.6).

3. THE SEM ALGORITHM

3.1 Definition

The SEM algorithm consists of three parts: (1) the evaluation of I_{oc}^{-1}, (2) the evaluation of DM, and (3) the evaluation of V from (2.3.5) and (2.3.6). Each of these parts is now discussed.

3.2 Evaluation of I_{oc}^{-1}

In most common practical applications of EM, the complete-data density $f(Y \mid \theta)$ is from an exponential family, that is

$$f(Y \mid \theta) = b(Y) \exp\{S(Y)C'(\theta)\}/a(\theta),$$

where $S(Y)$ is a $1 \times k$ ($k \geq d$) vector of complete-data sufficient statistics, $C(\theta)$ is a $1 \times k$ vector function of θ, and $b(Y)$ and $a(\theta)$ are scalar functions. This form for $f(Y \mid \theta)$ implies from (2.3.3) that $I_o(\theta \mid Y) = I_o(\theta \mid S(Y))$ is a linear function of $S(Y)$. Thus, from (2.3.4),

$$I_{oc} = I_o(\theta^* \mid S^*(Y_{obs})), \tag{3.2.1}$$

where $S^*(Y_{obs}) = E[S(Y) \mid Y_{obs}, \theta^*]$ is obtained at the last E step since the complete-data log-likelihood $L(\theta \mid Y) = L(\theta \mid S(Y))$ is also a linear function of $S(Y)$. In other words, I_{oc}^{-1} can be obtained simply by substituting the conditional expectation of $S(Y)$ found at the last E step of EM for $S(Y)$ in $I_o^{-1}(\theta^* \mid S(Y))$, which is the standard complete-data variance–covariance matrix evaluated at $\theta = \theta^*$ as a function of the complete-data sufficient statistics.

It is important to emphasize that the complete-data information matrix $I_o(\theta \mid S(Y))$ used previously is the complete-data observed information matrix, not the Fisher information matrix $I(\theta) = E[I_o(\theta \mid Y) \mid \theta]$. The Fisher information matrix is appropriate for our calculations only when $f(Y \mid \theta)$ is from a regular exponential family [that is, when $k = d$ and the Jacobian of $C(\theta)$ is full rank], because in this case it is easy to verify that

$$I_{oc} = I(\theta^*). \tag{3.2.2}$$

When $f(Y \mid \theta)$ is not an exponential family density, the complete-data log-likelihood $L(\theta \mid Y)$ is no longer a linear function of its sufficient statistics, with the result that the expected complete-data log-likelihood, $Q(\theta \mid \theta^{(t)})$, is generally not in closed form as a function of θ. This typically raises some difficulties in directly implementing the E step, because in principle the E step requires us to evaluate $Q(\theta \mid \theta^{(t)})$ separately for each θ in the parameter space Θ. One legitimate way to avoid this difficulty with large samples is to use a Taylor series expansion to linearize $L(\theta \mid Y)$ in terms of large-sample sufficient statistics. Once the linearization is formulated for the E step, our method can be applied directly to compute I_{oc}^{-1}, since $I_o(\theta \mid Y)$ is also a linear function of these large-sample sufficient statistics. In some nonexponential family cases, our method can be applied without resorting to linearization arguments because the conditional density $f(Y_{mis} \mid Y_{obs}, \theta)$ is still from an exponential family. Thus $I_o(\theta \mid Y)$ is a linear function of the conditional sufficient statistics from $f(Y_{mis} \mid Y_{obs}, \theta)$, and these conditional sufficient statistics are the only ones that involve missing data and need updating at each E step.

3.3 Computation of DM

For a vector θ, the simple procedure described in Section 1.3 cannot be used to produce the entire DM matrix because the observed component-wise rates of convergence of EM [e.g., $\lim_{t\to\infty}(\theta_i^{(t+1)} - \theta_i^*)/(\theta_i^{(t)} - \theta_i^*)$, $i = 1, \ldots, d$] provide only a few eigenvalues (in most cases, simply the largest eigenvalue) of DM, not the matrix itself (DLR; Meng 1990). Each element of DM, however, is the component-wise rate of convergence of a "forced EM" in the following sense.

Let r_{ij} be the (i, j)th element of DM and define $\theta^{(t)}(i)$ to be

$$\theta^{(t)}(i) = (\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(t)}, \theta_{i+1}^*, \ldots, \theta_d^*), \tag{3.3.1}$$

that is, only the ith component in $\theta^{(t)}(i)$ is active in the sense that the other components are fixed at their MLE's. By the definition of r_{ij}, we have

$$\begin{aligned} r_{ij} &= \frac{\partial M_j(\theta^*)}{\partial \theta_i} \\ &= \lim_{\theta_i \to \theta_i^*} \frac{M_j(\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i, \theta_{i+1}^*, \ldots, \theta_d^*) - M_j(\theta^*)}{\theta_i - \theta_i^*} \\ &= \lim_{t\to\infty} \frac{M_j(\theta^{(t)}(i)) - \theta_j^*}{\theta_i^{(t)} - \theta_i^*} \equiv \lim_{t\to\infty} r_{ij}^{(t)}. \end{aligned} \tag{3.3.2}$$

Because $M(\theta)$ is implicitly defined by the output of E and M steps, all quantities in (3.3.2) can be obtained using only the code for EM. This motivates the following algorithm for computing $r_{ij}^{(t)}$ ($t = 1, \ldots$).

First obtain θ^* by EM or any other procedure, such as closed-form answers obtained by factoring the likelihood with special patterns of missing data (Little and Rubin 1987, chap. 6). Then run a sequence of SEM iterations from some starting point not equal to θ^* in any component. At iteration $(t + 1)$ of SEM, perform the following steps—the first step can be eliminated if the starting value for SEM is the same as one of the parameter inputs to the original EM (i.e., some $\theta^{(t)}$) and the subsequent sequence of original EM iterates has been saved:

INPUT: θ^* and $\theta^{(t)}$.

Step 1. Run the usual E and M steps to obtain $\theta^{(t+1)}$.

Repeat steps 2–3 for $i = 1, \ldots, d$.

Step 2. Calculate $\theta^{(t)}(i)$ from (3.3.1), and treating it as the current estimate of θ, run one iteration of EM to obtain $\tilde{\theta}^{(t+1)}(i)$.

Step 3. Obtain the ratio

$$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \theta_j^*}{\theta_i^{(t)} - \theta_i^*}, \qquad \text{for } j = 1, \ldots, d. \tag{3.3.3}$$

OUTPUT: $\theta^{(t+1)}$ and $\{r_{ij}^{(t)}, i, j = 1, \ldots, d\}$.

We obtain r_{ij} when the sequence $r_{ij}^{(t^*)}, r_{ij}^{(t^*+1)}, \ldots$ is stable for some t^*. This process may result in using different values of t^* for different r_{ij} elements, as illustrated in Section 5. When all elements in the ith row of DM have been obtained, there is no need to repeat steps 2 and 3 for that i

in subsequent iterations. Notice that this method works even if DM is a deficient rank matrix, but is not defined if any denominators in (3.3.3) are zero; this can happen but is easily addressed, as discussed in Section 3.4.

This procedure for calculating r_{ij} from (3.3.3) is essentially numerically differentiating the vector function $M(\theta)$, and there are other ways for doing this. For example, once θ^* is obtained, one can simply perturb it by d linearly independent vectors ϵ_i $(i = 1, \ldots, d)$, and then solve a set of linear equations $M(\theta^* + \epsilon_i/2) - M(\theta^* - \epsilon_i/2) = C\epsilon_i$, $i = 1, \ldots, d$. The solution C provides an approximation to DM when all $\|\epsilon_i\|$ are small, but the accuracy of this approximation typically depends on sensible choice of ϵ_i $(i = 1, \ldots, d)$, which may not be an easy task in practice because the magnitude of different elements of DM may vary substantially. Just as with the scalar case, one can also approximate DM without first obtaining θ^* (Dennis and Schnabel 1983; Lanskey and Casella 1990). The basic idea is to approximate r_{ij} at iteration t by

$$\tilde{r}_{ij}^{(t)} = \frac{M_j(\theta_1^{(t-1)}, \ldots, \theta_{i-1}^{(t-1)}, \theta_i^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) - \theta_j^{(t)}}{\theta_i^{(t)} - \theta_i^{(t-1)}} \tag{3.3.4}$$

rather than by $r_{ij}^{(t)}$ of (3.3.3). Unlike in the scalar case, however, when $d > 1$, computing r_{ij} without first obtaining θ^* may actually require more computational time. The reason is that the number of E and M steps required for evaluating (3.3.4) is identical to that for (3.3.3) once θ^* has been obtained, and the extra E and M steps needed to obtain θ^* first can be easily compensated for by starting the evaluation of (3.3.3) from an initial value that is closer to θ^* than the original starting value of EM (more details are given in Section 5). The quantity obtained from (3.3.4) can also be quite unstable when different components of $\theta^{(t)}$ have different rates of convergence (see our example in Section 4.3). For example, in (3.3.4), if $\theta_i^{(t-1)}$ is quite close to θ_i^*, but $\theta_l^{(t-1)}$ for some $l \neq i$ is still relatively far from θ_l^*, then there might be no suitable t^* such that $\tilde{r}_{ij}^{(t^*)}$ is a good approximation to r_{ij}. Nevertheless, whether or not the procedure defined by (3.3.3) is the best way to approximate DM in general deserves further investigation.

3.4 Extensions When Some Components Have No Missing Information

When there is no missing information on a particular component of θ, EM will converge in one step for that component from any starting value, with the result that the corresponding component of $M(\theta)$ will be a constant with zero derivative. In this case with $\{M_j(\theta), j = 1, \ldots, d_1\}$ constant, the method described in Section 3.3 cannot be applied to compute $r_{ij}^{(t)}$ for $i = 1, \ldots, d_1$, because the corresponding denominators in (3.3.3) are zero. But then, because $r_{ij} = 0$ for $j = 1, \ldots, d_1$ and $i = 1, \ldots, d$, we can write

$$DM = \begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} 0 & A \\ 0 & DM^* \end{pmatrix} \end{array}, \qquad d_1 + d_2 = d. \tag{3.4.1}$$

The method described in Section 3.3 can then be directly applied to compute the $d_2 \times d_2$ submatrix DM^*, and the following identity, proved in the Appendix, shows that DM^* is sufficient for obtaining V.

Letting

$$I_{oc}^{-1} = \begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} G_1 & G_2 \\ G_2' & G_3 \end{pmatrix} \end{array}, \tag{3.4.2}$$

we have

$$V = I_{oc}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \Delta V^* \end{pmatrix} = \begin{pmatrix} G_1 & G_2 \\ G_2' & G_3 + \Delta V^* \end{pmatrix}, \tag{3.4.3}$$

where

$$\Delta V^* = (G_3 - G_2'G_1^{-1}G_2)DM^*(I - DM^*)^{-1}, \tag{3.4.4}$$

and I in (3.4.4) is the $d_2 \times d_2$ identity matrix.

Expressions (3.4.3) and (3.4.4) are special cases of (2.3.5) and (2.3.6), and the intuition underlying them is quite clear. Write $\theta = (\vartheta_1, \vartheta_2)$, where $\vartheta_1 = (\theta_1, \ldots, \theta_{d_1})$ and $\vartheta_2 = (\theta_{d_1+1}, \ldots, \theta_d)$. Consider first the special case when $G_2 = 0$. In this case, with complete data, $\vartheta_1 - \vartheta_1^*$ and $\vartheta_2 - \vartheta_2^*$ are (asymptotically) independent, and, therefore, the missing information for ϑ_2 does not increase the variance of $\vartheta_1 - \vartheta_1^*$. Thus we only need to use DM^*, the submatrix of DM corresponding to ϑ_2, to inflate $G_3 = V_c(\vartheta_2^*)$, the complete-data asymptotic variance of $\vartheta_2 - \vartheta_2^*$. This gives the formulas (3.4.3) and (3.4.4) when $G_2 = 0$.

More generally, we can decompose $\vartheta_2 - \vartheta_2^*$ into two (asymptotically with complete data) independent terms,

$$\vartheta_2 - \vartheta_2^* = [(\vartheta_2 - \vartheta_2^*) - E(\vartheta_2 - \vartheta_2^* \mid \vartheta_1 - \vartheta_1^*)] + E(\vartheta_2 - \vartheta_2^* \mid \vartheta_1 - \vartheta_1^*),$$

where the second term on the right side is purely a function of $\vartheta_1 - \vartheta_1^*$, with no increase in variance due to missing information (its complete-data variance is $G_2'G_1^{-1}G_2$). This suggests that we only need to use DM^* to inflate the complete-data variance of the first term, which is equal to $V_c(\vartheta_2^* \mid \vartheta_1^*) = G_3 - G_2'G_1^{-1}G_2$, the complete-data asymptotic conditional variance of $\vartheta_2 - \vartheta_2^*$ given $\vartheta_1 - \vartheta_1^*$. Hence the variance of $\vartheta_2 - \vartheta_2^*$ is

$$(G_3 - G_2'G_1^{-1}G_2)(I - DM^*)^{-1} + G_2'G_1^{-1}G_2 = G_3 + \Delta V^*,$$

which is exactly the term appearing at the lower right corner of (3.4.3). The other elements of I_{oc}^{-1} remain unchanged because there is no missing information for ϑ_1.

As illustrated in Section 4.4, the task of identifying the components of θ with no missing information is often trivial from inspection. Otherwise these components can be identified from the SEM iterations themselves, at least whenever $\theta_i^{(0)} \neq \theta_i^*$ for all i $(i = 1, \ldots, d)$.

3.5 Evaluation of $V = I_{oc}^{-1} + \Delta V$

Having obtained I_{oc}^{-1} and DM, we apply (2.3.6) to calculate ΔV, and then add it to I_{oc}^{-1} to produce the desired variance–covariance matrix V. This matrix, however, is not numerically constrained to be symmetric even though it is mathematically symmetric. The asymmetry can arise because of numerical inaccuracies in computing either DM or I_{oc}^{-1}. For exponential families, I_{oc} is typically very accurately

computed, whereas for nonexponential families its accuracy typically depends on large-sample approximations based on linearization methods. In contrast, the accuracy of DM is determined by the accuracy of EM itself, which typically is excellent when both E and M steps are simple calculations and is adequate in most cases as suggested by experience. If the resulting V is quite asymmetric, which we have only observed in the presence of programming errors, it is an indication of either such programming errors or severe numerical imprecision in either DM or I_{oc}^{-1}. This feature of SEM is analogous to the feature of EM that each iteration must increase the observed-data likelihood, in the sense that a decrease in the likelihood function indicates either programming errors or severe numerical imprecision. We consider it a highly desirable property of SEM, which is not shared by other commonly used algorithms (e.g., quasi-Newton–Raphson) for obtaining the asymptotic variance–covariance matrix. More on this as a diagnostic for SEM is presented in Section 5.2.

In Section 5 we also discuss how to proceed if $I - DM$ cannot be inverted numerically because it is nearly singular, which can happen when EM is extremely slow to converge. As is shown there, in general, one can always apply SEM to compute the observed-data observed information matrix $I_o(\theta^* \mid Y_{obs})$. As suggested by our experience, if the resulting matrix is not (numerically) symmetric, then it is an indication of programming errors in either the EM code or the SEM code. If this matrix is symmetric but not positive semidefinite, then it indicates that EM has not converged to a (local) maximum but rather to a saddle point, and one should rerun EM starting near the last iterate but perturbed in the direction of the eigenvector corresponding to the most negative eigenvalue of $I_o(\theta^* \mid Y_{obs})$. In this sense, SEM can also be used to monitor the convergence of EM to a local maximum, which cannot be detected by monitoring the increase in the likelihood. In fact, as we have encountered in practice, monotone increases in the likelihood may fail to detect programming errors in EM that leave "EM" converging to a nonstationary point.

4. EXAMPLES OF SEM

4.1 Introduction to Examples

As we mentioned in Section 1.3, using a single parameter example from DLR, Smith (1977) illustrated using the rate of convergence of EM to obtain the asymtotic standard error of the MLE. Since this example is very easy to understand and gives the flavor of SEM, we briefly review it in Section 4.2 before giving two examples of multiparameter cases.

In Section 4.3 we present a univariate contaminated normal example, which has a special feature in that the large-sample component-wise rates of convergence of EM correspond to different eigenvalues of DM in contrast to the usual situation where they are all equal to the largest eigenvalue of DM (Meng 1990). Despite this peculiarity in the convergence of EM, SEM remains quite stable.

In Section 4.4 we use a bivariate normal data set with the first variable fully observed to demonstrate the use of formulas (3.4.3) and (3.4.4) to deal with the situation where only a subset of the parameters are affected by missing information.

As we will see in all these examples, SEM is easy to implement and the results obtained from it are quite satisfactory. Incidentally, Meilijson's (1989) technique cannot be applied to the examples of Sections 4.2 or 4.4 because the observed data do not have iid structures even though the complete data do.

4.2 A Multinomial Example

This example has frequently appeared in the literature of the EM algorithm since it was first used in DLR to introduce EM (e.g., Louis 1982; Little and Rubin 1987). Suppose the complete data $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ have a multinomial distribution with cell probabilities

$$(1/2, \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4), \qquad 0 \le \theta \le 1.$$

The objective here is to find the ML estimate for θ based on the observed counts $Y_{obs} = (Y_1 + Y_2, Y_3, Y_4, Y_5) = (125, 18, 20, 34)$. Notice that if Y were observed, the MLE of θ would be immediate:

$$\theta^* = (Y_2 + Y_5)/(Y_2 + Y_3 + Y_4 + Y_5). \qquad (4.2.1)$$

Also note that the log-likelihood $L(\theta \mid Y)$ is linear in Y, so in the E step we only need to replace the missing values by their corresponding conditional expectations. Thus at the tth iteration, we have for the E step

$$Y_2^{(t)} = 125 \frac{\theta^{(t)}/4}{1/2 + \theta^{(t)}/4}; \qquad (4.2.2)$$

and for the M step, from (4.2.1), we have

$$\theta^{(t+1)} = (Y_2^{(t)} + 34)/(Y_2^{(t)} + 72). \qquad (4.2.3)$$

Formulas (4.2.2) and (4.2.3) together define the mapping $\theta^{(t+1)} = M(\theta^{(t)})$, the EM algorithm. In fact, by substituting $Y_2^{(t)}$ from (4.2.2) into (4.2.3), and letting $\theta^{(t+1)} = \theta^{(t)} = \theta^*$, we can explicitly solve a quadratic equation for the MLE of θ:

$$\theta^* = (15 + \sqrt{53{,}809})/394 \cong .6268214980.$$

Table 1 displays the convergence to this solution from the initial value $\theta^{(0)} = .5$. The second column in the table gives the corresponding values of $\phi = \arcsin \sqrt{\theta}$, which is the well known variance-stabilizing transformation for the binomial proportion θ. The third and fourth columns give their corresponding deviations $d_\theta^{(t)} = \theta^{(t)} - \theta^*$ and $d_\phi^{(t)} = \phi^{(t)} - \phi^*$. The fifth and sixth columns are the corresponding ratios of successive deviations. The ratios are essentially constant for $t \ge 3$, which implies that the rate of convergence for EM is $r = .1328$. This rate of convergence is invariant under any one-to-one differentiable transformation of the parameter.

Since the complete-data density is

$$f(Y \mid \theta) \propto \theta^{Y_2+Y_5}(1 - \theta)^{Y_3+Y_4},$$

the complete-data variance for $\theta - \theta^*$ is simply the ordinary binomial variance $\theta(1 - \theta)/n$, where $n = Y_2 + Y_3 +$

Table 1. The EM Iterations for the Example in Section 4.2

t	$\theta^{(t)}$	$\phi^{(t)}$	$d_\theta^{(t)}$	$d_\phi^{(t)}$	$d_\theta^{(t+1)}/d_\theta^{(t)}$	$d_\phi^{(t+1)}/d_\phi^{(t)}$
0	.50000000	.78539816	−.12682150	−.12822228	.1465	.1490
1	.60824742	.89450953	−.01857408	−.01911092	.1346	.1352
2	.62432105	.91103720	−.00250045	−.00258324	.1330	.1331
3	.62648888	.91327661	−.00033262	−.00034383	.1328	.1328
4	.62677732	.91357478	−.00004418	−.00004567	.1328	.1328
5	.62681563	.91361438	−.00000587	−.00000606	.1328	.1328
6	.62682072	.91361964	−.00000078	−.00000081	.1328	.1328
7	.62682140	.91362034	−.00000010	−.00000011	·	·
8	.62682149	.91362043	−.00000001	−.00000001	·	·
9	.62682150	.91362044	−.00000000	−.00000000	·	·

$Y_4 + Y_5$. Thus, from (1.3.1), as given in Smith (1977), the asymptotic standard error of $\theta - \theta^*$ based on the observed counts Y_{obs} is

$$\left[\frac{\theta^*(1-\theta^*)}{n^*(1-r)}\right]^{1/2} \cong .051,$$

where $n^* = Y_2^* + 72 = 101.83$ is the expectation of n given the MLE of θ. Similarly, since the complete-data variance of $\phi - \phi^*$ is $1/(4n)$, the asymptotic standard error of $\phi - \phi^*$ based on Y_{obs} is

$$\left[\frac{1}{4n^*(1-r)}\right]^{1/2} \cong .053.$$

These numerical results can be verified by direct computation (for example, see Little and Rubin 1987, p. 138).

4.3 A Univariate Contaminated Normal Example

This example is used in Little and Rubin (1987) to illustrate the idea of treating mixture models as incomplete-data problems. Suppose $x_1, \ldots, x_n$ represent an independent sample from the univariate contaminated normal model

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

where $0 < \pi < 1$ and $\lambda > 0$ are both known. The objective here is to find the MLE $\theta^* = (\mu^*, \log \sigma^{2*})$ for $\theta = (\mu, \log \sigma^2)$, which is, as mentioned in Section 1.1, an appropriate parameterization for normal theory inference and for SEM computations.

This problem can be treated as an incomplete-data problem although there are no missing data in the usual sense. Let

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

then $X = (x_1, \ldots, x_n)$ can be considered as an independent sample from a population such that

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

where the q_i are unobserved iid random variables with known density $h(q_i)$. In this setting we can apply the EM algorithm to compute θ^*, treating $Q = (q_1, \ldots, q_n)$ as the missing data Y_{mis}, X as the observed data Y_{obs}, and (X, Q) as the complete data Y.

Implementing EM is quite straightforward. Since the complete-data log-likelihood is linear in the q_i, the E step reduces to finding the conditional expectation

$$w_i^{(t)} \equiv w_i(\theta^{(t)}) = E(q_i \mid x_i, \theta^{(t)}), \tag{4.3.1}$$

where $\theta^{(t)}$ is the current estimate of θ. A simple application of Bayes's theorem yields

$$\begin{aligned} w_i(\theta) &\equiv E(q_i \mid x_i, \theta) \\ &= \frac{1-\pi+\pi\lambda^{3/2}\exp\{(1-\lambda)z_i^2/2\}}{1-\pi+\pi\lambda^{1/2}\exp\{(1-\lambda)z_i^2/2\}}, \end{aligned} \tag{4.3.2}$$

where $z_i^2 = (x_i - \mu)^2/\sigma^2$. The M step is also trivial since the complete-data MLE of $\theta = (\mu, \log \sigma^2)$ is in closed form:

$$\mu^* = \sum_{i=1}^n q_i x_i \Big/ \sum_{i=1}^n q_i, \tag{4.3.3}$$

and

$$\log \sigma^{2*} = \log\left\{\frac{1}{n}\sum_{i=1}^n q_i(x_i - \mu^*)^2\right\}. \tag{4.3.4}$$

Therefore the new estimate, $\theta^{(t+1)} = (\mu^{(t+1)}, \log \sigma^{2(t+1)})$, can be easily obtained from (4.3.3) and (4.3.4) with q_i replaced by $w_i^{(t)}$ from (4.3.1) and (4.3.2).

To illustrate SEM, we performed a simulation with $\mu = 0$, $\sigma^2 = 1$, $\lambda = 2$, $\pi = .1$, and sample size 100. Table 2

Table 2. The EM Iterations for the Example in Section 4.3

t	$\mu^{(t)}$	$d_1^{(t)}$	$d_1^{(t+1)}/d_1^{(t)}$	$\log \sigma^{2(t)}$	$d_2^{(t)}$	$d_2^{(t+1)}/d_2^{(t)}$
0	.19866916	−.00061984	.05187	.22943020	.00923662	.03502
1	.19925685	−.00003215	.04890	.22051706	.00032348	.03496
2	.19928743	−.00000157	.04708	.22020489	.00001131	.03497
3	.19928893	−.00000007	.04590	.22019397	.00000040	.03498
4	.19928900	−.00000000	.04509	.22019359	.00000001	.03499
5	.19928900	−.00000000	.04443	.22019358	.00000000	.03497
6	.19928900	−.00000000	.04223	.22019358	.00000000	.03386

Table 3. SEM Iterations for DM for the Example in Section 4.3

t	$r_{11}^{(t)}$	$r_{12}^{(t)}$	$r_{21}^{(t)}$	$r_{22}^{(t)}$
0	.04251717	−.00112649	−.00063697	.03494620
1	.04251677	−.00112375	−.00063697	.03492154
2	.04251674	−.00112359	−.00063666	.03492066
3	.04251658	−.00112333	−.00063663	.03492059
True	.04251675	−.00112360	−.00063666	.03492063

gives the EM output with initial values $\mu^{(0)} = \bar{x}$ and $\sigma^{2(0)} = s^2/(1 - \pi + \pi/\lambda)$ (which is an unbiased estimate of σ^2), where $d_1^{(t)} = \mu^{(t)} - \mu^*$ and $d_2^{(t)} = \log \sigma^{2(t)} - \log \sigma^{2*}$. The first four rows of Table 2 give the corresponding output for $r_{ij}^{(t)}$ $(i, j = 1, 2)$, for $t = 0, \ldots, 3$, obtained by SEM using $\theta^* = \theta^{(6)}$ and starting from $\theta^{(0)}$; the last row gives the true values of r_{ij} $(i, j = 1, 2)$ obtained by direct computation using analytic expressions (Meng 1990).

As we can see from Table 3, using only six iterations initially to estimate θ^*, we can approximate r_{ij} very well by $r_{ij}^{(t)}$ for "moderate" t. More precise values for r_{ij} can be obtained by using more initial iterations of EM to obtain a more precise value for θ^*. From Table 3, we can take

$$DM = \begin{pmatrix} .04252 & -.00112 \\ -.00064 & .03492 \end{pmatrix}. \qquad (4.3.5)$$

The complete-data variance–covariance matrix I_{oc}^{-1} is readily available from standard calculations for the bivariate normal distribution [notice that the complete-data density $f(x, q \mid \theta)$ is from an irregular exponential family]. In particular, using (2.3.4), we have

$$I_{oc}^{-1} = \frac{1}{n}\begin{pmatrix} \sigma^{2*}/\overline{w}^* & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} .01133 & 0 \\ 0 & .02 \end{pmatrix}, \qquad (4.3.6)$$

where $n\overline{w}^* = \Sigma_i w_i(\theta^*) = 109.97696$ is obtained at the last E step. Thus, from (4.3.5) and (4.3.6), using (2.4.6) [or (2.3.5) and (2.3.6)], we have

$$V = \begin{pmatrix} .01184 & -.00001 \\ -.00001 & .02072 \end{pmatrix}. \qquad (4.3.7)$$

The symmetry of the resulting matrix indicates numerical accuracy, as discussed in Section 3.5.

It is seen from the third and sixth columns of Table 2 that the two components of θ converge at different rates, corresponding to two different eigenvalues of DM. This special feature occurs because the two components of θ are asymptotically independent with both complete data and observed data.

The above simulation has been repeated with many different choices of λ, π, and sample size n, including some extreme cases (e.g., $\lambda = .01$, $\pi = .5$). In all these cases, SEM obtains quite stable and accurate values for DM, and hence for V, since the computation of I_{oc}^{-1} using (4.3.6) is very accurate.

4.4 A Bivariate Normal Example

In Section 3.4 we mentioned that in some cases there is no missing information for some particular components, and then we need to identify these components so that we can inflate the appropriate submatrix of the complete-data variance–covariance matrix. The following example, which is used in Little and Rubin (1987, pp. 101–106) for the purpose of comparing different interval estimates, has this feature.

Table 4. Data for the Example in Section 4.4

Y_1	8	6	11	22	14	17	18	24	19	23	26	40	4	4	5	6	8	10
Y_2	59	58	56	53	50	45	43	42	39	38	30	27	—	—	—	—	—	—

The data given in Table 4 are assumed to follow a bivariate normal distribution with parameter $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, where ρ is the correlation coefficient. As is well known, a normalizing parameterization in this case is $\theta = (\mu_1, \mu_2, \log \sigma_1^2, \log \sigma_2^2, Z_\rho)$, where $Z_\rho = .5 \log\{(1 + \rho)/(1 - \rho)\}$ is the Fisher Z tranformation of ρ. Since the first variable is fully observed, the MLE's for μ_1 and $\log \sigma_1^2$ are simply the sample mean and the log of the sample variance (with divisor n) of the first variable, respectively. Thus EM will converge in one step for these two components from any starting values, with the result that the corresponding components of $M(\theta)$ are constant functions. The implementation of EM for the normal distribution has been fully described in the literature, for instance, using the SWEEP operator in Little and Rubin (1987, chap. 8).

The first row of Table 5 gives the MLE for $\theta_2 = (\mu_2, \log \sigma_2^2, Z_\rho)$, using $\theta_2^* = \theta_2^{(65)}$ (In this case, the closed-form value of θ_2^* can be obtained by factoring the likelihood; see, for example, Little and Rubin 1987, pp. 98–100). The second row gives asymptotic standard errors for $\theta_2 - \theta_2^*$, obtained by direct computation in Little and Rubin (1987, p. 106) (and transformed via the appropriate Jacobian), and the third row gives the corresponding standard errors obtained by SEM.

The SEM results are obtained as follows, using the method of Section 3.4. First, using the algorithm in Section 3.3, we obtain DM^*, the submatrix of DM corresponding to $\theta_2 = (\mu_2, \log \sigma_2^2, Z_\rho)$

$$DM^* = \begin{array}{c} \\ \mu_2 \\ \log \sigma_2^2 \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccc} \mu_2 & \log \sigma_2^2 & Z_\rho \end{array} \\ \begin{pmatrix} .33333 & .05037 & -.02814 \\ 1.44444 & .29894 & .01921 \\ -.64222 & .01529 & .32479 \end{pmatrix} \end{array}. \qquad (4.4.1)$$

Since the complete-data distribution is from a regular exponential family (the standard bivariate normal), by (3.2.2), to obtain I_{oc}^{-1} we only need to compute the inverse of the complete-data Fisher information matrix $I^{-1}(\theta^*)$. It is particularly easy to do this for the bivariate normal:

Table 5. MLE's for θ_2 and Their Asymptotic Standard Errors (S.E.)

	μ_2	$\log \sigma_2^2$	Z_ρ
MLE $(\theta_2^{(65)})$	49.33	4.74	−1.45
S.E. from Little and Rubin (1987)	2.73	.37	.274
S.E. from SEM calculations	2.73	.37	.274

$$I_{oc}^{-1} = I^{-1}(\theta^*)$$

$$= \begin{array}{c} \\ \mu_1 \\ \mu_2 \\ \log \sigma_1^2 \\ \log \sigma_2^2 \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccccc} \mu_1 & \mu_2 & \log \sigma_1^2 & \log \sigma_2^2 & Z_\rho \end{array} \\ \begin{pmatrix} 4.9741 & -5.0387 & 0 & 0 & 0 \\ -5.0387 & 6.3719 & 0 & 0 & 0 \\ 0 & 0 & .1111 & .0890 & -.0497 \\ 0 & 0 & .0890 & .1111 & -.0497 \\ 0 & 0 & -.0497 & -.0497 & .0556 \end{pmatrix} \end{array}. \quad (4.4.2)$$

After a rearrangement that makes the first two rows and columns correspond to the parameters for the first component for which there is no missing information, the right side of (4.4.2) becomes

$$\begin{array}{c} \\ \mu_1 \\ \log \sigma_1^2 \\ \mu_2 \\ \log \sigma_2^2 \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccccc} \mu_1 & \log \sigma_1^2 & \mu_2 & \log \sigma_2^2 & Z_\rho \end{array} \\ \left(\begin{array}{ccccc} 4.9741 & 0 & -5.0387 & 0 & 0 \\ 0 & .1111 & 0 & .0890 & -.0497 \\ -5.0387 & 0 & 6.3719 & 0 & 0 \\ 0 & .0890 & 0 & .1111 & -.0497 \\ 0 & -.0497 & 0 & -.0497 & .0556 \end{array}\right) \end{array}$$

$$\equiv \begin{pmatrix} G_1 & G_2 \\ G_2' & G_3 \end{pmatrix}, \quad (4.4.3)$$

using the notation of (3.4.2). Applying formula (3.4.4), we obtain

$$\Delta V^* = \begin{array}{c} \\ \mu_2 \\ \log \sigma_2^2 \\ Z_\rho \end{array} \begin{array}{c} \begin{array}{ccc} \mu_2 & \log \sigma_2^2 & Z_\rho \end{array} \\ \begin{pmatrix} 1.0858 & .1671 & -.0933 \\ .1671 & .0286 & -.0098 \\ -.0933 & -.0098 & .0194 \end{pmatrix} \end{array}, \quad (4.4.4)$$

which is the increase in the variance of $\theta_2 - \theta_2^*$ due to missing information.

To obtain the asymptotic variance–covariance matrix for $\theta_2 - \theta_2^*$, we only need to add ΔV^* to G_3 of (4.4.3). For example, for the standard error of $\mu_2 - \mu_2^*$, we have from (4.4.3) and (4.4.4)

$$(6.3719 + 1.0858)^{1/2} \cong 2.73,$$

as given in the third row of Table 5.

5. REMARKS ON THE IMPLEMENTATION OF SEM

5.1 Starting Values and Stopping Criterion

The choice of starting values for SEM is based on considerations of both numerical accuracy and computational cost. As for numerical accuracy, it is almost always safe to use the original EM initial values as the starting values for SEM computations for DM. This choice does not require any additional work but may result in some unnecessary iterations because the original starting values may be far from the MLE. Based on our limited experience, we suggest using either a suitable (e.g., the second) iterate of the original EM or two complete-data standard deviations from the MLE, where the complete-data standard deviations are simply the square roots of the diagonal elements of I_{oc}^{-1}. Of course, sophisticated users of our method may be able to choose other starting values that preserve numerical accuracy but reduce the number of iterations.

The stopping criterion for SEM should be less stringent than that for the original EM because the method for computing DM is essentially numerical differentiation of a function, which is known to be less accurate than evaluating the function itself. We use the square root of the stopping criterion of the original EM as the stopping criterion for SEM (e.g., if the stopping criterion for EM is 10^{-8}, then the stopping criterion for SEM is 10^{-4}). Such a stopping criterion typically will stop the iterations at different places for different elements of DM, as illustrated in the next section.

5.2 Diagnostics

The symmetry of the resulting variance–covariance matrix V and the numbers of iterations needed for SEM to converge for different elements of DM provide very informative diagnostics for programming errors and numerical precision.

As an illustration, Table 6 gives some details in implementing SEM for the example in Section 4.4. The first column gives the stopping criteria used for EM and SEM. The second column gives the corresponding total number of iterations for EM, and the next three columns give a 3×3 SEM index matrix, where each element is the total number of iterations needed for SEM to converge for the corresponding element of DM^* using the original EM starting value as the initial value for SEM. The three columns under the heading DM^* and ΔV^* give the corresponding convergent values of DM^* and ΔV^*.

It is evident from the table that the more stringent the criterion we use, the more symmetry we have in the final

Table 6. Numbers of EM and SEM Iterations and Convergent Values of DM and ΔV* Under Three Different Criteria in the Example in Section 4.4*

Stopping criterion EM; SEM	Number of EM iterations for θ^*	Number of SEM iterations for DM^*			DM^*			ΔV^*		
10^{-4}; 10^{-2}	27	2	2	2	.333332	.048028	−.036111	1.080332	.154940	−.125674
		3	2	2	1.392507	.285944	.007518	.160118	.026170	−.012917
		8	2	6	−.651442	.017030	.334558	−.092877	−.008900	.021572
10^{-8}; 10^{-4}	46	2	8	11	.333333	.050273	−.028245	1.085304	.166662	−.093749
		17	14	13	1.444338	.298837	.019051	.167033	.028514	−.009822
		16	9	14	−.642371	.015366	.324922	−.093329	−.009754	.019380
10^{-12}; 10^{-6}	65	2	17	20	.333333	.050374	−.028144	1.085838	.167083	−.093350
		26	23	23	1.444443	.298944	.019210	.167087	.028555	−.009778
		26	17	23	−.642220	.015291	.324793	−.093344	−.009777	.019352

matrix ΔV^*, the increase in variance–covariance due to missing information. Thus, from the symmetry of the resulting variance–covariance matrix, we can basically deduce how many digits in the final result are trustworthy. If the resulting variance–covariance matrix is quite asymmetric, one should first increase the accuracy of the original EM (that is, use a more stringent stopping criterion) to see if the symmetry is improved. If it is not improved, our experience suggests the existence of programming errors in either the EM code or the SEM code.

The fact that SEM converges at different steps for different elements of DM suggests that some standard computational methods (e.g., quasi-Newton–Raphson) that provide the entire symmetric variance–covariance matrix at the same iteration might not provide as accurate results. The SEM index matrix tells us when SEM has converged for each element, and so if one observes that some elements in DM require the total number of iterations, the indication is that these elements have not converged, and one should use more stringent stopping criteria for EM and SEM.

5.3 Stability

As we mentioned earlier, the numerical differentiation method described in Section 3.3 is not used to obtain the variance–covariance matrix, but rather the increases in variance–covariance due to missing data that are to be added to the complete-data variance–covariance matrix. As a result, SEM typically is stable for the following reasons. When some increase in variance is large, the convergence of EM is slow because of the large fraction of missing information. This slow convergence of EM, however, provides an excellent sequence of iterates from which its linear convergence rate can be recovered, and thus the results obtained from SEM are typically quite accurate. In contrast, when the increases in variance are relatively small, the complete-data variance–covariance matrix, which is usually very accurately calculated, dominates the increases due to missing data. Therefore, the resulting matrix, as the sum of this accurately obtained complete-data variance–covariance matrix and the matrix of small increases due to missing data, is still quite satisfactory even if the numerical differentiation part of SEM used for calculating the increases is not as accurate as it is when the convergence of EM is slow. Thus, SEM is typically more stable than pure numerical differentiation procedures because of this automatic "self-adjustment."

As we mentioned in Section 3.3, there are other ways to obtain DM using other sequences of values for numerical differentiation, and they could be used in place of our method. A nice feature of our method is that it, like EM, is easy to implement and works reliably without extra attention because the EM iterates automatically choose "appropriate" step sizes for numerically differentiating the vector function $M(\theta)$. Just as a sophisticated user can develop a special purpose algorithm superior to EM in special cases, a sophisticated user can write a special purpose numerical differentiation program that is superior to our method in special cases. SEM is a general algorithm for computing asymptotic variance–covariance matrices using EM, which preserves the simplicity and stability of the EM approach.

5.4 Computational Effort and Storage Requirements

The total amount of computation involved in SEM with $d > 1$ is greater than that for the corresponding EM. Assuming the square root stopping rule will be used with the SEM computations for DM, our experience is that the maximum number of iterations of SEM for each row of DM is less than one-half the number of EM iterations; for example, see Table 6. Each of these SEM iterations is approximately $d + 1$ times as computationally expensive as an EM iteration. (In addition, SEM requires a $d \times d$ matrix inversion and a $d \times d$ matrix multiplication to compute V, which will be ignored in the following discussion since the iterative part of SEM is usually the dominant expense.) Consequently, SEM requires roughly $(d + 1)/2$ times as much computational time as EM itself. This factor, however, applies only to standard computing environments. In parallel computing environments, SEM can be faster than the corresponding EM because each of the d rows of DM can be evaluated independently by d parallel processors assuming access to the current EM iterate $\theta^{(t)}$, available either by saving initial EM iterates or by distributing $\theta^{(t)}$ to each of the d parallel processors from another processor running EM in parallel. Thus, if the computer being used has $d + 1$ parallel processors, then SEM is roughly twice as fast as the corresponding EM under the square root stopping rule. In this sense, SEM is ideally suited for modern parallel computing environments because the bulk of the computations can be done in parallel and then combined to produce the desired answer.

Another concern in practice is storage requirements, again a problem of decreasing importance in many modern computing environments. The matrices involved (V, DM, I_{oc}^{-1}) are $d \times d$, and for large d are large. But this is a feature of the problem if all of V is desired; it is not a problem created by SEM. If only a subset of values of V are desired, a version of SEM that involves only this subset would be attractive, but we have not been able to find a computationally effective version of SEM for this. Of course, if storage is no problem, some computational time [or parallel processor $(d + 1)$] can be cut when computing DM by saving the original sequence of EM iterates.

5.5 How To Proceed When $I - DM$ Is Nearly Singular

As shown in Section 2.3, the calculation of ΔV requires the inversion of $I - DM$, which can be nearly singular when the convergence of EM is extremely slow (i.e., when the largest eigenvalue of DM is very close to 1). Statistically this implies that the observed-data likelihood function is flat along some directions and thus that the observed-data observed information matrix $I_o(\theta^* \mid Y_{obs})$ [defined by (2.3.2)] is nearly singular. As with the storage requirements for V, this issue is a feature of the likelihood function and data, and is not a problem created by SEM. In fact, even in these ill-conditioned cases, SEM can be very helpful for identi-

fying the directions with little information and for finding the asymptotic variance–covariance matrix for the linear combinations of $\theta - \theta^*$ about which the observed data do provide information.

More specifically, using (2.4.4) and (2.4.5), one can first apply SEM to obtain the observed-data observed information matrix, sometimes called the precision matrix $P = I_o(\theta^* \mid Y_{obs}) = (I - DM)I_{oc}$. As discussed in Section 3.5 and Section 5.2, any lack of (numerical) symmetry in P indicates lack of convergence or existence of programming errors. Assuming (numerical) symmetry of P, standard matrix operations can then be applied to find the spectral decomposition of P,

$$P = B \cdot \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) \cdot B',$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of P and the columns of B are the corresponding orthonormalized eigenvectors. If λ_d is identified as (numerically) negative, as we mentioned in Section 3.5, it indicates that EM has not converged to a (local) maximum but to a saddle point, and then EM should continue in the direction corresponding to λ_d. Otherwise, suppose the first m ($\leq d$) eigenvalues are positive (and the last $d - m$ eigenvalues are identified as numerically close to zero), and their corresponding eigenvectors form the $d \times m$ submatrix B_1 of B; that is, $B = (B_1, B_2)$. Then the $m \times m$ diagonal matrix $\text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_m^{-1})$ gives the asymptotic variance–covariance matrix for the linear combination $B_1'(\theta - \theta^*)$ about which the observed data do have information. The asymptotic variance–covariance matrix for the linear combination $B_2'(\theta - \theta^*)$ is infinity or nearly infinity because the observed data have little or no information about it; in other words, the observed-data likelihood function is flat along the directions determined by B_2.

APPENDIX

Proof of (3.4.3) and (3.4.4)

By the definition of ΔV in (2.3.6), from (3.4.1) and (3.4.2), we have

$$\begin{aligned}\Delta V &= I_{oc}^{-1}DM[I - DM]^{-1} \\ &= \begin{pmatrix} G_1 & G_2 \\ G_2' & G_3 \end{pmatrix}\begin{pmatrix} 0 & A \\ 0 & DM^* \end{pmatrix}\begin{pmatrix} I_{d_1} & A(I_{d_2} - DM^*)^{-1} \\ 0 & (I_{d_2} - DM^*)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} 0 & (G_1A + G_2DM^*)(I_{d_2} - DM^*)^{-1} \\ 0 & (G_2'A + G_3DM^*)(I_{d_2} - DM^*)^{-1} \end{pmatrix}. \qquad \text{(A.1)}\end{aligned}$$

But the upper right corner of the last matrix in (A.1) is zero, since ΔV is symmetric, which implies that

$$A = -G_1^{-1}G_2DM^*. \qquad \text{(A.2)}$$

Substituting (A.2) into (A.1) we obtain

$$\begin{aligned}\Delta V &= \begin{pmatrix} 0 & 0 \\ 0 & (G_3 - G_2'G_1^{-1}G_2)DM^*(I_{d_2} - DM^*)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & \Delta V^* \end{pmatrix}, \qquad \text{(A.3)}\end{aligned}$$

where ΔV^* is defined in (3.4.4). Thus (3.4.3) follows from (2.3.5) and (A.3).

[*Received January 1990. Revised March 1991.*]

REFERENCES

Carlin, J. B. (1987), "Seasonal Analysis of Economic Time Series," unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.

Celeux, G., and Diebolt, J. (1985), "The SEM Algorithm: A Probabilistic Teacher Derived From the EM Algorithm for the Mixture Problem," *Computational Statistics Quarterly*, 2, 73–82.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

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

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Goodman, L. A. (1974), "Exploratory Latent Structure Models Using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 315–331.

Lanskey, D., and Casella, G. (1990), "Improving the EM Algorithm," in *Computing Science and Statistics: Proceedings of the Twenty-Second Symposium on the Interface*, American Statistical Association.

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

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Meilijson, I. (1989), "A Fast Improvement to the EM Algorithm on its Own Terms," *Journal of the Royal Statistical Society*, Ser. B, 51, 127–138.

Meng, X. L. (1990), "Towards Complete Results for Some Incomplete-Data Problems," Ph.D. dissertation, Harvard University, Dept. of Statistics. Printed by U.M.I., Ann Arbor, MI.

Meng, X. L., and Rubin, D. B. (1989), "Obtaining Asymptotic Variance–Covariance Matrices by the EM Algorithm," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 41–45.

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.

Rubin, D. B. (1976), "Noniterative Least Squares Estimates, Standard Errors and F-Test for Any Analysis of Variance With Missing Data," *Journal of the Royal Statistical Society*, Ser. B, 38, 270–274.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Smith, C. A. B. (1977), discussion of "Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm," by A. P. Dempster, N. M. Laird, and D. B. Rubin, *Journal of the Royal Statistical Society*, Ser. B, 39, 24–25.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 805–811.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley.