The Restricted EM Algorithm for Maximum Likelihood Estimation Under Linear Restrictions on the Parameters

Author(s): Dong K. Kim and Jeremy M. G. Taylor

Source: *Journal of the American Statistical Association*, Vol. 90, No. 430 (Jun., 1995), pp. 708-716
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2291083
Accessed: 25/05/2011 17:26

The Restricted EM Algorithm for Maximum Likelihood Estimation Under Linear Restrictions on the Parameters

Dong K. KIM and Jeremy M. G. TAYLOR*

The EM algorithm is one of the most powerful algorithms for obtaining maximum likelihood estimates for many incomplete-data problems. But when the parameters must satisfy a set of linear restrictions, the EM algorithm may be too complicated to apply directly. In this article we propose maximum likelihood estimation procedures under a set of linear restrictions for situations in which the EM algorithm could be used if there were no such restrictions on the parameters. We develop a modification to the EM algorithm, which we call the restricted EM algorithm, incorporating the linear restrictions on the parameters. This algorithm is easily updated by using the code for the complete data information matrix and the code for the usual EM algorithm. Major applications of the restricted EM algorithm are to construct likelihood ratio tests and profile likelihood confidence intervals. We illustrate the procedure with two models: a variance component model and a bivariate normal model.

KEY WORDS: Bivariate normal model; Incomplete data problems; Lagrange multipliers; Likelihood ratio test; Profile likelihood confidence interval; Restricted maximum likelihood estimation; Variance component model.

1. INTRODUCTION

The EM algorithm is one of the most powerful algorithms for maximum likelihood estimation in incomplete data problems. Because the EM algorithm is computationally simple and numerically stable, it is used for a broad range of applications, such as variance component models in normal data, finite mixture models, and multivariate normal models with missing data (Dempster, Laird, and Rubin 1977; Little and Rubin 1987).

The EM algorithm handles complicated missing-data problems by using complete-data tools. It is particularly useful in many situations in which there are no actual missing data but the problem can be reformulated as a missing-data problem such that the EM algorithm can be used. In the EM algorithm it is usually necessary to find the conditional distribution in the E step, then use standard maximum likelihood estimation for the complete-data problem in the M step. When there are no restrictions on the parameters, each step of the EM algorithm is usually simple and straightforward. But when the parameters must satisfy a set of linear restrictions, the M step will usually involve complicated procedures to find the solution, and no closed form may exist. In this case, constrained maximization routines are needed.

For data without missing information, we can use constrained maximum likelihood estimation methods to find solutions under linear restrictions. For a computationally convenient method, Nyquist (1991) proposed iteratively reweighted least squares to estimate parameters under a set of linear restrictions and applied the method to generalized linear models.

Recently, Kim (1991) proposed a modification to the EM algorithm, called the restricted EM algorithm, that incorporates linear restrictions on the parameters. In this work Kim developed the restricted EM algorithm for the specific case of finite mixture models. He illustrated the numerical properties of the algorithm and applied it to an overdispersed binomial data set from radiobiology.

In this article we construct the restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. We show that this algorithm is applicable, not only to the finite mixture model, but also to general statistical problems in which the EM algorithm could be used. Furthermore, we apply the restricted EM algorithm to test hypotheses concerning linear combinations of parameters and to construct confidence intervals. Although there are several ways to construct hypothesis tests and confidence intervals from the EM algorithm (Louis 1982; Meng and Rubin 1991), these approaches produce Wald-type symmetric confidence intervals based on the asymptotic variance–covariance matrix. In this article we apply the restricted EM algorithm to produce the profile of the log-relative likelihood for a parameter. From this likelihood-based approach, we can construct a profile likelihood confidence interval that is not forced to be symmetric.

Section 2 describes maximum likelihood estimation of the observed data under linear restrictions on the parameters. Section 3 proposes the restricted EM algorithm for incomplete data problems when the parameters must satisfy linear restrictions. Section 4 describes some theoretical properties of the restricted EM algorithm; and Section 5 applies the restricted EM algorithm to obtain likelihood ratio tests and profile likelihood confidence intervals. Section 6 illustrates use of the restricted EM algorithm with a variance component model and a bivariate normal model. Section 7 contains a discussion and conclusions. Proofs can be found in the Appendix.

2. MAXIMUM LIKELIHOOD ESTIMATION UNDER LINEAR RESTRICTIONS ON PARAMETERS

In this section, we describe restricted maximum likelihood estimation assuming that the observation vector has no

* Dong K. Kim is Assistant Professor, Department of Mathematics, Statistics, and Computer Science, and Department of Surgical Oncology, University of Illinois, Chicago, IL 60607. Jeremy M. G. Taylor is Professor, Department of Biostatistics, University of California, Los Angeles, CA 90024. This work was partly supported by National Institutes of Health Grants CA 45216 and AI 07370. The authors thank the referees, the associate editor, and the editor for helpful comments and suggestions.

Journal of the American Statistical Association
June 1995, Vol. 90, No. 430, Theory and Methods

missing information. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be an observation vector and let $\boldsymbol{\theta}$ be a $p \times 1$ parameter vector of interest. Let $f(\mathbf{y} \mid \boldsymbol{\theta})$ be the known probability density of y indexed by unknown parameter $\boldsymbol{\theta}$. Denote the log-likelihood of n observations by $l(\boldsymbol{\theta} \mid \mathbf{y})$. If there are no restrictions on the parameters, a fast and popular algorithm for maximizing $l(\boldsymbol{\theta} \mid \mathbf{y})$ is a Newton–Raphson algorithm.

The score function and the observed information matrix for the Newton–Raphson algorithm are given by

$$\mathbf{S}_U = \frac{\partial l(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \mathbf{I}_U = -\frac{\partial^2 l(\boldsymbol{\theta} \mid \mathbf{y})}{\partial \boldsymbol{\theta}^2},$$

where $\mathbf{I}_U$ is assumed to be positive definite. So the unrestricted maximum likelihood estimate of $\boldsymbol{\theta}$ is a solution of a set of iterations given by

U1. $l \leftarrow 0$; choose a starting value for $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_{U(0)}$.

U2. $\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{U(l)}] \leftarrow \boldsymbol{\theta}_{U(l)} + \mathbf{I}_U^{-1}\mathbf{S}_U$, where $\mathbf{I}_U$ and $\mathbf{S}_U$ are evaluated at $\boldsymbol{\theta}_{U(l)}$. Stop if $\boldsymbol{\theta}_{U(l)}$ has converged.

U3. $\boldsymbol{\theta}_{U(l+1)} \leftarrow \boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{U(l)}]$; $l \leftarrow l + 1$; go to U2.

In U2, $\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{U(l)}]$ denotes the $(l + 1)$th term in the Newton–Raphson sequence for the unrestricted problem obtained by taking one Newton–Raphson step from $\boldsymbol{\theta}_{U(l)}$.

Now suppose that there are Q linearly independent restrictions on the parameter $\boldsymbol{\theta}$, such that

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{a}, \tag{1}$$

where $\mathbf{A}$ is a known linearly independent $Q \times p$ matrix defining the restriction, with $\operatorname{rank}(\mathbf{A}) = Q < p$, and $\mathbf{a}$ is a known $Q \times 1$ vector.

There are a number of approaches to maximizing the log-likelihood under the restriction (1). For generalized linear models, Nyquist (1991) proposed restricted maximum likelihood estimation using a quadratic penalty function. He considered situations in which iteratively reweighted least squares would be used if there was no restriction, and he showed how the algorithm could be adapted to satisfy the restriction. The resulting procedure can also be derived using a Lagrange multiplier approach. In this article, using the Lagrange multiplier method, we derive an algorithm to find the restricted maximum likelihood estimate. We use the relationship between the restricted solution and the unrestricted solution assuming that a Newton–Raphson algorithm is used to maximize the log-likelihood.

When the Lagrange multiplier method is used to incorporate the restriction, the restricted log-likelihood is given by $l(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda}) = l(\boldsymbol{\theta} \mid \mathbf{y}) - \boldsymbol{\lambda}'(\mathbf{a} - \mathbf{A}\boldsymbol{\theta})$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_Q)$ are Lagrangean multipliers. When $\boldsymbol{\lambda}$ is given, the procedure for maximization of the restricted log-likelihood $l(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda})$ is the same as for the unrestricted maximization in U1–U3.

A simple adaptation of the Newton–Raphson iteration scheme leads to the restricted solution. The score function and the information matrix for the restricted log-likelihood can be expressed as

$$\mathbf{S}_R = \frac{\partial l(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \mathbf{S}_U + \mathbf{A}'\boldsymbol{\lambda} \quad \text{and}$$

$$\mathbf{I}_R = -\frac{\partial^2 l(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\lambda})}{\partial^2 \boldsymbol{\theta}} = \mathbf{I}_U. \tag{2}$$

From the relationship of the score functions and information matrices between the unrestricted and restricted problems, we can easily verify that the Lagrange multiplier is a function of the unrestricted solution and the unrestricted information matrix. A sequence $\boldsymbol{\theta}_{R(0)}, \boldsymbol{\theta}_{R(1)}, \boldsymbol{\theta}_{R(2)}, \ldots$ for the restricted problem is obtained by the following algorithm:

R1. $l \leftarrow 0$; choose a starting value, $\boldsymbol{\theta}_{R(0)}$.

R2. Calculate $\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]$ from U2 for the unrestricted problem.

R3. Calculate $\boldsymbol{\theta}_{R(l+1)}$ for the restricted problem from the following equation:

$$\boldsymbol{\theta}_{R(l+1)} = \boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] + \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}(\mathbf{a} - \mathbf{A}\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]),$$

where $\mathbf{I}_U$ are evaluated at $\boldsymbol{\theta}_{R(l)}$. Stop if $\boldsymbol{\theta}_{R(l)}$ has converged.

R4. $l \leftarrow l + 1$; go to R2.

The expression in R3 is derived in the Appendix. From R3, it is clear that each member of the sequence for the restricted problem is easily obtained in each iteration by using the unrestricted solution and the information matrix.

3. THE EM ALGORITHM UNDER LINEAR RESTRICTIONS ON PARAMETERS

3.1 Factorizing the Likelihood Under Parameter Restrictions

Following the notation of Little and Rubin (1987), suppose that we have a model for the complete-data Y with associated density $f(Y \mid \boldsymbol{\theta})$. We partition the complete-data Y into $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} represents the observed part of Y and Y_{mis} denotes the missing part of Y. The distribution of the complete-data Y can be factored as

$$f(Y \mid \boldsymbol{\theta}) = f(Y_{\text{obs}}, Y_{\text{mis}} \mid \boldsymbol{\theta}) = f(Y_{\text{obs}} \mid \boldsymbol{\theta}) f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}). \tag{3}$$

So, based on (3), the log-likelihood can be written as

$$l(\boldsymbol{\theta} \mid Y_{\text{obs}}) = l(\boldsymbol{\theta} \mid Y) - \ln f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}). \tag{4}$$

Taking expectations of both sides of (4) over the distribution of the missing data Y_{mis}, given Y_{obs} and the current estimate of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(m)}$, gives $l(\boldsymbol{\theta} \mid Y_{\text{obs}}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$, where

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = \int l(\boldsymbol{\theta} \mid Y) f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}^{(m)})\, dY_{\text{mis}} \quad \text{and}$$

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = \int \ln f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}) f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}^{(m)})\, dY_{\text{mis}}.$$

The E step of the EM algorithm consists of finding the conditional expectation of the complete-data log-likelihood given the current parameter estimate and the observed data—that is, to evaluate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$. The M step consists of maximizing this conditional expectation over values of $\boldsymbol{\theta}$.

Now assume that for this incomplete data problem, the parameters must satisfy a set of linear restrictions given by (1). Consider a sequence of values obtained from an iterative algorithm $\boldsymbol{\theta}_R^{(0)}, \boldsymbol{\theta}_R^{(1)}, \ldots, \boldsymbol{\theta}_R^{(m)}, \ldots$, where $\boldsymbol{\theta}_R^{(m+1)} = M(\boldsymbol{\theta}_R^{(m)})$ for some function $M(\ \ )$ and where each member of the sequence $\boldsymbol{\theta}_R^{(m)}$ should satisfy the linear restrictions (1).

The difference in values of $l(\theta_R | Y_{\text{obs}})$ at successive iterates is given by

$$\begin{aligned} l(\theta_R^{(m+1)} | Y_{\text{obs}}) - l(\theta_R^{(m)} | Y_{\text{obs}}) &= (Q(\theta_R^{(m+1))} | \theta_R^{(m)}) - Q(\theta_R^{(m)} | \theta_R^{(m)})) \\ &\quad - (H(\theta_R^{(m+1)} | \theta_R^{(m)}) - H(\theta_R^{(m)} | \theta_R^{(m)})). \end{aligned} \tag{5}$$

From Jensen's inequality, we have

$$H(\theta_R^{(m+1)} | \theta_R^{(m)}) \le H(\theta_R^{(m)} | \theta_R^{(m)}). \tag{6}$$

Thus it is a property of the EM algorithm that if $\theta_R^{(m+1)}$ is chosen to increase $Q(\theta | \theta_R^{(m)})$ with respect to θ, this will ensure that the log-likelihood under the set of restrictions on the parameters also increases.

3.2 The Restricted EM Algorithm

Let $D^{10}Q(\theta^{(m+1)} | \theta^{(m)})$ and $D^{20}Q(\theta^{(m+1)} | \theta^{(m)})$ be the first and second partials of $Q(\theta | \theta^{(m)})$ with respect to θ evaluated at $\theta^{(m+1)}$; that is,

$$D^{10}Q(\theta^{(m+1)} | \theta^{(m)}) = \left. \frac{\partial Q(\theta | \theta^{(m)})}{\partial \theta} \right|_{\theta = \theta^{(m+1)}} \quad \text{and}$$

$$D^{20}Q(\theta^{(m+1)} | \theta^{(m)}) = \left. \frac{\partial^2 Q(\theta | \theta^{(m)})}{\partial \theta^2} \right|_{\theta = \theta^{(m+1)}}.$$

To find $\theta_R^{(m+1)}$, which maximizes $Q(\theta | \theta_R^{(m)})$ under the linear restrictions (1), we use the restricted maximization technique described in Section 2. That is, we use Newton–Raphson iterations, replacing $\mathbf{S}_U$ and $\mathbf{I}_U$ by $D^{10}Q(\theta | \theta_R^{(m)})$ and $-D^{20}Q(\theta | \theta_R^{(m)})$, and then apply the equation in R3. This gives the $(m+1)$th M-step solution, $\theta_R^{(m+1)}$, which maximizes $Q(\theta | \theta_R^{(m)})$ under the restriction (1). Thus we propose the following restricted EM algorithm:

The restricted EM algorithm (I)
E step: Evaluate $Q(\theta | \theta_R^{(m)})$.
RM step:

1. $l \leftarrow 0$; use the starting value $\theta_{R(l)} = \theta_R^{(m)}$.
2. Find $\mathbf{S}_U = D^{10}Q(\theta_{R(l)} | \theta_R^{(m)})$ and $\mathbf{I}_U = -D^{20} Q(\theta_{R(l)} | \theta_R^{(m)})$.
3. Find the restricted solution, $\theta_{R(l+1)}$, from R2–R3 and evaluate $Q(\theta_{R(l+1)} | \theta_R^{(m)})$. If $Q(\theta_{R(l+1)} | \theta_R^{(m)})$ has reached its maximum value as defined by a convergence criterion, then $\theta_R^{(m+1)} = \theta_{R(l+1)}$.
4. If $Q(\theta_{R(l+1)} | \theta_R^{(m)}) > Q(\theta_{R(l)} | \theta_R^{(m)})$, then $l \leftarrow l + 1$ and go to Step 2. If not, then do step halving on $\Delta\theta_{U(l+1)}[\theta_{R(l)}]$, where $\Delta\theta_{U(l+1)}[\theta_{R(l)}] = \theta_{U(l+1)}[\theta_{R(l)}] - \theta_{R(l)}$ and go to Step 2.

For this algorithm, we will assume that the function to be maximized is sufficiently well behaved and the starting value is appropriately chosen so that the foregoing Newton–Raphson algorithm modified by the step-halving procedure converges to the global maximum. When we apply R3 to find the restricted solution in the RM step, this restricted solution may not increase $Q(\theta | \theta_R^{(m)})$. To guarantee increase of this quantity, we propose a step-halving procedure applied to the Newton–Raphson estimate. In this procedure, if $Q(\theta_{R(l+1)} | \theta_R^{(m)}) < Q(\theta_{R(l)} | \theta_R^{(m)})$, then the intermediate value $\theta_{U(l+1)}[\theta_{R(l)}] = \theta_{R(l)} + \Delta\theta_{U(l+1)}[\theta_{R(l)}]$ in R3 is replaced by $\theta_{R(l)} + \frac{1}{2}\Delta\theta_{U(l+1)}[\theta_{R(l)}]$. If $Q(\theta_{R(l+1)} | \theta_R^{(m)})$ is still less than $Q(\theta_{R(l)} | \theta_R^{(m)})$, then $\theta_{U(l+1)}[\theta_{R(l)}]$ in R3 is replaced by $\theta_{R(l)} + \frac{1}{4}\Delta\theta_{U(l+1)}[\theta_{R(l)}]$. This procedure continues until a value of θ satisfying the restriction is found such that $Q(\theta | \theta_R^{(m)}) > Q(\theta_{R(l)} | \theta_R^{(m)})$. Such a value of θ is guaranteed to exist because of the local properties of $Q(\theta | \theta_R^{(m)})$ when small increments in the intermediate value $\theta_{U(l+1)}[\theta_{R(l)}]$ in R3 are taken in the direction $\Delta\theta_{U(l+1)}[\theta_{R(l)}]$ from $\theta_{R(l)}$. The theoretical support for using this step-halving procedure is given in the next section.

After convergence of the successive iterations, the converged $(m+1)$th RM step solution is denoted by $\theta_R^{(m+1)}$. Because $\theta_R^{(m+1)}$ is a maximum point under the restriction, we have $Q(\theta_R^{(m+1)} | \theta_R^{(m)}) > Q(\theta_R^{(m)} | \theta_R^{(m)})$. Thus, prior to convergence, the observed data log-likelihood is increased at each step of the restricted EM algorithm (I). But the restricted EM algorithm (I) has an unattractive aspect, because it involves iterations within each EM iteration. Whereas the EM algorithm chooses $\theta_R^{(m+1)}$ to maximize $Q(\theta | \theta_R^{(m)})$ with respect to θ, the GEM (generalized EM) algorithm (Dempster et al. 1977; Wu 1983) chooses any $\theta_R^{(m+1)}$ so that $Q(\theta_R^{(m+1)} | \theta_R^{(m)})$ is greater than $Q(\theta_R^{(m)} | \theta_R^{(m)})$. Thus, following the spirit of GEM, it is not necessary to carry out a full Newton–Raphson algorithm to converge in the RM step. One-step iteration of the Newton–Raphson algorithm (Lange 1991) or several steps of iteration that increase $Q(\theta | \theta_R^{(m)})$ enough to be greater than $Q(\theta_R^{(m)} | \theta_R^{(m)})$ can be used in the RM step.

When, as frequently occurs, there is a closed-form expression for the unrestricted M-step solution, then this solution can be used in the RM step. Thus we propose a second restricted EM algorithm that is computationally simpler, because it does not require any Newton–Raphson iteration.

The restricted EM algorithm (II)
E step: Evaluate $Q(\theta | \theta_R^{(m)})$.
RM step:

1. Find the unrestricted solution, θ_U^+, which has a closed form, and calculate $\mathbf{I}_U = -D^{20}Q(\theta_R^{(m)} | \theta_R^{(m)})$.
2. From R3, obtain the restricted solution, θ_R^+; that is, $\theta_R^+ = \theta_U^+ + \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}(\mathbf{a} - \mathbf{A}\theta_U^+)$.
3. If $Q(\theta_R^+ | \theta_R^{(m)}) > Q(\theta_R^{(m)} | \theta_R^{(m)})$, then $\theta_R^{(m+1)} = \theta_R^+$.
4. If not, then do step halving on $\Delta\theta_U^+$, where $\Delta\theta_U^+ = \theta_U^+ - \theta_R^{(m)}$.

To find the restricted solution, $\overset{(m+1)}{\theta_R}$, this algorithm uses the complete-data information matrix, $\mathbf{I}_U = -D^{20}Q(\theta_R^{(m)} | \theta_R^{(m)})$, and the closed-form unrestricted M-step solution, θ_U^+. In the restricted EM algorithm (II), if the procedure for obtaining $\theta_R^{(m+1)}$ does not increase the likelihood, we suggest applying the step-halving procedure on $\Delta\theta_U^+$. The theoretical support for using step halving in the restricted EM (II) is based on the result in the next section.

One convenient aspect of the restricted EM algorithms (I) and (II) is that they do not require complicated constrained maximization procedures to find the maximum likelihood estimates. The restricted EM algorithms (I) and (II) we developed here can be easily carried out by using the code of the complete-data information matrix in addition to the code of the usual EM algorithm without any further calculation.

4. THEORETICAL PROPERTIES OF THE RESTRICTED EM ALGORITHM

Because the restricted EM algorithms (I) and (II) are adaptations of the EM and the GEM algorithms, they have some of the same theoretical properties as the EM and the GEM algorithms. We illustrate the theoretical properties of algorithms (I) and (II) adapting the GEM algorithm results of Dempster et al. (1977) and Wu (1983). Let Ω_R be the parameter space under restrictions on the parameters. From (5) and (6), we obtain the following theorem.

Theorem 1. Suppose that $\boldsymbol{\theta}_R^{(m+1)}$, $m = 0, 1, 2, \ldots$, is a sequence of iterations of the restricted EM algorithms (I) and (II); then we have

$$l(\boldsymbol{\theta}_R^{(m+1)} \mid Y_{\mathrm{obs}}) \geq l(\boldsymbol{\theta}_R^{(m)} \mid Y_{\mathrm{obs}}),$$

with equality if and only if $Q(\boldsymbol{\theta}_R^{(m+1)} \mid \boldsymbol{\theta}_R^{(m)}) = Q(\boldsymbol{\theta}_R^{(m)} \mid \boldsymbol{\theta}_R^{(m)})$ and $H(\boldsymbol{\theta}_R^{(m+1)} \mid \boldsymbol{\theta}_R^{(m)}) = H(\boldsymbol{\theta}_R^{(m)} \mid \boldsymbol{\theta}_R^{(m)})$ almost everywhere.

Corollary 1. Suppose that $\boldsymbol{\theta}_R^{(m+1)}$, $m = 0, 1, 2, \ldots$, converges to $\boldsymbol{\theta}_R^*$, where $l(\boldsymbol{\theta}_R^* \mid Y_{\mathrm{obs}}) \geq l(\boldsymbol{\theta} \mid Y_{\mathrm{obs}})$ for all $\boldsymbol{\theta} \in \Omega_R$ in the restricted EM algorithms (I) and (II); then

$$M(\boldsymbol{\theta}_R^*) = \boldsymbol{\theta}_R^*, \quad \text{and thus} \quad l(M(\boldsymbol{\theta}_R^*) \mid Y_{\mathrm{obs}}) = l(\boldsymbol{\theta}_R^* \mid Y_{\mathrm{obs}}),$$

$$Q(M(\boldsymbol{\theta}_R^*) \mid Y_{\mathrm{obs}}) = Q(\boldsymbol{\theta}_R^* \mid Y_{\mathrm{obs}}), \quad \text{and}$$

$$f(Y_{\mathrm{mis}} \mid Y_{\mathrm{obs}}, M(\boldsymbol{\theta}_R^*)) = f(Y_{\mathrm{mis}} \mid Y_{\mathrm{obs}}, \boldsymbol{\theta}^*).$$

Theorem 1 shows that $l(\boldsymbol{\theta} \mid Y_{\mathrm{obs}})$ is nondecreasing in each iteration of the restricted EM algorithms (I) and (II) and is strictly increasing in any iteration such that $Q(\boldsymbol{\theta}_R^{(m+1)} \mid \boldsymbol{\theta}_R^{(m)}) > Q(\boldsymbol{\theta}_R^{(m)} \mid \boldsymbol{\theta}_R^{(m)})$. Corollary 1 says that for each algorithm, the maximum likelihood estimate of θ under parameter restrictions obtained at convergence is a fixed point in the restricted parameter space.

In the restricted EM algorithm (I), each M step involves iterations, in which we find the unrestricted Newton–Raphson step first and then the restricted solution. Also, we use a step-halving procedure applied to the Newton–Raphson algorithm. The theoretical support for this updating scheme is given by the following theorem and corollary. Let $R(\mathbf{A})$ be a subspace spanned by the rows of $\mathbf{A}$ and let $N(\mathbf{A})$ be a null space of $\mathbf{A}$.

Theorem 2: Restricted EM (I). The incremental step $\Delta\boldsymbol{\theta}_{R(l+1)} = \boldsymbol{\theta}_{R(l+1)} - \boldsymbol{\theta}_{R(l)}$, $l = 1, 2, \ldots$ lies in a feasible direction defined by the linear restriction; that is, $\Delta\boldsymbol{\theta}_{R(l+1)} \in N(\mathbf{A})$. Furthermore, prior to convergence, we have $D^{10}Q(\boldsymbol{\theta}_{R(l)} \mid \boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_{R(l+1)} > 0$.

Corollary 2: Restricted EM (I). Step halving applied to the step $\Delta\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] = \boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] - \boldsymbol{\theta}_{R(l)}$ provides a means of finding a value of $\boldsymbol{\theta}$ to ensure that $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{R(l)})$ increases.

When we have a closed form for the unrestricted solution, we can apply the restricted EM algorithm (II) directly. The theoretical support for this updating scheme and the step-halving procedures are based on the following theorem and corollary.

Theorem 3: Restricted EM (II). The incremental step $\Delta\boldsymbol{\theta}_R^+ = \boldsymbol{\theta}_R^+ - \boldsymbol{\theta}_R^{(m)}$ lies along a feasible direction defined by the linear restriction; that is, $\Delta\boldsymbol{\theta}_R^+ \in N(\mathbf{A})$. Furthermore, when $\|\Delta\boldsymbol{\theta}_U^+\|$ is small enough, prior to convergence, we have $D^{10}Q(\boldsymbol{\theta}_R^{(m)} \mid \boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_R^+ > 0$.

Corollary 3: Restricted EM (II). Step halving applied to the step $\Delta\boldsymbol{\theta}_U^+ = \boldsymbol{\theta}_U^+ - \boldsymbol{\theta}_R^{(m)}$ provides a means of finding a value of $\boldsymbol{\theta}$ to ensure that $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_R^{(m)})$ increases.

In this article we propose step halving as a strategy to find an acceptable step size. More effective strategies for adjusting the step size have been discussed by Dennis and Schnabel (1983, p. 126).

5. APPLICATION OF THE RESTRICTED EM ALGORITHM TO HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

Illustrations of the restricted EM algorithm with various types of linear restrictions on the parameters in a finite mixture model for a specific application were given by Kim (1991) and will be presented in a later work. In this section we illustrate use of the algorithm to construct test statistics and confidence intervals for parameters of interest. The construction of test statistics and confidence intervals are important issues when the EM algorithm is applied. This is because the EM algorithm does not automatically produce the variance–covariance matrix for parameters. Therefore, to test a null hypothesis or to construct confidence intervals for parameters, additional steps are needed to find the variance–covariance matrix (Louis 1982; Meng and Rubin 1991).

Assume that we are interested in testing the following hypothesis:

$$H_0\colon \mathbf{A}\boldsymbol{\theta} = \mathbf{a}, \tag{7}$$

where $\mathbf{A}$ and $\mathbf{a}$ are defined in Section 2.1. Because the null hypothesis is a statement about a linear combination of parameters, we can apply the restricted EM algorithm to estimate $\boldsymbol{\theta}_R^*$ under the null hypothesis. The likelihood ratio test statistic, to compare the full model to the reduced model, is defined by

$$r = -2(l(\boldsymbol{\theta}_R^* \mid Y_{\mathrm{obs}}) - l(\boldsymbol{\theta}_F^* \mid Y_{\mathrm{obs}})), \tag{8}$$

where $\boldsymbol{\theta}_F^*$ is the maximum likelihood estimate under the full model. Here r, under suitable regularity conditions, has an asymptotic χ^2 distribution, with the degree of freedom determined by the difference in number of parameters between the full model and the reduced model.

The profile likelihood confidence interval for a scalar component of $\boldsymbol{\theta}$ can be obtained directly from the likelihood function by inverting the likelihood ratio test. Consider the likelihood ratio test of the hypothesis $H_0\colon \theta_j = \theta_0$, where θ_j is the jth element of $\boldsymbol{\theta}$. This is a special case of the general linear hypothesis (7). The likelihood ratio test statistic for the null hypothesis is to reject at level α if $r > \chi^2_{\alpha,1}$, which implies that the $100(1 - \alpha)\%$ likelihood-based confidence interval for θ_j is

$$\left(\theta_0 \colon l(\theta_R^* \mid Y_{\text{obs}}) - l(\theta_F^* \mid Y_{\text{obs}}) > -\frac{1}{2}\chi^2_{\alpha,1}\right), \qquad (9)$$

where $\boldsymbol{\theta}_R^*$ is the restricted EM solution under $H_0\colon \theta_j = \theta_0$.

Thus the procedure for finding the likelihood-based confidence interval is to apply the restricted EM algorithm to find $\boldsymbol{\theta}_R^*$ over a grid of values of θ_0 and plot the profile of the log-relative likelihood. One reason to prefer this likelihood-based approach compared to the Wald-type confidence interval is that profile likelihood confidence interval for $\boldsymbol{\theta}$ is not forced to be symmetric in situations in which the log-likelihood function is asymmetric. If the log-likelihood is quadratic in the parameter, then the Wald-type confidence interval is asymptotically equivalent to the likelihood-based confidence interval (Aitken, Anderson, Francis, and Hinde 1989).

6. EXAMPLES

6.1 A Variance Component Model

Variance component models are used when the factor levels are not of primary interest in themselves but are considered a sample from a large population of factor levels. Table 1 shows Apex Enterprises data from Neter, Wasserman, and Kutner (1985, p. 648). These data are from a study of the evaluation ratings of potential employees by personnel officers. Five personnel officers were selected at random, and four candidates were randomly chosen and evaluated by all of the personnel officers. A single-factor variance component model for these data is given by

$$y_{ij} = b_i + e_{ij}, \qquad i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, J,$$

where b_i's are independent $N(\mu, \sigma_b^2)$, e_{ij} are independent $N(0, \sigma^2)$, and b_i and e_{ij} are independent random variables.

Dempster et al. (1977) and Little and Rubin (1987, p. 149) illustrated how the EM algorithm can be applied for this model. The unknown parameters to be estimated are $\boldsymbol{\theta} = (\mu, \sigma_b^2, \sigma^2)$. Regarding the y_{ij} as observed data and treating the unobserved random variables $b_1, b_2, \ldots, b_I$ as missing data, we can use the EM algorithm to obtain maximum likelihood estimates of $\boldsymbol{\theta}$. Let $z = (\mathbf{y}, \mathbf{b})$ be the complete data. The complete-data likelihood can be expressed as the product of two factors, the first corresponding to the distribution of y_{ij}, given b_i and $\boldsymbol{\theta}$, and the second corresponding to the distribution of b_i, given $\boldsymbol{\theta}$. The complete data log-likelihood is given by

$$l(\boldsymbol{\theta} \mid z) = -\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(y_{ij} - b_i)^2}{\sigma^2} - \frac{1}{2}\sum_{i=1}^{I}\frac{(b_i - \mu)^2}{\sigma_b^2}$$
$$-\frac{IJ}{2}\ln(\sigma^2) - \frac{I}{2}\ln(\sigma_b^2) - \frac{IJ}{2}\ln(2\pi) - \frac{I}{2}\ln(2\pi).$$

The EM algorithm for this model is as follows:

E Step: Evaluate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = E(l(\boldsymbol{\theta} \mid z) \mid \mathbf{y}, \boldsymbol{\theta}^{(m)})$.

Table 1. Apex Enterprises Data

Officer (i)	Candidate (j) 1	2	3	4
A	76	64	85	75
B	58	75	81	66
C	49	63	62	46
D	74	71	85	90
E	66	74	81	79

In this step we need to evaluate the conditional distribution of b_i, given $\mathbf{y}$. From Bayes's theorem applied to the joint distribution of b_i and $\mathbf{y}$, we obtain

$$[b_i \mid \mathbf{y}, \boldsymbol{\theta}] \sim \text{independent } N(w\mu + (1 - w)\bar{y}_{i\cdot}, v),$$

where

$$\bar{y}_{i\cdot} = \frac{1}{J}\sum_{i=1}^{J} y_{ij}, \qquad w = \frac{\sigma^2}{\sigma^2 + J\sigma_b^2}, \quad \text{and} \quad v = \sigma_b^2 w.$$

M step: $\boldsymbol{\theta}^{(m+1)}$ is given by

$$\mu^{(m+1)} = \sum_{i=1}^{I} E(b_i \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}),$$

$$(\sigma_b^2)^{(m+1)} = \sum_{i=1}^{I} E((b_i - \mu)^2 \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}),$$

and

$$(\sigma^2)^{(m+1)} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{\cdot\cdot})^2$$
$$+ \frac{1}{I}\sum_{i=1}^{I} E((\bar{y}_{i\cdot} - b_i)^2 \mid \mathbf{y}, \boldsymbol{\theta}^{(m)}).$$

Now suppose that we are interested in the ratio of variances σ^2 and σ_b^2—for example, we want to test $H_0\colon \sigma_b^2 = .5\sigma^2$. The ratio of variances is useful for determining the balance between the number of officers and candidates (Neter et al. 1985). The restricted EM algorithm under H_0 will be straightforward. Because we have a closed form for the unrestricted solution, algorithm (II) can be used with $\mathbf{A} = (0\ 1\ {-.5})$ and $\mathbf{a} = 0$ in (2). The solution is $\mu_R^* = 71.0$, $(\sigma_b^2)_R^* = 41.9278$, $(\sigma^2)_R^* = 83.8556$. The likelihood ratio test for H_0 is $r = -2(l(\boldsymbol{\theta}_R^* \mid Y_{\text{obs}}) - l(\boldsymbol{\theta}^* \mid Y_{\text{obs}})) = -2(-75.1195 - (-75.0456)) = .1478$. Because $r = .1478 < 3.84$, we cannot reject the null hypothesis.

To find confidence intervals, we apply the restricted EM algorithm at a set of grid points for each parameter and calculate the log-relative likelihood of (9). Figure 1 shows the profile of the log-relative likelihood for each component of the parameter $\boldsymbol{\theta}$. We can see that the profile of the mean looks symmetric, but that profiles of others are far from symmetric. Table 2 shows the comparison between the likelihood-based approach and the traditional least squares method. In the least squares method, the confidence interval for μ is based on the t distribution and those for others are based on the χ^2 distribution. Detailed calculations were shown by Neter et al. (1985). Notice that likelihood-based confidence intervals are narrower than those from the least squares method, and the confidence intervals of the ratio of two variances are substantially different. Moreover, the

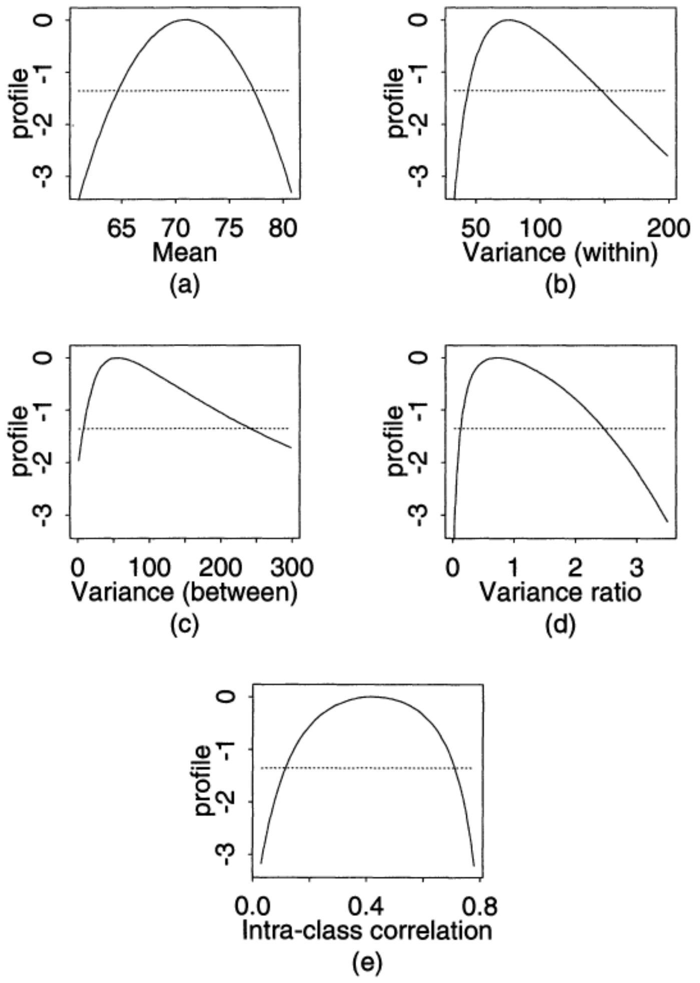

Figure 1. Profile Likelihoods of the Parameters in the Variance Component Model: (a) μ; (b) σ^2; (c) σ_b^2; (d) σ_b^2/σ^2; (e) $\sigma_b^2/(\sigma_b^2 + \sigma^2)$.

likelihood-based approach produces confidence intervals for the between variance, σ_b^2, which is not available from the least squares method.

6.2 Bivariate Normal Model

We assume that the data in Table 3 (Little and Rubin 1987, p. 101) follow a bivariate normal distribution, with parameter $\theta = (\mu_1, \sigma_{11}, \mu_2, \sigma_{22}, \rho)$.

We see that y_{i1} and y_{i2}, $i = 1, 2, \ldots, 12$ are fully observed, but y_{i2}, $i = 13, \ldots, 18$ are missing. The log-likelihood based on the complete data is given by

$$l(\theta|\mathbf{y}) = -\frac{n}{2}\ln(\sigma_{11}\sigma_{22}(1-\rho^2)) - \frac{1}{2(1-\rho^2)} \times \sum_{i=1}^{n}\left(\frac{(y_{i1}-\mu_1)^2}{\sigma_{11}} - 2\rho\,\frac{(y_{i1}-\mu_1)(y_{i2}-\mu_2)}{\sqrt{\sigma_{11}\,\sigma_{22}}} + \frac{(y_{i2}\ -\mu_2)^2}{\sigma_{22}}\right).$$

The E step is to find $E(l(\theta|\mathbf{y})|Y_{\text{obs}}, \theta^{(m)})$. This requires calculating

$$T_{i2}^{(1)} = y_{i2}, \qquad T_{i2}^{(2)} = y_{i2}^2 \quad \text{if } y_{i2} \text{ is observed,}$$

$$\text{and} \quad T_{i2}^{(1)} = E(y_{i2}\,|\,y_{i1}, \theta^{(m)}),$$

$$T_{i2}^{(2)} = E(y_{i2}^2\,|\,y_{i1}, \theta^{(m)}) \quad \text{if } y_{i2} \text{ is missing.}$$

Then,

$$Q(\theta|\theta^{(m)}) = -\frac{n}{2}\ln(\sigma_{11}\sigma_{22}(1-\rho^2)) - \frac{1}{2(1-\rho^2)} \times \sum_{i=1}^{n}\left(\frac{(y_{i1}-\mu_1)^2}{\sigma_{11}} - 2\rho\,\frac{(y_{i1}-\mu_1)(T_{i2}^{(1)}-\mu_2)}{\sqrt{\sigma_{11}\,\sigma_{22}}} + \frac{T_{i2}^{(2)} - 2\mu_2 T_{i2}^{(1)} + \mu_2^2}{\sigma_{22}}\right).$$

The required conditional expectations are given by $T_{i2}^{(1)} = \beta_{20\cdot 1} + \beta_{21\cdot 1}y_{i1}$ and $T_{i2}^{(2)} = (\beta_{20\cdot 1} + \beta_{21\cdot 1}y_{i1})^2 + \sigma_{22\cdot 1}$, where

$$\beta_{21\cdot 1} = \frac{s_{12}}{s_{11}}, \qquad \beta_{20\cdot 1} = \frac{\sum_{i=1}^{n} T_{i2}^{(1)}}{n} - \beta_{21\cdot 1}\frac{\sum_{i=1}^{n} y_{i1}}{n},$$

$$\sigma_{22\cdot 1} = s_{22} - \frac{s_{21}^2}{s_{11}}.$$

The unrestricted M-step solution has a closed form. Let

$$s_1 = \sum_{i=1}^{n} y_{i1}, \qquad s_2 = \sum_{i=1}^{n} T_{i2}^{(1)}, \qquad s_{11} = \sum_{i=1}^{n} y_{i1}^2,$$

$$s_{22} = \sum_{i=1}^{n} T_{i2}^{(2)}, \qquad s_{12} = \sum_{i=1}^{n} y_{i1}T_{i2}^{(1)}.$$

Then the new estimates are given by

$$\mu_1^{(m+1)} = \frac{s_1}{n}, \qquad \mu_2^{(m+1)} = \frac{s_2}{n},$$

$$\sigma_{11}^{(m+1)} = \frac{s_{11}}{n} - (\mu_1^{(m+1)})^2, \qquad \sigma_{22}^{(m+1)} = \frac{s_{22}}{n} - (\mu_2^{(m+1)})^2,$$

$$\rho^{(m+1)} = \frac{\sigma_{12}^{(m+1)}}{\sqrt{\sigma_{11}^{(m+1)}\,\sigma_{22}^{(m+1)}}},$$

where $\sigma_{12}^{(m+1)} = s_{12}/n - \mu_1^{(m+1)}\mu_2^{(m+1)}$.

We calculate the complete-data information matrix, $\mathbf{I}_U = -D^{20}Q(\theta^{(m)}|\theta^{(m)})$, and use the restricted EM algorithm (II) to construct confidence intervals for the parameters. Figure 2 shows the profiles of the log-relative likelihood of

Table 2. Point Estimates and 90% Confidence Intervals for Apex Enterprises Data

	Likelihood-based approach	Least squares approach
Parameters	*Estimate (90% CI)*	*Estimate (90% CI)*
μ	71.0 (64.69, 77.31)	71.0 (61.83, 80.17)
σ_b^2	55.1 (7.80, 241.54)	73.6 (—, —)
σ^2	75.6 (43.77, 147.43)	75.6 (45.36, 165.20)
σ_b^2/σ^2	.729 (.131, 2.466)	.974 (.150, 6.947)
$\sigma_b^2/(\sigma_b^2 + \sigma^2)$	.422 (.116, .712)	.493 (.130, .873)

Table 3. Bivariate Normal Data (n = 18)

y_{i1}	8	6	11	22	14	17	18	24	19	23	26	40	4	4	5	6	8	10
y_{i2}	59	58	56	53	50	45	43	42	39	38	30	27	*	*	*	*	*	*

the five parameters. The profiles of the means, μ_1 and μ_2, look symmetric; in contrast, the profiles of σ_{11}, σ_{22}, and ρ are not symmetric. Table 4 shows the maximum likelihood estimates and 95% confidence intervals from the likelihood-based approach and from the Wald-type approach. The Wald-type confidence intervals are based on the inverse of the observed information matrix (Little and Rubin 1987, p. 106). For the means, the two confidence intervals are similar; however, for the variances and correlation, these two confidence intervals are not the same. The Wald-type confidence intervals are symmetric around the points estimates, whereas the likelihood-based approach allows asymmetric confidence intervals. In particular, we notice that the profile of the correlation is extremely asymmetric. The likelihood-based confidence interval for ρ is $(-.9508, -.7009)$, which is wider than the Wald-type confidence interval $(-1.0018, -.7882)$ and does not include impossible values less than -1.

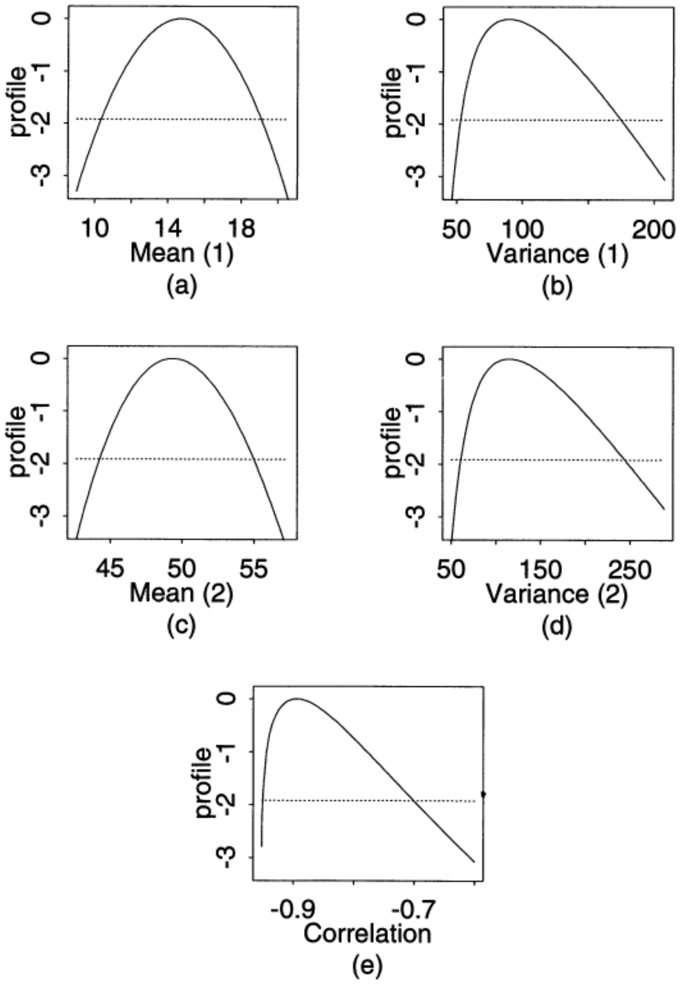

Figure 2. Profile Likelihoods of the Parameters in the Bivariate Normal Model: (a) μ_1; (b) σ_{11}; (c) μ_2; (d) σ_{22}; (e) ρ.

7. DISCUSSION AND CONCLUSIONS

We have studied maximum likelihood estimation under linear restrictions on parameters for incomplete-data problems. We have proposed the restricted EM algorithms (I) and (II) for the estimation procedure. In both algorithms, each step is easily updated by using the usual EM code and the complete-data information matrix without further complicated calculation.

We show that the likelihood is nondecreasing on each iteration of the restricted EM algorithms (I) and (II), and the sequence of values defined by the algorithm converges to a fixed point. We also find that step halving is a convenient modification of the step size when the RM step in the restricted EM algorithm (I) consists of one or more iterates of a Newton–Raphson algorithm. We show that a step-halving procedure applied to the direction defined by the closed-form solution in the restricted EM algorithm (II) would guarantee that the complete-data log-likelihood in the RM step would increase. To apply the step-halving procedure for several parameters, one might try to reduce the step for each parameter in turn; however, this would require more calculation. Reducing the step size for all parameters by the same factor is a simple procedure that will work.

To find the maximum likelihood estimate under the linear restriction, an alternative approach is to maximize the observed data log-likelihood under the restrictions. But this requires a complicated constrained optimization routine. The EM algorithm will tend to be more numerically stable, because the complete-data log-likelihood of the EM setting typically has a simpler form than the observed-data log-likelihood. Moreover, frequently we can find a closed-form expression for the maximum of the complete-data log-likelihood. Another possible approach is to apply the EM algorithm after reparameterization. When the restriction is just a simple linear combination of parameters, reparameterization might be easy; however, when the restriction is a set of linear restrictions of parameters, this might be complicated.

The most obvious application of the restricted EM algorithm is to find the maximum likelihood estimate for models

Table 4. Point Estimates and Confidence Intervals for Bivariate Normal Model

		95% confidence intervals	
Parameters	MLE	Likelihood-based interval	Wald type interval
μ_1	14.722	(10.352, 19.092)	(10.351, 19.094)
σ_{11}	89.534	(53.01, 174.71)	(31.318, 147.750)
μ_2	49.333	(44.242, 54.986)	(43.98, 54.69)
σ_{22}	114.695	(60.79, 243.99)	(30.68, 198.71)
ρ	−.895	(−.9508, −.7009)	(−1.0018, −.7882)

where there are linear restrictions on the parameters. Other important applications of the restricted EM algorithm, which have been the focus of this article, are to construct hypothesis tests and confidence intervals for situtations where the EM algorithm is used. We illustrated some applications of this algorithm to obtain likelihood ratio tests and likelihood-based confidence intervals. Other approaches to hypothesis testing and construction of confidence intervals following the EM algorithm have been described by Louis (1982), who suggested a formula for the observed information matrix, and Meng and Rubin (1991), who used the rate of convergence of the EM algorithm. These asymptotic approaches are less likely to be accurate if the log-likelihood is not a quadratic function of the parameters.

A possible extension of the restricted EM algorithm is to the situation where the restriction involves inequalities or is nonlinear. We are currently investigating whether the algorithm can be adapted to accommodate these types of constraints.

APPENDIX: DERIVATION AND PROOFS

Derivation of the Result R3

Let $\boldsymbol{\theta}_{R(l)}$, $l = 0, 1, \ldots$ be the sequences of the Newton–Raphson iteration scheme. The $(l + 1)$th element of the Newton–Raphson sequence is obtained from the restricted score function and the information matrix. From (2) and U2, we can derive

$$\begin{aligned}\boldsymbol{\theta}_{R(l+1)} &= \boldsymbol{\theta}_{R(l)} + \mathbf{I}_R^{-1}\mathbf{S}_R = \boldsymbol{\theta}_{R(l)} + \mathbf{I}_U^{-1}(\mathbf{S}_U + \mathbf{A}'\boldsymbol{\lambda})\\ &= \boldsymbol{\theta}_{R(l)} + \mathbf{I}_U^{-1}\mathbf{S}_U + \mathbf{I}_U^{-1}\mathbf{A}'\boldsymbol{\lambda} = \boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] + \mathbf{I}_U^{-1}\mathbf{A}'\boldsymbol{\lambda},\end{aligned}$$

where $\mathbf{I}_U$ and $\mathbf{S}_U$ are evaluated at $\boldsymbol{\theta}_{R(l)}$. From (1), we have

$$\mathbf{a} - \mathbf{A}(\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] + \mathbf{I}_U^{-1}\mathbf{A}'\boldsymbol{\lambda}) = 0,$$

which implies

$$\boldsymbol{\lambda} = (\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}(\mathbf{a} - \mathbf{A}\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]).$$

Thus we have

$$\boldsymbol{\theta}_{R(l+1)} = \boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] + \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}(\mathbf{a} - \mathbf{A}\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]).$$

This result can also be obtained using the extension to general likelihood problems of the penalty function approach developed by Nyquist (1991).

Proof of Theorem 2

Let $\mathbf{S}_U = D^{10}Q(\boldsymbol{\theta}_{R(l)}|\boldsymbol{\theta}_R^{(m)})$, $\mathbf{I}_U = -D^{20}Q(\boldsymbol{\theta}_{R(l)}|\boldsymbol{\theta}_R^{(m)})$ and $\mathbf{B} = \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}$. We have

$$\begin{aligned}\Delta\boldsymbol{\theta}_{R(l+1)} &= \boldsymbol{\theta}_{R(l+1)} - \boldsymbol{\theta}_{R(l)} = \boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]\\ &\quad + \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}(\mathbf{a} - \mathbf{A}\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]) - \boldsymbol{\theta}_{R(l)}\\ &= (\mathbf{I} - \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}\mathbf{A})(\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] - \boldsymbol{\theta}_{R(l)})\\ &= (\mathbf{I} - \mathbf{B}\mathbf{A})\Delta\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}] = (\mathbf{I} - \mathbf{B}\mathbf{A})\mathbf{I}_U^{-1}\mathbf{S}_U.\end{aligned}$$

Because $\mathbf{ABA} = \mathbf{A}$, $\mathbf{B}$ is a generalized inverse of $\mathbf{A}$ and $(\mathbf{I} - \mathbf{BA})$ is idempotent matrix. Thus $(\mathbf{I} - \mathbf{BA})$ is an orthogonal projector onto $N(\mathbf{A})$ along $R(\mathbf{A})$. Therefore, we have $\mathbf{A}\Delta\boldsymbol{\theta}_{R(l+1)} = \mathbf{A}(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1}\mathbf{S}_U = 0$, because $\mathbf{A}(\mathbf{I} - \mathbf{BA}) = 0$, which implies $\Delta\boldsymbol{\theta}_{R(l+1)} \in N(\mathbf{A})$. The next step is to show that prior to convergence, $\mathbf{H} = D^{10}Q(\boldsymbol{\theta}_{R(l)}|\boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_{R(l+1)} = \mathbf{S}_U'(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1}\mathbf{S}_U > 0$. Simple algebra gives

$$(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1} = \mathbf{I}_U^{-1} - \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}\mathbf{A}\mathbf{I}_U^{-1}.$$

Let $\mathbf{F}$ be $p \times d$ $(d = p - Q)$ matrix of rank d such that $\mathbf{AF} = 0$. From Seber (1984, p. 536), if $\mathbf{F}$ is any $p \times d$ matrix of rank d $(d = p - Q)$ that satisfies $\mathbf{AF} = 0$, then it can be shown that $(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1} = \mathbf{F}(\mathbf{F}'\mathbf{I}_U\mathbf{F})^{-1}\mathbf{F}' = \mathbf{G}$, where $\mathbf{G}$ is a positive semidefinite with rank d. Let $M_1 = \{\mathbf{x}\colon \mathbf{x}'\mathbf{G}\mathbf{x} = 0\}$ and $M_2 = \{\mathbf{x}\colon \mathbf{x}'\mathbf{G}\mathbf{x} > 0\}$. We will show that prior to convergence, $\mathbf{S}_U \in M_2$. To see this, in the set M_1 we have $\mathbf{x}'\mathbf{G}\mathbf{x} = \mathbf{x}'\mathbf{F}(\mathbf{F}'\mathbf{I}_U\mathbf{F})^{-1}\mathbf{F}'\mathbf{x} = 0$, which implies $\mathbf{x}'\mathbf{F} = 0$, and hence $\mathbf{x}'$ is a linear combination of the rows of $\mathbf{A}$, because $\mathbf{AF} = 0$ and $\operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{F}) = p$. Therefore, if $\mathbf{S}_U \in M_1$, then we have $\mathbf{S}_U' \in R(\mathbf{A})$, which implies $\Delta\boldsymbol{\theta}_{R(l+1)} = (\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1}\mathbf{S}_U = 0$, indicating that $\boldsymbol{\theta}_{R(l)}$ is a stationary point. Thus we have shown that $\mathbf{S}_U \in M_2$ prior to convergence of the algorithm, which implies $\mathbf{S}_U'(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1}\mathbf{S}_U > 0$.

Proof of Corollary 2

Let $\boldsymbol{\theta}_{U(l+1),t}[\boldsymbol{\theta}_{R(l)}] = \boldsymbol{\theta}_{R(l)} + t\Delta\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]$, $0 < t < 1$ and $\boldsymbol{\theta}_{R(l+1),t}$ be the restricted solution corresponding to $\boldsymbol{\theta}_{U(l+1),t}[\boldsymbol{\theta}_{R(l)}]$ obtained from R3. Let $\mathbf{S}_U$, $\mathbf{I}_U$, and $\mathbf{B}$ be defined in the same way as in the proof of Theorem 2. We assume that $\mathbf{I}_U$ is positive definite. A Taylor series expansion around $t = 0$ of $Q(\boldsymbol{\theta}_{R(l+1),t}|\boldsymbol{\theta}_{R(l)})$ at $\boldsymbol{\theta}_{R(l)}$ gives

$$\begin{aligned}&Q(\boldsymbol{\theta}_{R(l+1),t}|\boldsymbol{\theta}_{R(l)})\\ &\qquad = Q(\boldsymbol{\theta}_{R(l)}|\boldsymbol{\theta}_{R(l)}) + t\mathbf{S}_U'(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1}\mathbf{S}_U + O(t^2). \quad \text{(A.1)}\end{aligned}$$

From the proof of Theorem 2, prior to convergence we have $\mathbf{S}_U'(\mathbf{I} - \mathbf{BA})\mathbf{I}_U^{-1}\mathbf{S}_U > 0$. Therefore, if we perform step halving on $\Delta\boldsymbol{\theta}_{U(l+1)}[\boldsymbol{\theta}_{R(l)}]$, t is reduced, and, for small enough t, the positive $O(t)$ term will be the dominant term on the right side of (A.1). Therefore, prior to convergence, we can find a positive t such that $Q(\boldsymbol{\theta}_{R(l+1),t}|\boldsymbol{\theta}_{R(l)}) \geq Q(\boldsymbol{\theta}_{R(l)}|\boldsymbol{\theta}_{R(m)})$ with equality holding only if we are at convergence.

Proof of Theorem 3

Let $\mathbf{I}_U = -D^{20}Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})$ and $\mathbf{C} = \mathbf{I}_U^{-1}\mathbf{A}'(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}')^{-1}$. Simple algebra gives $\Delta\boldsymbol{\theta}_R^+ = \boldsymbol{\theta}_R^+ - \boldsymbol{\theta}_R^{(m)} = (\mathbf{I} - \mathbf{CA})(\boldsymbol{\theta}_U^+ - \boldsymbol{\theta}_R^{(m)}) = (\mathbf{I} - \mathbf{CA})\Delta\boldsymbol{\theta}_U^+$. Here $(\mathbf{I} - \mathbf{CA})$ is an orthogonal projector onto $N(\mathbf{A})$ along $R(\mathbf{A})$. Thus we have $\mathbf{A}\Delta\boldsymbol{\theta}_R^+ = \mathbf{A}(\mathbf{I} - \mathbf{CA})\Delta\boldsymbol{\theta}_U^+ = 0$, because $\mathbf{A}(\mathbf{I} - \mathbf{CA}) = 0$. Next we want to show that prior to convergence, $\mathbf{H} = D^{10}Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_R^+ > 0$. Using similar arguments to those in the proof of Theorem 2, we can write $(\mathbf{I} - \mathbf{CA})\mathbf{I}_U^{-1} = \mathbf{F}(\mathbf{F}'\mathbf{I}_U\mathbf{F})^{-1}\mathbf{F}' = \mathbf{G}$, where $\mathbf{G}$ is a positive semidefinite with rank d and $\mathbf{F}$ is a $p \times d$ $(d = p - Q)$ matrix of rank d such that $\mathbf{AF} = 0$. Because $\mathbf{I}_U$ is a positive definite, $(\mathbf{I} - \mathbf{CA})$ is positive semidefinite with rank d. We will complete the proof that $\mathbf{H} > 0$ by showing that prior to convergence, we have (a) $\Delta\boldsymbol{\theta}_U^{+\prime}\Delta\boldsymbol{\theta}_R^+ > 0$ and (b) $D^{10}Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_U^+ > 0$. Since $(\mathbf{I} - \mathbf{CA})$ is a positive semidefinite, we have $\Delta\boldsymbol{\theta}_U^{+\prime}\Delta\boldsymbol{\theta}_R^+ = \Delta\boldsymbol{\theta}_U^{+\prime}(\mathbf{I} - \mathbf{CA})\Delta\boldsymbol{\theta}_U^+ \geq 0$. Using the same argument as the proof of Theorem 2, the equality gives us $\Delta\boldsymbol{\theta}_U^{+\prime} \in R(\mathbf{A})$, which implies $\Delta\boldsymbol{\theta}_R^+ = (\mathbf{I} - \mathbf{CA})\Delta\boldsymbol{\theta}_U^+ = 0$ indicating $\boldsymbol{\theta}_R^+$ is a stationary point; thus, (a) is proved. To see (b), we apply a Taylor series expansion of $Q(\boldsymbol{\theta}_U^{(+)}|\boldsymbol{\theta}_R^{(m)})$ at $Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})$. We have $Q(\boldsymbol{\theta}_U^{(+)}|\boldsymbol{\theta}_R^{(m)}) = Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)}) + D^{10}Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_U^+ - \frac{1}{2}\Delta\boldsymbol{\theta}_U^{+\prime}\mathbf{I}_U\Delta\boldsymbol{\theta}_U^+ + O(\|\Delta\boldsymbol{\theta}_U^+\|^3)$. Because $\mathbf{I}_U$ is positive definite and $Q(\boldsymbol{\theta}_U^{(+)}|\boldsymbol{\theta}_R^{(m)}) > Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})$, we have $D^{10}Q(\boldsymbol{\theta}_R^{(m)}|\boldsymbol{\theta}_R^{(m)})'\Delta\boldsymbol{\theta}_U^+ > 0$ when $\|\Delta\boldsymbol{\theta}_U^+\|$ is small enough to ignore $O(\|\Delta\boldsymbol{\theta}_U^+\|^3)$.

Proof of Corollary 3

Let $\boldsymbol{\theta}_{U,t}^+ = \boldsymbol{\theta}_R^{(m)} + t\Delta\boldsymbol{\theta}_U^+$, let $0 < t < 1$, and let $\boldsymbol{\theta}_{R,t}^+$ be the restricted solution corresponding to $\boldsymbol{\theta}_{U,t}^+$. Let $\mathbf{I}_U$ and $\mathbf{C}$ be defined in the same way as in the proof of Theorem 3. We assume that $\mathbf{I}_U$ is positive definite. A Taylor series expansion around $t = 0$ of $Q(\boldsymbol{\theta}_{R,t}^+|\boldsymbol{\theta}_R^{(m)})$ at $\boldsymbol{\theta}_R^{(m)}$ gives

$$Q(\boldsymbol{\theta}^{+}_{R,t}\,|\,\boldsymbol{\theta}_R^{(m)}) = Q(\boldsymbol{\theta}_R^{(m)}\,|\,\boldsymbol{\theta}_R^{(m)}) + tD^{10}Q(\boldsymbol{\theta}_R^{(m)}\,|\,\boldsymbol{\theta}_R^{(m)})'(\mathbf{I} - \mathbf{CA})\Delta\boldsymbol{\theta}_U^{+} + O(t^2).$$

From the proof of Theorem 3, prior to convergence we have $D^{10}Q(\boldsymbol{\theta}_R^{(m)}\,|\,\boldsymbol{\theta}_R^{(m)})'(\mathbf{I} - \mathbf{CA})\Delta\boldsymbol{\theta}_U^{+} > 0$. Thus step halving on $\Delta\boldsymbol{\theta}_U^{+}$ is a means of finding a value of $\boldsymbol{\theta}$ that ensures that $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_R^{(m)})$ increases.

[Received January 1993. Revised June 1994.]

REFERENCES

Aitkin, H., Anderson, D., Francis, B., and Hinde, J. (1989), *Statistical Modeling in GLIM,* Oxford, U.K.: Clarendon Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood 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.

Kim, D. K. (1991), "Regression Models for Overdispersed Binomial Data," unpublished Ph.D. dissertation, University of California, Los Angeles, Dept. of Biostatistics.

Lange, K. (1991), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," unpublished manuscript, University of California, Los Angeles, Dept. of Biomathematics.

Little, R. J. A., and Rubin, D. (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.

Meng, X., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association,* 86, 899–909.

Neter, J., Wasserman, W., and Kutner, M. H. (1985), *Applied Linear Statistical Models* (2nd ed.), Homewood, IL: Richard D. Irwin.

Nyquist, H. (1991), "Restricted Estimation of Generalized Linear Models," *Applied Statistics,* 40, 133–141.

Seber, G. A. F. (1984), *Multivariate Observations,* New York: John Wiley.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics,* 11, 95–103.