Journal of Econometrics 49 (1991) 305–341. North-Holland

Extensions of estimation methods using the EM algorithm*

Paul A. Ruud
University of California, Berkeley, CA 94720, USA

Received December 1988, final version received May 1990

The EM algorithm described by Dempster, Laird, and Rubin (1977) is reviewed with the purpose of clarifying several misconceptions in the statistical and econometric literature. The clarifications lead to several applications of the algorithm to models that have appeared to be less tractable. The relationship between the EM algorithm and the method of scoring is also explained, providing estimators of the score and the information from the EM algorithm. The EM algorithm is extended to missing-data problems and an estimation method based on simulations.

1. Introduction

The EM algorithm has proved to be a useful method of computation in econometric and statistical models, particularly since its general structure was explained by Dempster, Laird, and Rubin (1977) (DLR). Due to the pervasiveness of exponential models, the expectation and maximization steps of this algorithm yield tractable and appealing iterative schemes for computing maximum-likelihood estimators in many popular econometric models. Many special cases anticipated the DLR exposition. Indeed, the encompassing character of their discovery was a key contribution. Such cases include econometric models for limited dependent-variable models [Hartley (1977) and Fair (1977)]. Factor analysis [Rubin and Thayer (1982)] and time-series models [Shumway and Stoffer (1982), Watson and Engle (1983)] have seen its use since DLR's paper. Statistical models with mixtures are a natural setting for the EM algorithm, and semi-parametric models with the structure of mixtures have rekindled interest in the EM algorithm [Jewell (1982), Heckman and Singer (1984)].

*The research assistance of Douglas Steigerwald and the comments of Angelo Melino are gratefully acknowledged.

Despite its wide application, the EM algorithm enjoys limited understanding and actual use. Casual summaries of this algorithm describe its expectation step as replacing latent data with its expectation conditional on observed data and guesses of the parameter values; the maximization step is described as estimating the parameters by treating the synthetic latent data as though it were the latent data itself. While special cases do fit this loose, but intuitively appealing, description, many interesting cases do not. Following this introduction, we give a review of the EM algorithm of DLR, emphasizing its limitations and possibilities. In the process, we describe several new applications of the algorithm to limited dependent-variable models popular in econometrics: ordered Probit, multinomial Probit, truncated regression, and nonrandom sample selection.

In addition to its limitations, researchers have not appreciated the full usefulness of the EM algorithm either. It is frequently mislabelled a nonderivative algorithm, and although its simplified maximization step appears to be an advantage over other methods in some respects, this has proven to be a major failing because the algorithm does not seem to provide an estimate of the information matrix. Such derivative-based algorithms as quadratic optimization methods generally compute matrices that approximate the Hessian of the log-likelihood function, and thereby provide estimators of the information matrix as a by-product. The EM algorithm apparently only provides a local critical value of the log-likelihood function. Yet a simple demonstration shows that this is untrue. We will show that the score of the log-likelihood function of the data is readily available in the EM algorithm, so that two estimators of the information matrix are also convenient to compute.

The EM algorithm is also widely recognized as slow to converge, its rate being linear (see DLR). Such quadratic methods as Newton–Raphson converge much faster in the neighborhood of the log-likelihood maximum where the function is well approximated by a quadratic function. Balanced against this slowness is an 'impressive numerical stability' [Haberman (1977)] such that overstepping in areas of the parameter space distant from the likelihood maximum does not occur. Using our observation about the score function, we can make some direct comparisons between the EM and Scoring steps, providing insight into the sluggishness of the former. We also suggest a procedure for using the two algorithms in tandem.

In the final sections of this paper, we suggest that the EM algorithm can be generalized in two important ways. Use of the algorithm has been restricted to models that have complete data sets; that is, no observations can be nonrandomly missing. We give a partial extension of the EM algorithm to such missing data problems, as opposed to the partial observation problems that already use this algorithm. In addition, we show how simulated method-of-moments estimation applies naturally to the EM algorithm.

2. An introduction to the EM algorithm

Consider the problem of maximum-likelihood estimation given the observations on the vector of random variables y drawn from a population with the cumulative distribution function (c.d.f.) $F(\theta; Y) = \mathrm{P}\{y \leq Y\}$. Let the corresponding density function with respect to Lebesgue measure be $f(\theta, y)$. The parameter vector θ is unknown, finite-dimensional, and $\theta \in \Theta$, where Θ is a compact subset of $\mathbb{R}^K$. Estimation of θ by maximum likelihood (ML) involves the maximization of this $f(\theta; y)$, or its logarithm, over Θ. In the limited dependent-variable models we will consider in this paper, F will be a mixture of discrete and continuous distributions so that f may consist of nonzero probabilities for discrete values of y and continuous probability densities for intervals of y.

In general, one can represent the data-generating process for y as an 'incomplete data' or 'partial observability' process in which the observed-data vector y is an indirect observation on a latent-data vector y^* according to the many-to-one mapping $y = \tau(y^*)$. This mapping is often called the 'observation rule' and though it may not be monotonic, τ is generally piece-wise continuous. One must specify a c.d.f. $F(\theta; Y^*)$ for y^* such that

$$F(\theta; Y) = \int_{\{y^*:\, \tau(y^*) \leq Y\}} \mathrm{d}F(\theta; y^*). \tag{2.1}$$

For our purposes, $F(\theta; y^*)$ will take the form

$$F(\theta; y^*) = \int f(\theta; y^*)\, \mathrm{d}y^*,$$

where $f(\theta; y^*)$ is an ordinary continuous probability density function (p.d.f.) $f(\theta; y^*)$. Using this latent structure, the EM algorithm provides an iterative method for maximizing the discrete and continuous mixture $f(\theta; y)$ over values of θ.

The algorithm consists of two steps: the E, or expectation, step and the M, or maximization, step. In the E-step, one takes the expectation of the *log-likelihood function* (not y^*) for the latent data y^* conditional on the observed data y and an initial value for θ, say θ_0. Let $F(\theta; y^*|y)$ denote the conditional c.d.f. of y^* given that $\tau(y^*) = y$:

$$F(\theta; Y^*|y) = \lim_{\varepsilon \to 0^+} \frac{\mathrm{P}\{y^* \leq Y^*, y < \tau(y^*) \leq y + \varepsilon\}}{\mathrm{P}\{y < \tau(y^*) \leq y + \varepsilon\}}. \tag{2.2}$$

This distribution can also be a discrete and continuous mixture. We let

$$\mathrm{E}_0[g(y^*)|y] = \int g(y^*)\,\mathrm{d}F(\theta_0; y^*|y) \tag{2.3}$$

denote the expectation of a function $g(y^*)$ with respect to the conditional c.d.f. $F(\theta; y^*|y)$ of y^* given y when evaluated at a particular value θ_0. Then the expected log-likelihood function is

$$Q(\theta, \theta_0; y) = \mathrm{E}_0[\log f(\theta; y^*)|y]. \tag{2.4}$$

In the M-step, one maximizes $Q(\theta, \theta_0; y)$ with respect to θ, inducing the mapping

$$\theta_1 = M(\theta_0; y) \equiv \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta, \theta_0; y). \tag{2.5}$$

Under conditions described by Boyles (1983) and Wu (1983), the iterative application of this mapping yields a sequence of values for θ which converges to a fixed point of the mapping, which is also a local maximum of $\log f(\theta; y)$. The difference between the $Q(\theta, \theta_0; y)$ that one maximizes and the actual log-likelihood function $\log f(\theta; y)$ is

$$Q(\theta, \theta_0; y) - \log f(\theta; y) = H(\theta, \theta_0; y) \equiv \mathrm{E}_0[\log f(\theta; y^*|y)], \tag{2.6}$$

where $f(\theta; y^*|y)$ is the density associated with $F(\theta; y^*|y)$ in (2.2). The expression in (2.6) is the expected log-likelihood of the latent y^* conditional on $\tau(y^*) = y$. According to the information inequality

$$\int_{\mathscr{A}} [\log f(\theta_0; y^*|y) - \log f(\theta; y^*|y)]\,\mathrm{d}F(\theta_0; y^*|y) \geq 0,$$

if

$$\int_{\mathscr{A}} [\mathrm{d}F(\theta_0; y^*|y) - \mathrm{d}F(\theta; y^*|y)] \geq 0 \tag{2.7}$$

and

$$f(\theta_0; y^*|y) > 0, \qquad \forall y^* \in \mathscr{A},$$

$H(\theta, \theta_0; y)$ is maximized at $\theta = \theta_0$ so that any value of θ that increases Q, $\theta_1 = M(\theta_0)$ in particular, also increases $\log f(\theta; y)$. Following DLR, we label such iterative solution methods generalized EM (GEM) algorithms.

Definition. An iterative algorithm with mapping M is a generalized EM algorithm (a GEM algorithm) if

$$\{Q[M(\theta;y),\theta;y] \geq Q(\theta,\theta;y)\}$$
$$\Rightarrow \{\log f[M(\theta;y);y] \geq \log f(\theta;y)\}, \tag{2.8}$$

for every θ in Θ.

The reason for this generalized form is to emphasize that useful alternatives to the EM algorithm can be formed by replacing the M-step with a step that only increases Q. Such replacements may accelerate the rate of convergence or save computational time per iteration.

For the rest of the paper, it is helpful to introduce one more notational device. We will need to distinguish between two very similar expectations: a *conditional* expectation using θ_0, the initial value, as the parameter value in the distribution function and a *marginal* expectation using θ, the argument of the likelihood function. We have already used E_0 to denote the first kind of expectation [see (2.3)]. We will treat the operator E to mean

$$\mathrm{E}[g(y^*)] = \int g(y^*)\, f(\theta;y^*)\, \mathrm{d}y^*. \tag{2.9}$$

A similar notation will be used for variances.

Exponential distributions

The EM algorithm is especially attractive for the ML estimation of θ for the family of regular exponential likelihood functions given by

$$f(\theta;y^*) = b(y^*)\exp[\theta' t(y^*) - a(\theta)]. \tag{2.10}$$

Many discrete and continuous distributions belong to this family including the multinomial, Poisson, uniform, gamma, and normal. The statistic $\sum_{n=1}^{N} t(y_n)$ is a sufficient statistic for θ given a random sample $\{y_n; n = 1,\ldots,N\}$. When the likelihood of the latent data y^* has this form, then

$$\frac{\partial Q(\theta,\theta_0;y)}{\partial\theta} \equiv Q_1(\theta,\theta_0;y) = \mathrm{E}_0[t(y^*)|y] - \mathrm{E}[t(y^*)], \tag{2.11}$$

and the score for ML estimation of the latent model is similar:

$$\frac{\partial \log f(\theta; y^*)}{\partial \theta} = t(y^*) - \mathrm{E}[t(y^*)]. \tag{2.12}$$

As a result, each step of the EM algorithm corresponds to the maximization of the latent-data log-likelihood function after replacing $t(y^*)$ with $\mathrm{E}_0[t(y^*)|y]$, the expectation of $t(y^*)$ conditional on the available information and an initial value for θ. Frequently this maximization step is simple and familiar because of the simple form of $\mathrm{E}[t(y^*)]$ as a function of θ.

Some authors describe the EM algorithm as replacing the unobserved y^* with its expectation conditional on y and θ_0 and then estimating θ as though this expectation were the actual y^*. According to our description of the EM algorithm, this is accurate when the family of distributions is the exponential and

$$\mathrm{E}[t(y^*)|y] = t\{\mathrm{E}[y^*|y]\}, \tag{2.13}$$

but not otherwise. For the multinomial distribution, condition (2.13) holds and so the EM algorithm takes a simple, intuitive form. Mixture models like switching regressions are another popular application [see Hartley (1977a), Kiefer (1978)]. For the ordinary normal regression model, the sufficient statistics for the slope parameters also satisfy (2.13) so that Probit and Tobit estimation for these coefficients also proceeds this way. But the sufficient statistics for variance parameters in normal regression do not satisfy (2.13), and if one were to attempt Tobit estimation using the EM algorithm without taking this into account, the calculations would not converge to the MLE. We will consider the multivariate normal regression model in more detail below because so many familiar models are generated by it.

Additional comments

It has also been observed that the EM algorithm may be poorly suited to observed likelihood functions that do not exhibit simple sufficient statistics. Our discussion should make clear that sufficient statistics for the *latent* process, rather than the observed process, are the key to a simple algorithm. In some cases, there is a clear connection between the sufficient statistics of the corresponding likelihood functions, but not always. ARMA models are an example [see Ruud (1988)].

In our description of the EM algorithm, we also wish to emphasize a limitation to its application: the transformation τ from latent to observable data cannot depend on parameters to be estimated in such a way that the support of y^* conditional on y depends on θ. If this condition is violated, the

information inequality that establishes the unimodality of H in (2.6) will not apply. If the region of integration depends on θ, then $Q(\theta, \theta_0; y)$ and $\log f(\theta; y)$ are integrals over different regions [compare (2.1) and (2.4)] and their difference no longer has an interpretation as an expected log-likelihood function. As a result, the GEM property will not generally obtain. Such violations occur in econometric models and we will examine such a case below.

Finally, we note that the EM algorithm affords a convenience in parameterizations that is sometimes overlooked: the parameter vector θ need not be identifiable with respect to the likelihood of the observable y for the EM algorithm to work. It can be much easier to overparameterize the likelihood in terms of a 'natural' parameter vector θ than to impose restrictions that are necessary for identifiability. The parameterization of the latent likelihood is not unique. In general, there are many ways, some attractive and some not, to pick a latent likelihood for any given observable likelihood. Some examples of this will be given below, too.

The computational efficiency of the EM algorithm is apparently mixed in practice. Some applications seem to work quite well while others do not. Working in favor of the EM algorithm are that it does not require the calculation of (1) the log-likelihood function, (2) a line search in the parameter space, or (3) the Hessian of the log-likelihood function at each iteration. Working against the EM algorithm are its typical choice of 'safe but stupid' iterations in the parameter values. This last characteristic, however, can be an advantage in problems where quadratic methods tend to overshoot, especially when the starting parameter values are distant from the maximizing values.

In their comparison of the EM and the method of scoring in DYMIMIC models, Watson and Engle (1983) found that the speed of each iteration more than compensated for the extra iterations that the EM algorithm required. They also found the EM algorithm to reach the neighborhood of the maximum quite quickly, but to move slowly in the neighborhood itself. Fair (1977) found that an EM-like algorithm was much faster than scoring for the Tobit model, again because each iteration was so rapid. Hartley (1977a) reported similar success for switching regressions, but Maddala (1983, p. 303) has questioned these results.

3. Applications of the EM algorithm

In econometrics, the EM algorithm can be applied to likelihood functions derived from a latent normal distribution, a special case of the exponential family. The maximization of the likelihood for the latent data is a generalized-least-squares (GLS) calculation, and so, therefore, is each iteration of the EM algorithm.

Let the latent model of J equations with N observations be

$$y_{nj}^* = x_n'\beta_j + u_{nj}, \qquad n = 1,\ldots,N, \quad j = 1,\ldots,J, \tag{3.1}$$

where $u_n = [u_{nj}] \sim N(0_J, \Omega)$, x_n is a vector of K explanatory variables, β_j is a vector of K unknown coefficients, and Ω is an unknown $J \times J$ nonsingular covariance matrix. Let the M-dimensional observable dependent variable be

$$y_{nm} = \tau_m(y_n^*), \qquad m = 1,\ldots,M, \tag{3.2}$$

where $y_n^* = [y_{nj}^*]$. The likelihood function for the latent dependent variable has the seemingly unrelated regression (SUR) form

$$f(B,\Omega; y^*, X) = [\det(2\pi\Omega)]^{-N/2} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left[\Omega^{-1}(Y^* - XB)'(Y^* - XB)\right]\right\}, \tag{3.3}$$

where Y^* is the $N \times J$ matrix $[y_{nj}^*]$, B is a $K \times J$ matrix with β_j in its jth column, and X is an $N \times K$ matrix of all the explanatory variables in the SUR system. Taking $\theta = [\mathrm{vec}(B)', \mathrm{vech}(\Omega^{-1})']'$, sufficient statistics for this likelihood function are

$$t(Y^*) = \begin{bmatrix} X'Y^* \\ \mathrm{vech}(Y^{*\prime}Y^*) \end{bmatrix}, \tag{3.4}$$

conditional on the ancillary statistic $X'X$. The expected log-likelihood (without unnecessary constants) and the expected sufficient statistics are

$$Q(\theta,\theta_0; y) = -\tfrac{1}{2}N\log[\det(\Omega)] - \tfrac{1}{2}\mathrm{tr}\left\{\Omega^{-1}\left[\mathscr{V} + (Y^{*0} - XB)'(Y^{*0} - XB)\right]\right\} \tag{3.5}$$

and

$$\mathrm{E}_0\left[t(Y^*)|Y\right] = \begin{bmatrix} X'Y^{*0} \\ \mathrm{vech}\left[\mathscr{V} + Y^{*0\prime}Y^{*0}\right] \end{bmatrix}, \tag{3.6}$$

where $Y^{*0} \equiv \mathrm{E}_0(Y^*|Y)$ and $\mathscr{V} \equiv \sum_{n=1}^N \mathrm{V}_0(y_n^*|Y)$.

When B is unrestricted, the EM normal equations corresponding to (2.11) for B are

$$X'(Y^{*0} - XB)\Omega_0^{-1} = 0 \Leftrightarrow B = (X'X)^{-1}X'Y^{*0}, \tag{3.7}$$

so that the maximization step for B is not only ordinary least squares (OLS) equation by equation, but also the dependent variable is the conditional expectation of Y^* given Y and the current guess at θ.

The EM normal equations corresponding to Ω^{-1} are

$$\mathscr{V} + (Y^{*0} - XB)'(Y^{*0} - XB) - N \cdot \Omega = 0. \tag{3.8}$$

Thus, the maximization step for the covariance parameters has a simple, intuitive form, but it does not correspond to the maximization of $\log f(\theta; Y^*)$ after replacing Y^* by its conditional expectation, Y^{*0}. This occurs because the restriction in (2.13) is not satisfied when the sufficient statistics include sample moments of second order. Therefore, even though the maximization step for B does not involve conditional expectations of second order, the overall maximization step of the EM algorithm does. Simply replacing Y^* by its conditional expectation and proceeding as though Y^* were observed at each iteration will not implement the EM algorithm.

When there are exclusion restrictions on B, as there often are, then the maximization step for the slope coefficients becomes a GLS calculation, just as it would for the latent model:

$$S'(\Omega_0^{-1} \otimes I)[X'Y^{*0} - X'XB] = 0 \tag{3.9}$$

$$\Leftrightarrow \beta = (Z'\Sigma^{-1}Z)^{-1} Z'\Sigma^{-1} \operatorname{vec}(Y^{*0}), \tag{3.10}$$

where $\beta \equiv S' \operatorname{vec}(B)$ are the unrestricted coefficients, $Z \equiv (I \otimes X)S$ is the matrix of selected regressors, and $\Sigma \equiv (\Omega_0 \otimes I)$ is the covariance matrix of the stacked system of equations.

Just as in SUR, it is not possible to solve the normal equations for both β and Ω in a single iteration when there are exclusion restrictions. Tsur (1983) has pointed out that (3.8) yields a simple solution for Ω if B is taken to be B_0:

$$\Omega = N^{-1}\left[\sum_{n=1}^{N} \mathrm{V}_0(u_n|Y) + \mathrm{E}_0(U|Y)'\mathrm{E}_0(U|Y)\right]. \tag{3.11}$$

Again analogous to SUR computations, EM computations are simplified if one alternates between the M-step for β and the M-step for Ω (which is sometimes called *iterated* GLS), rather than carrying out a single M-step for both simultaneously. Note that the parameter restrictions do not change the basic analogy between the maximization step in the EM algorithm and the maximization of the latent likelihood function shown in (2.11) and (2.12).

We now turn to several applications of the EM algorithm along the lines of (3.10) and (3.11) to illustrate the general points made above.

Multinomial Probit

Note that the expressions for B and Ω in (3.10) and (3.11) can be employed whether all of their elements are identifiable or not. In a multinomial Probit model for example, the transformation from y^* to y takes the form

$$y_j = \left\{ \begin{array}{lll} 1 & \text{if} & y_j^* = \max(y_1^*, \ldots, y_J^*) \\ 0 & \text{if} & \text{otherwise} \end{array} \right\}. \tag{3.12}$$

Neither B nor Ω is identifiable. Because the scale of y^* is not identifiable, B and Ω must be normalized in some way. In addition, each y gives information only about certain contrasts of the elements of y^*,

$$y_i = 1 \Rightarrow y_j^* - y_i^* \geq 0, \qquad j = 1, \ldots, J,$$

so that only certain linear combinations of the elements of B and Ω are identifiable. Unfortunately, a convenient normalization is not available and working with restricted forms is somewhat laborious.

Fortunately, the EM algorithm can be applied to the unrestricted matrices. If the starting value for Ω is nonsingular and Z is full rank, then (3.11) clearly preserves the nonsingularity of Ω and (3.10) will yield values for every element of B. Upon convergence, the EM algorithm stops at a point that locally maximizes the log-likelihood function but that is not locally identifiable. Identifiable functions of the parameters can then be computed.

This example illustrates one role that the latent model plays in the EM algorithm. The latent model determines the direction in the parameter space that the algorithm will move, including situations where identifiability is lacking and many directions are observationally equivalent. The latent model is an alternative to normalizations, which also specify the search direction in such cases: a normalization fixes a subset of parameters.

Ordered Probit

The support of y^* conditional on y must not depend on the parameters to be estimated, but in the latent models of limited dependent-variable models this can happen quite easily. A simple example is the ordered Probit model [see Rosett and Nelson (1975), Maddala (1983) cites McKelvey and Zavoina (1975)]. In this model, one can choose τ to be the step function

$$y = g \quad \text{when } \alpha_g \leq y^* < \alpha_{g+1},$$

where

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_G < \alpha_{G+1} = \infty, \tag{3.13}$$

where α_g, $g = 1, \ldots, G$, are parameters to be estimated. In this case, τ depends on α. Although the ordinary Probit model with $G = 2$ and $\alpha_1 = 0$ can be estimated using the EM algorithm, the ordered Probit generalization cannot be directly implemented. Reparameterization and a different, less intuitive, latent model are required.

Consider a single observation for which $y = k \in [2, \ldots, G-1]$. If

$$y^* \sim N\left[-\alpha_k(\alpha_{k+1} - \alpha_k)^{-1}, (\alpha_{k+1} - \alpha_k)^{-2}\right], \tag{3.14}$$

then

$$\text{Prob}\{y^* \in [0,1)\} = \Phi(\alpha_{k+1}) - \Phi(\alpha_k), \tag{3.15}$$

which is the desired likelihood for this observation so that we may specify τ as the indicator function for the unit interval, which does not depend upon θ. Let us also reparameterize some of the α's in terms of the interval lengths, denoted by λ's, so that

$$\alpha_k = \alpha_1 + \sum_{g=2}^{k} \lambda_g, \qquad k = 2, \ldots, G. \tag{3.16}$$

We also specify

$$y^* \sim N(\alpha_1, 1) \quad \text{if} \quad y = 0, \tag{3.17}$$

$$y^* \sim N(\alpha_G, 1) \quad \text{if} \quad y = G, \tag{3.18}$$

as we would in Ordinary Probit which does not have double truncation. Now the expected log-likelihood function (3.5) is (without constants)

$$\begin{aligned} &\sum_{\{n:\, y_n = 1\}} -\tfrac{1}{2}\left\{\varepsilon_n + \left[y_n^{*0} - \alpha_1\right]^2\right\} \\ &+ \sum_{g=3}^{G} \sum_{\{n:\, y_n = k\}} \log(\lambda_g) - \tfrac{1}{2}\left\{\varepsilon_n \lambda_g^2 + \left[y_n^{*0}\lambda_g + \alpha_1 + \sum_{m=1}^{g-1} \lambda_m\right]^2\right\} \\ &+ \sum_{\{n:\, y_n = G\}} -\tfrac{1}{2}\left\{\varepsilon_n + \left[y_n^{*0} + \alpha_1 + \sum_{m=2}^{k} \lambda_m\right]^2\right\}. \end{aligned} \tag{3.19}$$

The conditional expectations $y_n^{*0} = \mathrm{E}_0(y_n^* | y_n)$ and $\varepsilon_n = \mathrm{V}_0(y_n^* | y_n)$ are given in Johnson and Kotz (1970, pp. 81–83). The M-step for α_1 alone is an OLS

calculation using the dependent variable

$$\begin{Bmatrix} y_n^{*0} & \text{if} \quad y_n = 1 \\ -y_n^{*0}\lambda_k - \sum_{m=1}^{k-1} \lambda_m & \text{otherwise} \end{Bmatrix}, \tag{3.20}$$

where we set $\lambda_{G+1} = 1$.

The M-step for the λ's alone is also fairly simple. It is the recursive solution of the quadratic equations

$$1 - \left[y_n^{*0} \left(\alpha_1 + \sum_{m=1}^{k-1} \lambda_m \right) \right] \lambda_k - \left[\varepsilon_n + \left(y_n^{*0} \right)^2 \right] \lambda_k^2 = 0, \tag{3.21}$$

$$k = 1, \ldots, G,$$

where the larger, positive root is taken each time.

The latent model in this application of the EM algorithm differs from others in several important respects. First of all, the choice of latent model depends on the observed y. Secondly, y^* falls in the same unit interval for observations with values of $y \in \{1, \ldots, G-1\}$. These characteristics underline the fictional character of the latent model in the EM algorithm. We operate as though the distribution for y^* were fixed before y is realized, whereas the opposite actually holds. In cases like ordered Probit, the latent model may have substantive meaning, but that meaning should not be attached to the role of the latent model in the EM algorithm. Coping with the limits of the algorithm may require the use of less intuitive latent models. In fact, one may not wish to proceed with the EM algorithm at all in such cases. Our purpose here is to illustrate its flexibility. A more appealing example appears in Ruud (1988) for the moving-average component of ARMA models.

Nonrandom-sample selection

The nonrandom-sample-selection model is another case where the latent model can be taken too seriously. In this bivariate model ($J = 2$), all sample points are observed, but one dependent variable is partially obscured:

$$y_1 = \mathbf{1}\{y_1^* > 0\}, \qquad y_2 = y_1 \cdot y_2^*. \tag{3.22}$$

It is tempting to try to replace both y_1^* and y_2^* with their conditional expectations in the E-step, as Tsur (1983) did. However, this leads to a

paradoxical situation: the EM algorithm seems to require the values of the regressors for y_2^* when it is not observed, whereas the likelihood for y does not depend on those regressors at all. This paradox is resolved by noting that for observations with $y_1 = 0$, only the *marginal* log-likelihood of y_1^* should appear in the E-step. Although y_2^* is missing nonrandomly, only information on y_1^* is available, namely that it is negative. Thus, the expected log-likelihood for the observations for which $y_{1n} = 1$ looks like (3.5), where

$$y_n^{*0} \equiv \mathrm{E}_0(y_n^* | y_n) = \begin{bmatrix} \mu_{0n} + h_{0n} \\ y_{2n} \end{bmatrix}, \tag{3.23}$$

$$\mu_{0n} \equiv x_{1n}'\beta_{0,1} + (y_{2n} - x_{2n}'\beta_{0,2})\omega_{0,12}/\omega_{0,22},$$

$$h_{0n} \equiv \sigma_0 \frac{\phi(\mu_{0n}/\sigma_0)}{\Phi(\mu_{0n}/\sigma_0)}, \quad \sigma_0 \equiv \sqrt{\omega_{0,11} - \omega_{0,12}^2/\omega_{0,22}},$$

and $\Omega = [\omega_{ij}]$. For the observations with $y_{1n} = 0$, we add

$$-\tfrac{1}{2} \sum_{\{n:\, y_{1n}=0\}} \log(\omega_{11}) + \{\mathrm{V}_0(y_{1n}^* | y_{1n}^* < 0)$$

$$+ [\mathrm{E}_0(y_{1n}^* | y_{1n}^* < 0) - x_{1n}'\beta_1]\}^2/\omega_{11}.$$

The M-step, then, involves bivariate SUR regression computations, but one dependent variable, the replacement for y_1^*, has a longer data series. Upon convergence, the identifiable functions of B and Ω are computed; for example, ω_{11} is normalized to one.

The contrast between this M-step and the Heckman–Lee two-step estimator is interesting [Heckman (1976), Lee (1976)]. The latter requires comparable conditional expectation computations, but the location of the expectations is completely different. The two-step estimator computes

$$\mathrm{E}(u_{2n} | y_{1n} = 1) = \frac{\phi(x_{1n}'\beta_1)}{\Phi(x_{1n}'\beta_1)} \tag{3.24}$$

as a function of an initial consistent estimator for β_1 and includes this as an additional regressor in a second-step, univariate regression for y_2 alone. While the efficiency of the two-step estimator can be improved by taking heteroscedasticity and parameter restrictions into account, this efficiency will never attain that of the MLE owing to its univariate nature.

4. The EM algorithm as a derivative-based method

It is common to refer to the EM algorithm as a derivative-free method because one does not explicitly compute the derivatives of the observed, incomplete-data, log-likelihood function [e.g., Bresnahan (1983), Watson and Engle (1983)]. In a simple sense, the EM algorithm is obviously an algorithm based on derivatives because the score of the expected, complete-data, log-likelihood function is solved in each M-step. However, such calculations do not appear to have a direct relationship with those of the method of scoring, for example. In particular, there has been wide-spread complaint, beginning with the comments on the original DLR paper, that the EM algorithm may find the MLE but it does not provide an estimator of the information matrix. As a result, one must inevitably return to the observed log-likelihood function in the end to compute estimates of the standard errors of the MLE.

On the face of it, this situation is paradoxical: to be able to solve iteratively for the MLE would seem to require solution of the normal equation and, therefore, the calculation of the score. Somehow, the EM algorithm is computing the score of the observed log-likelihood function. In fact, the relationship is quite simple [see also Meilijson (1989)].[1]

Lemma. *If* $f(\theta; y)$ *is differentiable with respect to* θ *and* $\tau(y)$ *does not depend on* θ, *then* $\partial \log f(\theta; y)/\partial\theta = Q_1(\theta, \theta; y)$.

Proof. The result follows from differentiating (2.6) and noting that (2.7) implies that $H_1(\theta, \theta; y) = 0$.

Thus, the EM algorithm calculations provide a direct calculation of the score. Given this lemma, several useful results follow directly. Even though the lemma is well-known, its implications have been generally ignored. First of all, two estimators of the information matrix are immediately available from the EM algorithm and the information matrix is as easy to determine as with the observed likelihood. Secondly, we can explicitly compare the EM algorithm parameter steps with those of such other algorithms as Newton–Raphson and Scoring.

Information-matrix estimators

To simplify our notation let $L(\theta; y) = \log f(\theta; y)$ denote the log-likelihood function of the random variable y for θ. There are three common ways to

[1]Similar results to this lemma, and some of the results that follow in this section, can also be found in Meilijson (1989). This reference came to my attention after the acceptance of this paper. I am thankful to my colleague George Judge for pointing this work out to me.

compute an estimate of the information matrix,

$$\mathscr{I}(\theta) \equiv N^{-1} \sum_{n=1}^{N} \mathrm{E}\left[L_1(\theta; y_n) L_1(\theta; y_n)' \right]. \tag{4.1}$$

The most obvious is to evaluate the information matrix at an estimate of θ:

$$\hat{\mathscr{I}}_1 = \mathscr{I}(\hat{\theta}). \tag{4.2}$$

An alternative is to dispense with the expectations operator and compute

$$\hat{\mathscr{I}}_2 = N^{-1} \sum_{n=1}^{N} L_1(\hat{\theta}; y_n) L_1(\hat{\theta}; y_n)'. \tag{4.3}$$

A third choice relies on the identity $\mathrm{E}(L_1 L_1') = -\mathrm{E}(L_{11})$ for regular likelihoods:

$$\hat{\mathscr{I}}_3 = -N^{-1} L_{11}(\hat{\theta}; y); \tag{4.4}$$

it is sometimes called the observed information [Efron and Hinkley (1978)]. Rothenberg (1981) has proved the following result about these estimators.

Theorem 1. $V[N^{1/2} \operatorname{vech}(\hat{\mathscr{I}}_1)] \leq V[N^{1/2} \operatorname{vech}(\hat{\mathscr{I}}_2)]$, $V[N^{1/2} \operatorname{vech}(\hat{\mathscr{I}}_3)]$, *where* $V(\cdot)$ *stands for the covariance matrix of the limiting multivariate normal distribution of the statistic as sample size* $N \to \infty$.

Therefore, on efficiency grounds, one has a distinct preference among the estimators. Note that this efficiency ordering does not imply a preference among test statistics using these various estimators.

Now when we are using the EM algorithm, we can calculate the second estimator of the information matrix quite simply as

$$\hat{\mathscr{I}}_2 = N^{-1} \sum_{n=1}^{N} Q_1(\hat{\theta}, \hat{\theta}; y_n) Q_1(\hat{\theta}, \hat{\theta}; y_n)'. \tag{4.5}$$

Louis (1982) derived this estimator for the special case of the multinomial distribution, apparently missing its broader applicability. Meilijson (1989) also makes this observation.

Alternatively, one can differentiate $Q_1(\theta, \theta; y)$, perhaps numerically, and compute

$$\hat{\mathscr{I}}_3 = -N^{-1} \partial Q_1(\theta, \theta; y) / \partial \theta \big|_{\theta = \hat{\theta}}. \tag{4.6}$$

Louis (1982) provided another form for this estimator:

$$\hat{\mathscr{I}}_3 = -\mathrm{E}\big[L_{11}(\hat{\theta}; y^*)|y, \hat{\theta}\big] + \mathrm{E}\big[L_1(\hat{\theta}; y^*)L_1(\hat{\theta}; y^*)'|y, \hat{\theta}\big], \tag{4.7}$$

which is intuitively appealing, but less convenient because it requires additional expectations. See also Meilijson (1989).

It is now obvious that the information matrix itself is as easy to find in the EM setting as in the observed-likelihood setting. To find its mathematical form one takes the expectation of either $\hat{\mathscr{I}}_2$ or $\hat{\mathscr{I}}_3$ for fixed θ:

$$\mathscr{I}(\theta) = N^{-1} \sum_{n=1}^{N} \int Q_1(\theta, \theta; y_n) Q_1(\theta, \theta; y_n)' f(\theta; y_n)\, \mathrm{d}y_n \tag{4.8}$$

is probably easiest.

Given these estimators, several additional features can easily be added to the EM algorithm in standard ways. First of all, it is convenient to calculate LM (or score tests) of such maintained hypotheses as excluded regressors and homoscedasticity. Secondly, we can discover nonidentifiable parameterizations that cause numerically singular information estimates, even though we can employ nonidentifiable structures in the EM algorithm itself. Thirdly, there are several convenient measures of convergence for the EM algorithm. Rather than judge convergence by the length of the parameter changes, one should measure the length of the gradient by a measure of curvature; for example,

$$\Delta = Q_1(\hat{\theta}, \hat{\theta}; y)' \hat{\mathscr{I}}_3^{-1} Q_1(\hat{\theta}, \hat{\theta}; y). \tag{4.9}$$

This quadratic form equals the directional derivative of the observed likelihood function along the search direction of the Newton–Raphson algorithm. Such measures are particularly important because the EM algorithm has a well-deserved reputation for converging very slowly in the neighborhood of the critical values of the log-likelihood function.

A comparison of the EM and quadratic algorithms

Because we know the relationship between the EM score and the observed score, we can directly compare the parameter steps of the EM algorithm and quadratic methods and gain insight into the relative speed of the latter in a neighborhood of the maximum. Eq. (2.5) and differentiability allow us to write the EM algorithm in a form reminiscent of such quadratic procedures as Newton–Raphson (NR). Suppose Q_{11} $(= \partial^2 Q(\theta, \theta_0; y)/\partial\theta\partial\theta')$ and L_{11} are nonsingular. Then

$$\theta_{\mathrm{EM}} = \theta_0 - Q_{11}^{-1} Q_1 + \mathrm{o}(\|\theta_{\mathrm{EM}} - \theta_0\|), \tag{4.10}$$

where Q_{11} and Q_1 are evaluated at $\theta = \theta_0$. This can be compared with the simplest form of NR which computes

$$\theta_{NR} = \theta_0 - [L_{11}(\theta_0; y)]^{-1} L_1(\theta_0; y).$$

According to our lemma, $L_1 = Q_1$, so that by further differentiation:

$$L_{11} = Q_{11} + Q_{12}, \tag{4.11}$$

and we can also write

$$\theta_{NR} = \theta_0 - (Q_{11} + Q_{12})^{-1} Q_1. \tag{4.12}$$

To a first-order approximation, the difference between EM and NR is the matrix which scales the score vector Q_1. By differentiating (2.6), we see that

$$Q_{12} = -H_{11}, \tag{4.13}$$

which is the negative of the information matrix for the conditional distribution $F(\theta; y^*|y)$. Therefore, although it is clearly a gradient method, the EM algorithm fails to use the Hessian of the log-likelihood function like NR; it substitutes a matrix that differs from the Hessian by a negative semi-definite matrix that measures the information loss due to partial observability.

Intuition suggests that this explains the slow rates of convergence exhibited by EM. In certain cases, it does follow from (16) and (17) that there is a neighborhood of the MLE in which the EM algorithm improves the log-likelihood function less than the NR algorithm. We use the following definition [Rothenberg (1971)]:

Definition. Let $M(\theta)$ be a matrix whose elements are continuous functions of θ everywhere in an open subset Θ. The point $\hat{\theta} \in \Theta$ is said to be a regular point of the matrix if there exists an open neighborhood of $\hat{\theta}$ in which $M(\theta)$ has constant rank.

Theorem 2. If the MLE $\hat{\theta}$ is a regular point of $L_{11}(\theta; y)$ and $L_{11}(\hat{\theta}; y)$ is nonsingular, then there is an open neighborhood of $\hat{\theta}$ such that

$$L(\theta_0; y) < L(\theta_{EM}; y) < L(\theta_{NR}; y).$$

A proof is given in the appendix [see Meilijson (1989) for a similar result]. Although NR takes faster steps than EM toward the MLE in its neighborhood, experience shows that EM often increases the log-likelihood function

more than NR outside such small neighborhoods. As a result, EM is often superior to NR at the outset of iterative numerical optimization because each iteration takes less time and increases the log-likelihood function more.

Within the exponential family of distributions for y^*, we can make comparisons between the EM algorithm and the method of scoring that yield similar results. If the distribution of y^* has a probability density function of the form (2.10), then $Q_{11}(\theta, \theta; y) = -\partial^2 a(\theta)/\partial\theta\partial\theta'$ does not depend on y and, therefore, equals the negative of the information of the latent marginal log-likelihood function. Taking the expectation over values of y, (4.11) becomes

$$\mathscr{I}(\theta) = -Q_{11}(\theta) - \mathscr{H}(\theta), \tag{4.14}$$

where $\mathscr{I}(\theta)$ is the information for θ and $\mathscr{H}(\theta)$ is a symmetric, positive, semi-definite matrix. Using the same argument that supports the previous theorem, we have:

Theorem 3. One iteration of the method of scoring is given by

$$\theta_S = \theta_0 + \mathscr{I}^{-1}Q_1 + o(\|\theta_S - \theta_0\|).$$

If the latent likelihood has the exponential form (2.10), the MLE $\hat{\theta}$ is a regular point of L_{11}, and L_{11} is nonsingular, then there is an open neighborhood of $\hat{\theta}$ such that

$$L(\theta_0; y) < L(\theta_{EM}; y) < L(\theta_S; y).$$

Intuitively speaking, the method of scoring performs a GLS calculation which takes the heteroscedasticity of the gradient terms into account. A statistical consequence of this difference is that one step of the method of scoring from a consistent estimator is asymptotically equivalent to the MLE, but one step of the EM algorithm is relatively inefficient. This relative statistical inefficiency in a neighborhood of the MLE parallels a computational inefficiency. In actual iterations, the method of scoring converges faster because it is maximizing a quadratic form that approximates the log-likelihood function. Indeed, (4.11) and (4.12) can be used to show that Louis' (1982) method for speeding up the EM algorithm in the neighborhood of the MLE is a version of the method of scoring.

The methods are similar, however, in the sense that they move into the same half space from any point in the parameter space. It appears, therefore, that if the EM algorithm avoids singularities in the likelihood function better than the method of scoring, as some practitioners report, this is due only to the tendency of the method of scoring to overshoot into numerically prob-

lematic parts of the parameter space. If this overshooting is corrected by the appropriate choice of a step length, then the two algorithms will converge to the same critical values from any starting point.

One need not be restricted to one algorithm or the other. Because the score can be computed so easily from the EM computations, the BHHH algorithm [Berndt et al. (1974)] is a particularly convenient quadratic method to use along with the EM algorithm. Using $\hat{\mathscr{I}}_2$ to estimate the information-matrix term, one effectively replaces $a_{\theta\theta} = \mathrm{V}[t(y^*)]$ with White's heteroskedastic-consistent estimator of $\mathrm{V}\{\mathrm{E}[t(y^*)|y]\}$. Since the EM algorithm often performs better in the first iterations, Watson and Engle (1983) suggest that a promising approach is to combine the algorithms by beginning with the EM algorithm and then switching to the BHHH algorithm when the EM approaches the MLE and begins to slow down. The convergence criterion that we mentioned above can serve as an indicator for switching, after replacing $\hat{\mathscr{I}}_3$ with $\hat{\mathscr{I}}_2$. Alternatively, a smooth combination of the algorithms may automate this switching process well:

$$\gamma = \gamma_0 + \left(\mathrm{e}^{-\Delta}\gamma_S + \mathrm{e}^{-\Lambda}\gamma_{\mathrm{EM}}\right)/\left(\mathrm{e}^{-\Delta} + \mathrm{e}^{-\Lambda}\right), \tag{4.15}$$

where Λ is a constant chosen by the practitioner.

5. Missing data

The EM algorithm does not apply to problems in which data is missing nonrandomly, as in cases where the observations are missing 'because of the values that would have been observed' (DLR, p. 11). Yet similar regression forms appear in the first-order conditions of such MLE's as that for the truncated normal regression model. Let the latent model remain (3.1) but let the observable data be

$$y_n = y_n^* \quad \text{if} \quad y_n^* \in \mathscr{B}, \quad \text{but } y \text{ is unobserved otherwise,} \tag{5.1}$$

instead of (3.2). There are no 'partially observed' y^*'s to replace by conditional expectations. In this case, the likelihood function is

$$f(B, \Omega; y, X) = \begin{cases} P(\mathscr{B})^{-1} f(B, \Omega; y^*, X), & y = y^* \in \mathscr{B} \\ 0, & y^* \notin \mathscr{B} \end{cases}, \tag{5.2}$$

where

$$P(\mathscr{B}) = \int_{\mathscr{B}} f(B, \Omega; y^*, X)\, \mathrm{d}y^*, \tag{5.3}$$

where $f(B, \Omega; y^*, X)$ is given in (3.3).

The normal equations are

$$X'[U - \mathrm{E}(U|\mathscr{B})] = 0 \tag{5.4}$$

and

$$N \cdot \Omega^{-1} - U'U - [\mathrm{E}(N \cdot \Omega - U'U|\mathscr{B}, X)] = 0, \tag{5.5}$$

when B and Ω are unrestricted and where $U \equiv Y - XB$. These normal equations are strikingly similar to (3.7) and (3.8). It appears that in the case of B, we should replace y with $y - \mathrm{E}_0(u|\mathscr{B})$ in order to avoid nonlinearities in the solution; this is in contrast to replacing the unobserved y_n^* with its conditional expectation $\mathrm{E}_0(y_n^*|y)$. For Ω the same logic suggests replacing $u_n u_n'$ with $u_n u_n' + \mathrm{E}_0(\Omega_0 - u_n u_n'|\mathscr{B})$.

Note that the log-likelihood function is the difference between the log-likelihood functions for two partially observed data problems:

$$y_1 = y^* \cdot \mathbf{1}\{y^* \in \mathscr{B}\} \quad \text{and} \quad y_2 = \mathbf{1}\{y^* \in \mathscr{B}\} \tag{5.6}$$

so that

$$\log f(\theta; y) = \log f(\theta; y_1) - \log f(\theta; y_2). \tag{5.7}$$

Since the EM algorithm works for each of the right-hand log-likelihood functions, one might hope that there is a counterpart for the left-hand log-likelihood function. We can find an analogy for the Q function of the EM algorithm implicit in the substitutions just suggested by integrating back:

$$\begin{aligned} Q(\theta, \theta_0; y) = {} & \mathrm{E}_0[\log f(\theta; y^*)|y_1] - \mathrm{E}_0[\log f(\theta; y^*)|y_2] \\ & + \mathrm{E}_0[\log f(\theta; y^*)]. \end{aligned} \tag{5.8}$$

Differentiation confirms that in general, not only for our truncated regression problem, this Q has the derivative property $Q_1(\theta, \theta; y) = L_1(\theta; y)$ possessed by the EM algorithm.

Unfortunately, we have not been able to establish also the GEM property for the general case. The corresponding H function given by (2.5) does not possess the concavity of an expected log-likelihood function because it involves the difference of such functions. In the appendix, we show by counterexample that H is not always maximized at the initial value θ_0. But this is not a necessary condition for the GEM property to hold. Indeed, we also show that the GEM property holds for our counterexample, the truncated normal location model.

The missing-data and partially-observed-data cases are easily combined. If we let

$$y = \tau(y^*), \tag{5.9}$$

$$\mathscr{A}(y) \equiv \{y^* | y = \tau(y^*)\},$$

$$\mathscr{B} \equiv \bigcup_y \mathscr{A}(y),$$

so that the log-likelihood of y is given by

$$\log f(\theta; y) = \log \int_{\mathscr{A}(y)} f(y^*, \theta)\, \mathrm{d}y^* - \log \int_{\mathscr{B}} f(y^*, \theta)\, \mathrm{d}y^*, \tag{5.10}$$

the proposed Q function is

$$Q(\theta, \theta_0; y) = Q(\theta, \theta_0; y_1) - Q(\theta, \theta_0; y_2) + Q(\theta, \theta_0), \tag{5.11}$$

where $y_1 \equiv [y, \mathbf{1}\{y^* \in \mathscr{B}\}]$, $y_2 \equiv \mathbf{1}\{y^* \in \mathscr{B}\}$, and $Q(\theta, \theta_0) \equiv \mathrm{E}_0[\log f(\theta; y^*)]$. Of course, the derivative property continues to hold.

For regular exponential likelihood functions, we have the EM normal equations

$$\mathrm{E}_0[t(y^*)|y] - \mathrm{E}_0[t(y^*)|\mathscr{B}] + \mathrm{E}_0[t(y^*)] - \mathrm{E}[t(y^*)] = 0, \tag{5.12}$$

so that we can interpret each step of the extended EM algorithm as the maximization of the latent-data log-likelihood after replacing $t(y^*)$ with its expectation conditional on the observed information minus its expectation conditional on truncation plus its marginal expectation, all evaluated at an initial value for θ. Using (4.10) and (4.11), we can show that the EM iteration will be

$$\begin{aligned} \gamma_{\mathrm{EM}} &= \gamma_0 + N^{-1} a_{\theta\theta}(\theta_0)\left[L_\gamma(\theta_0; y_1) - L_\gamma(\theta_0; y_2)\right] \\ &= \gamma_0 + N^{-1} a_{\theta\theta}(\theta_0) L_\gamma(\theta_0; y). \end{aligned} \tag{5.13}$$

Therefore, the search direction in exponential models points in the uphill direction and retains its relationship to the direction of the method of scoring.

6. The simulated EM algorithm

Recently McFadden (1989) proposed a method of estimation called simulated-moments estimation. Pakes and Pollard (1989) make an independent, similar contribution. Rather than compute exact moments for a method-of-moments estimator, McFadden suggests computing simulations that are unbiased estimators of the moments. This alternative is attractive when the exact moments are relatively expensive to compute compared with unbiased simulations. Several useful asymptotic properties of the estimators are preserved under this substitution of simulations for moments: consistency and asymptotic normality. Typically there is a loss in efficiency caused by the noise in the simulations.

The EM algorithm is a natural setting for this simulation methodology. Indeed, Bresnahan (1981) has already used this approach. Replacing the expectation step with Monte Carlo simulation makes a close theoretical link between such simulated-moments estimators as the Multinomial Probit estimator of McFadden and the maximum-likelihood estimator. It also provides a direct motivation for such estimators in many other settings. There are practical advantages also. For simulations that are discontinuous functions of the parameters, the M-step of the EM algorithm provides a search direction which does not require differentiation of the simulations. Large sample sizes provide enough smoothness so that the M-step will improve the parameter estimates at each iteration. The EM version of simulated-moments estimation also provides a simplification of McFadden's estimator. Whereas the latter must simulate the efficient instruments, these instruments are observed variables in the EM form. This computational advantage can translate into a statistical one because simulated instrumental variables are not efficient instrumental variables.

These advantages are often balanced by the following disadvantage: the simulations required by the EM algorithm may be as difficult to compute as the likelihood function itself. McFadden uses simulations drawn from the marginal distribution of the latent variable y^*, which are inexpensive to compute for exponential models. In contrast, the EM algorithm generally uses random draws from the conditional distribution of y^* given y. Such distributions involve truncation which can complicate simulation significantly. After describing the general approach to simulation, we will illustrate these points with the Multinomial Probit problem.

Suppose that the E-step of the EM algorithm is replaced by an S-, or simulation, step, creating the simulated EM (SEM) algorithm. That is, instead of analytically finding $Q(\theta, \theta_0; y)$, one calculates an unbiased estimator. One might draw one or more samples from the conditional distribution of y^* given y and the parameter values θ_0; let the M samples be $\{\hat{y}_1^*(\theta_0), \ldots, \hat{y}_M^*(\theta_0)\}$. A simulated version of Q defined in expression (2.3) is

the unbiased simulation

$$\hat{Q}(\theta, \theta_0; y) = M^{-1} \sum_{m=1}^{M} \log f\left[\theta; \hat{y}_m^*(\theta_0)\right]. \tag{6.1}$$

As the sample size approaches infinity, we can expect that under certain regularity conditions

$$\text{plim } N^{-1}\left[\hat{Q}(\theta, \theta_0; y) - Q(\theta, \theta_0; y)\right] = 0, \tag{6.2}$$

according to a law of large numbers. As a result, the SEM algorithm estimator shares some of the properties of the ML estimator, which depends on Q.

A more direct route to a simulated ML estimator would be to simulate the log-likelihood function of y directly. Note, however, that simulation cannot provide an unbiased predictor of

$$L(\theta, y) = \log\left[\int_{A(y)} f(\theta; y^*)\, \mathrm{d}y^*\right], \tag{6.3}$$

because of the logarithmic transformation. This difficulty is a motivation for the method-of-moments approach to estimators based on simulations. The simulation step of the SEM algorithm exploits a transformation of $\log f(\theta; y)$ that provides a direct route from the likelihood function, when it is available, to an estimator employing simulations.

In the simplest case, $\hat{y}^*(\theta)$ is differentiable and the likelihood function is unimodal. Then the estimator $\hat{\theta}$,

$$\hat{Q}_1(\hat{\theta}, \hat{\theta}; y) = 0, \tag{6.4}$$

so that under suitable regularity conditions,

$$\text{plim } \hat{\theta} = \theta^*, \tag{6.5}$$

$$\text{plim } N^{-1}\left[\hat{Q}_{11}(\hat{\theta}, \hat{\theta}; y) + \hat{Q}_{12}(\hat{\theta}, \hat{\theta}; y)\right] = -\mathscr{I},$$

and

$$N^{1/2}(\hat{\theta} - \theta^*) \xrightarrow{\mathscr{L}} N[0, \mathscr{I}^{-1} V \mathscr{I}^{-1}],$$

where θ^* is the value of θ in the population-data-generating process, V is the asymptotic variance of $N^{-1/2}\hat{Q}_1(\theta^*, \theta^*; y)$, and $\mathscr{I}$ is the information matrix evaluated at θ^*. For the exponential of section 2, the limiting variance

reduces to

$$\mathscr{I}^{-1}V\mathscr{I}^{-1} = \mathscr{I}^{-1} + \mathscr{I}^{-1}\{a_{\theta\theta} - \mathscr{I}\}\mathscr{I}^{-1}/M,$$

because (4.13) and (4.14) imply

$$\mathrm{E}\{\mathrm{V}[t(y^*)|y]\} = a_{\theta\theta} - \mathscr{I}. \tag{6.6}$$

As a result, one sees that the rate with which the inefficiency due to simulation falls as the number of simulations M increases depends on the loss of information in observing y rather than y^*.

Let us reconsider the Multinomial Probit model defined by (3.1) and (3.12). McFadden uses this as an example and it gives concrete illustrations of the trade-offs between the SEM algorithm and McFadden's alternative. McFadden notes that the score for β can be written as

$$\partial L(\theta; y)/\partial\beta = \sum_{n=1}^{N}\sum_{j=1}^{J} w_{nj}\left[y_{nj} - \mathrm{P}(j|\theta, X_n)\right], \tag{6.7}$$

where

$$w_{nj} = \partial \log \mathrm{P}(j|\theta, X_n)/\partial\beta, \tag{6.8}$$

and $\mathrm{P}(j|\theta, X)$ is the probability that $y_j = 1$ $(j = 1, \ldots, J)$. Although $\mathrm{P}(j|\theta, X)$ is difficult to compute for high-dimensional choice problems, it is quite easy to estimate by Bernoulli Monte Carlo experiments based on the marginal distribution for y^*.[2] There are several quick algorithms for computing independent standard normal pseudo-random variables. If one factors $\Omega = P'P$, then

$$\hat{Y}_m^*(\theta) = XB + Z_m P, \qquad m = 1, \ldots, M, \tag{6.9}$$

provides the mth simulation of y^* as a function of an $N \times J$ matrix of such pseudo-random variables Z_m. The observed data are simulated using (3.12) to compute $\hat{Y}_m(\theta)$ as a function of $\hat{Y}_m^*(\theta)$. Replacing each probability $\mathrm{P}(j|\theta, X_n)$ by the sample average of the simulations of y_{nj} yields a simulated-moment equation that is an unbiased substitute for the original score.

[2]McFadden also suggests other smooth, unbiased estimators and these are also based upon draws from unconditional distributions.

In contrast the EM algorithm writes the score as

$$\partial L(\theta; y)/\partial\beta = \sum_{n=1}^{N}\sum_{j=1}^{J} x_{jn}\left[\mathrm{E}\left(y_{jn}^{*}|y_n, \theta, X_n\right) - x_{jn}'\beta\right], \tag{6.10}$$

and the SEM algorithm would require Monte Carlo draws from the conditional distribution of y^* given y. That is, y^* would be drawn from the truncated multivariate normal distribution $N(x_n'B, \Omega)$ with support $\{y_{nj}^* \leq y_{nk}^*, j = 1, \ldots, J\}$ given that $y_{nk} = 1$. As far as we know, such simulations require rejection methods that make draws until a suitable simulation is obtained. The crudest example of this would be to draw from the nontruncated distribution of y_n^* (as above) until $\hat{y}_{nk} = 1$. A more sophisticated approach uses importance sampling. In any case, simulation is significantly more difficult.

An alternative approach is to abandon the method of independent simulation for each observation and to use one sample of untruncated simulations to estimate the truncated expectations for each observation. In our example, if $\{z_n;\ n = 1, \ldots, N\}$ is a simulated sample of N J-dimensional standard normal pseudo-random variables, then an unbiased predictor of $\mathrm{E}(y_n^*|y_n, \theta, X_n)$ is

$$\hat{y}_n^* = X_n\beta + \hat{u}_n(\beta), \tag{6.11}$$

where

$$\hat{u}_n(\beta) = \frac{\sum_{m=1}^{N} \mathbf{1}\left[z_m \in \mathscr{A}_n(\beta)\right] P'z_m}{\sum_{m=1}^{N} \mathbf{1}\left[z_m \in \mathscr{A}_n(\beta)\right]}$$

and

$$\mathscr{A}_n(\beta) = \left\{z | u = P'z, x_{nj}'\beta + u_j \leq x_{nk}'\beta + u_k, j = 1, \ldots, J\right\},$$

again given that $y_{nk} = 1$. In this way, one computes one simulation per observation and avoids the difficulties of rejection methods, but with the effect of introducing correlation among the predictors across observations. While it is possible, that no simulation would fall in the set $\mathscr{A}_n(\beta)$ is an improbable event such that the effects of any resampling are generally negligible.

The differences between the two simulated scores are important. McFadden's form requires the additional simulation of the instruments w, and because these will be estimated with noise, there will be a loss in

efficiency. The SEM form uses the instruments x, which are observed variables. The SEM simulations are more difficult, however, because they require draws from the conditional distribution of y^* given y, which is a truncated distribution. The SEM algorithm can be extended to missing data problems, as described in section 4. No additional complications arise. Indeed, the need to draw from truncated distributions seems unavoidable in this case.

7. An illustration

In conclusion, we offer a simple illustration of the extensions of the EM algorithm offered here using Fair's (1977) data on extramarital affairs. Fair proposed a method of computing the MLE for the normal censored regression, or Tobit, model that is similar to the EM and showed its computational gains over a more common procedure. After summarizing our computations for the censored case, we also apply the estimation methods described in the previous two sections to his data. In our comparisons of the EM algorithm versus various scoring algorithms, a familiar pattern emerges: the EM algorithm is a superior method in the early iterations of maximization but quadratic methods are computationally more efficient near the maximum. This continues to hold in the missing-data case in which we ignore censored data. We also apply the SEM estimator to the censored regression model and find no significant bias for the MLE.

First of all, we will compare the computational speeds of the EM algorithm with competing algorithms for estimating the parameters of a censored regression model. Within the notation of section 3, the latent normal regression model consists of a single latent normal regression equation ($J = 1$) and the observation rule is

$$y_n = \tau(y_n^*) = \mathbf{1}\{y_n^* \geq 0\} \cdot y_n^*, \qquad n = 1, \ldots, N,$$

so that only the positive observations of the dependent variable are observed, but no observations are missing. Let the homoscedastic variance of u_n be σ^2. In terms of the natural parameters (β, σ^2), the log-likelihood function has a unique maximum but exhibits nonconcavity away from the maximum. Thus, a scoring algorithm like (4.12) cannot use the negative of the Hessian matrix as an estimator of the information matrix. Various approaches are available; a particularly convenient choice is to use the estimator in (4.4) because it can be computed from the first derivatives alone. Among econometricians this method is commonly called the BHHH algorithm, after Berndt, Hall, Hall, and Hausman (1974).

Table 1

Parameter		Value	Standard error
β_1	Constant	7.6085	3.9060
β_2	Sex dummy (1 = male)	0.94579	1.0629
β_3	Age	−0.19270	0.08097
β_4	Number of years married	0.53319	0.14661
β_5	Number of children	1.0192	1.2796
β_6	Index of religiosity	−1.6990	0.40548
β_7	Index of education	0.025361	0.22767
β_8	Index of occupation	0.21298	0.32116
β_9	Index of marital happiness	−2.2733	0.41541
σ		8.2584	0.55458

Even for such simple problems as the normal regression model, the BHHH algorithm is known to be exceedingly slow when it is applied in this way, to all of the parameters simultaneously. A popular modification is to alternate between the maximization of the log-likelihood over the slope coefficients β and the variance parameter σ^2. This same strategy was advocated for the EM algorithm in (3.10) and (3.11). The method is called *iterated* generalized least squares when it is applied to the general linear model. We will also apply this improvement to the scoring calculations and call the algorithm iterated Newton–Raphson since the Hessian of the log-likelihood is used in place of the information matrix. Fair's method also alternates between the regression and variance parameters in this way.

Alternatively, Olsen (1978) has shown that that the log-likelihood is globally concave in terms of the reparameterization ($\delta = \beta/\sigma$, $\gamma = 1/\sigma$). This is a special circumstance within the family of limited dependent-variable models considered in section 3.1, but Newton–Raphson based on this parameterization is much faster than that in terms of (β, σ^2). Fair (1977) proposes another algorithm especially for censored regression that we will also compute.

Our comparison of these algorithms is made with the first test problem provided in Fair (1977). It consists of nine explanatory variables and 601 observations, 150 nonzero and 451 zero. The dependent variable is a measure of the time spent by individuals in extramarital affairs. The sample comes from a survey in 1969 by *Psychology Today* and is described in more detail in Fair (1978). The maximum of the log-likelihood function is −704.731 and occurs at the values in table 1, which agree with Fair's table 3.[3] The standard errors were computed using (4.3). All of the algorithms were started at the same point in the parameter space: $\beta = 0$ and $\sigma = 1$.

[3]In his table 3, Fair apparently mislabelled the parameter σ with σ^2.

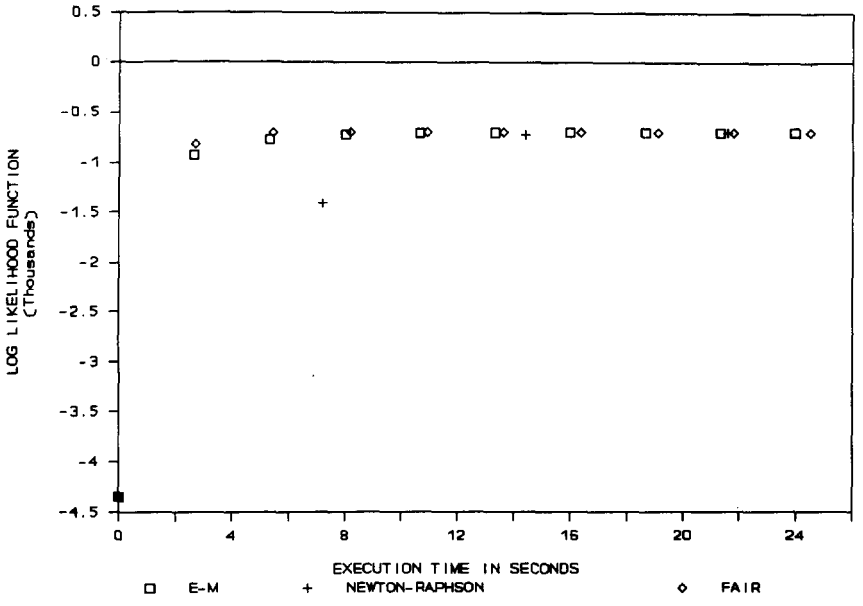

Fig. 1. Globally concave parameterization of Tobit.

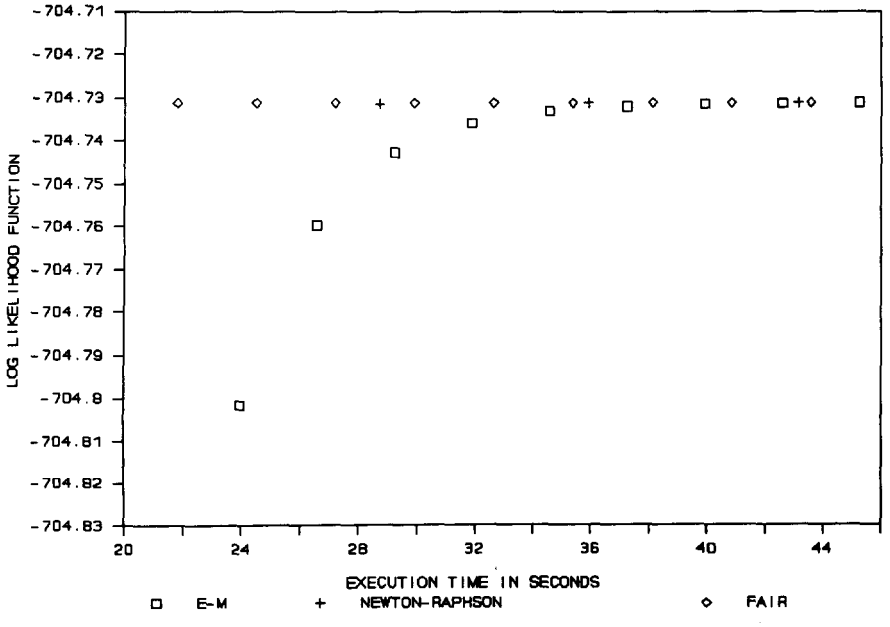

Fig. 2. Globally concave parameterization of Tobit.

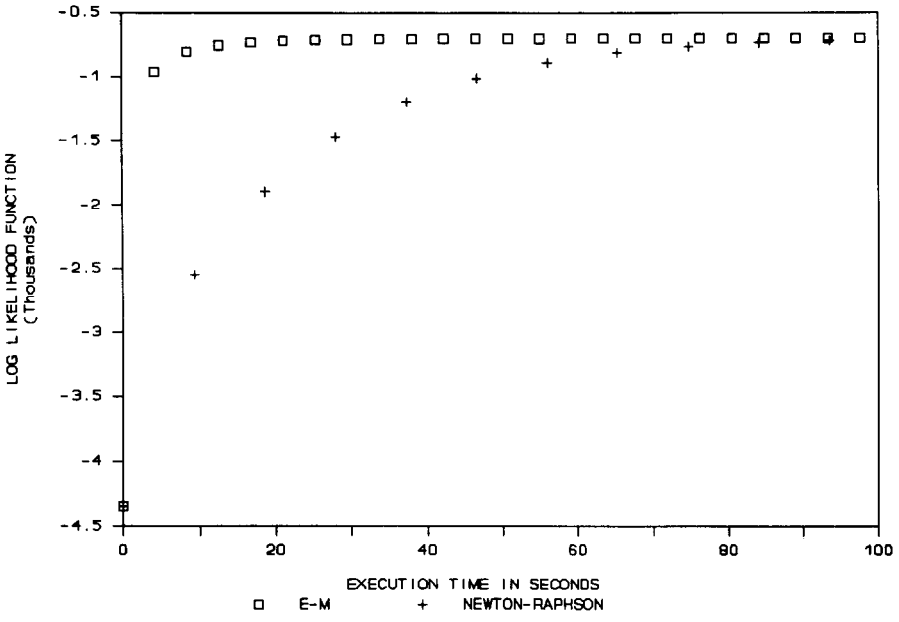

Fig. 3. Comparison with iterated Newton–Raphson.

Figs. 1 to 6 are graphs of the convergence of the various algorithms.[4] Each point is the result of an iteration of the indicated algorithm so that slower algorithms have points spaced farther apart. Figs. 1 and 2 show the speediest algorithms against the EM algorithm. Fair's algorithm is unambiguously the best, but one should note that this speed depends upon his optimal choice of a damping parameter. The EM algorithm clearly reaches the neighborhood of the maximum much faster than the Newton–Raphson algorithm, but by the fourth iteration the latter has overtaken the EM and actually converges more quickly in the end. This behavior was explained in section 4, where we also advocated a combination of the algorithms with the EM beginning and Newton–Raphson completing the iterative process.

Figs. 3 and 4 show a more representative situation for such limited dependent-variable models, in which special algorithms and parameterizations are not employed. Instead, one iterates between the β and σ parameters. The EM algorithm appears to enjoy a greater advantage in this case, but it still exhibits difficulties in the neighborhood of the maximum. Figs. 5 and 6 give a particularly dramatic illustration of the benefits of switching algo-

[4]The computations were made on an IBM PC AT with an 8087 coprocessor using Fortran program compiled with the Microsoft Fortran Compiler. Fortran code for matrix inversion and pseudo-random number generation came from the IMSL. The data and the Fortran code are available from the author upon request.

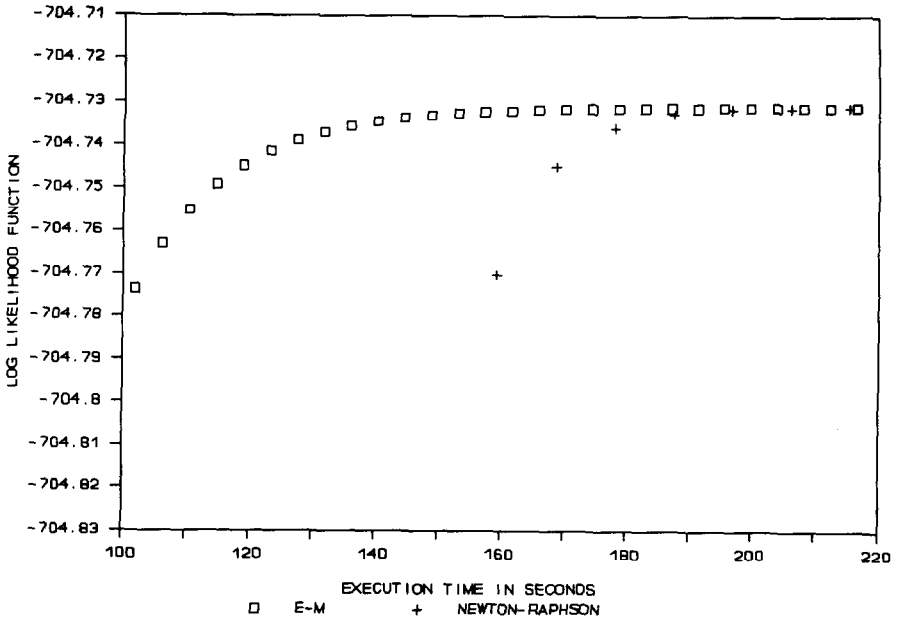

Fig. 4. Comparison with iterated Newton–Raphson.

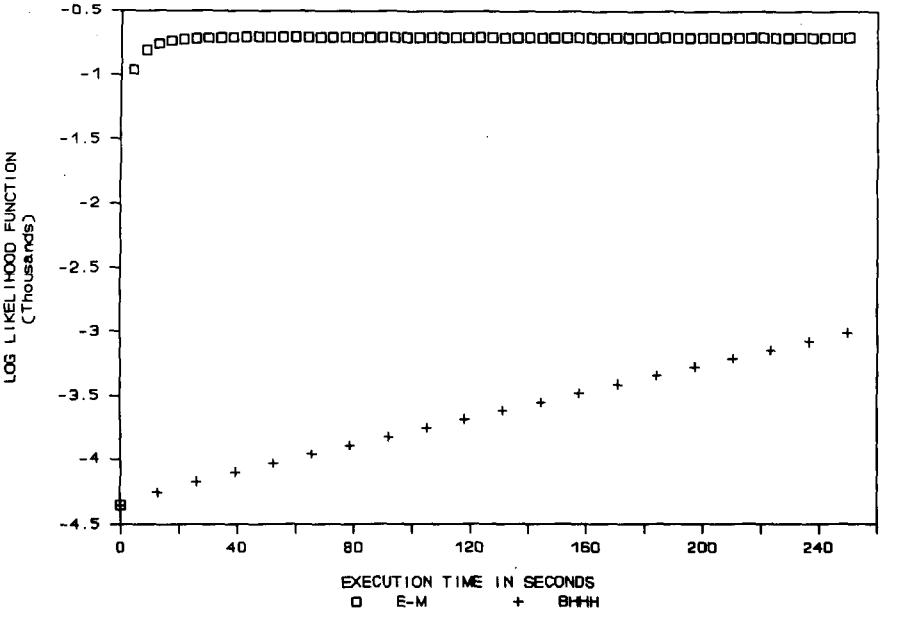

Fig. 5. EM versus BHHH for the Tobit model.

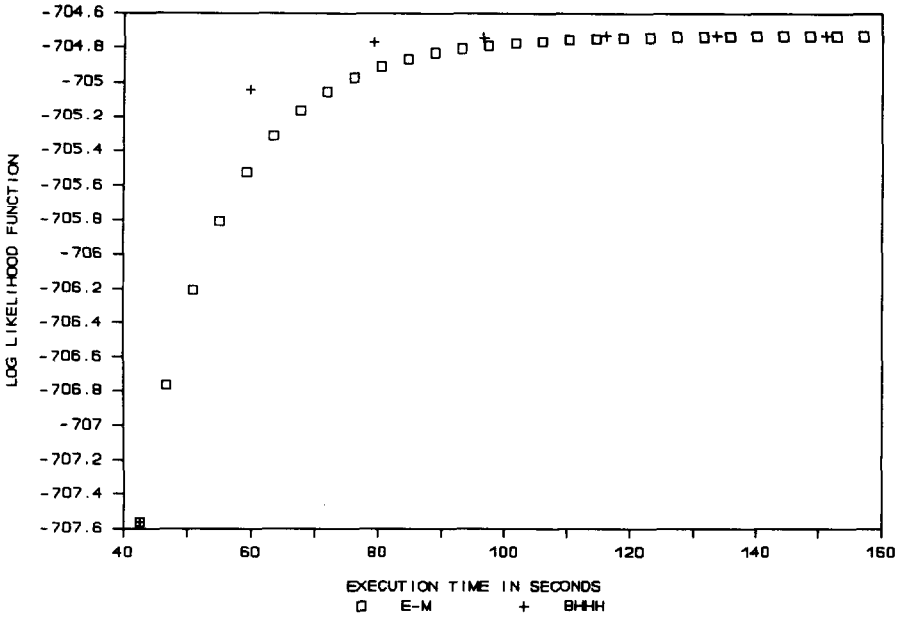

Fig. 6. Switching from EM to BHHH.

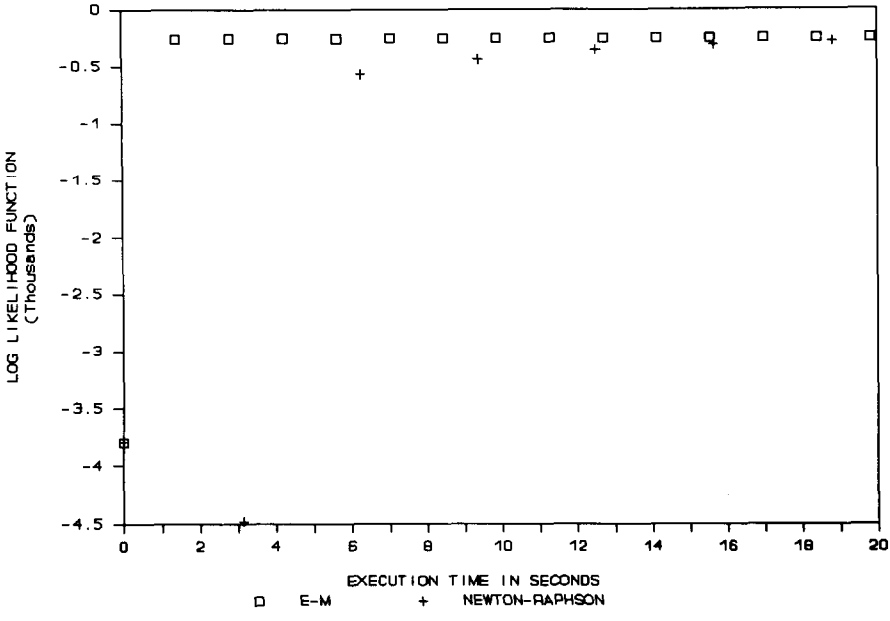

Fig. 7. Truncated regression.

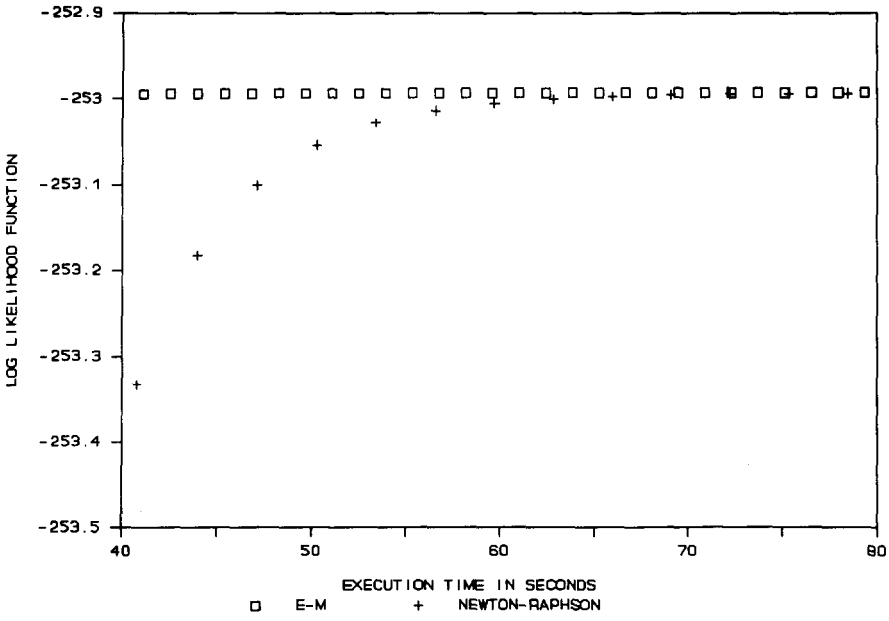

Fig. 8. Truncated regression.

rithms. Fig. 5 shows that the BHHH algorithm is ponderously slow at the outset, but if one were to switch to this algorithm after the fourth iteration of the EM algorithm, faster convergence would be obtained.

Figs. 7 and 8 describe convergence for the extension of the EM algorithm to missing data in section 5. In this case, the maximum-likelihood estimator is calculated using the likelihood for the 150 positive observations alone. The corresponding model is often called truncated regression. It is this model that is discussed in the appendix to section 5. The EM steps are compared with the iterated Newton–Raphson steps, as in figs. 3 and 4 for censored regression. The results are essentially the same, although the overall speed is faster because there are fewer observations.

Finally, we computed an SEM estimator from section 6 for the complete data set. This relatively simple problem is interesting because we know the MLE. More difficult estimation problems are straightforward to compute using the SEM estimator, but then the MLE would not be available.

The simulations were computed using (6.11). Since the estimator varies with the simulations, we present the results of a small Monte Carlo experiment of 155 trials in table 2. Comparison of the average values of the SEM estimators with the values of the MLE in table 1 shows a remarkable similarity. Given the standard errors in the last column, there is no evidence of bias in the SEM estimators, as estimators of the MLE, although we do

Table 2

Parameter		Average value	Sampling standard deviation	Standard error of the mean
β_1	Constant	7.6048	0.35375	0.02851
β_2	Sex dummy (1 = male)	0.95571	0.22913	0.01846
β_3	Age	−0.19397	0.03117	0.00251
β_4	Number of years married	0.53558	0.07412	0.00597
β_5	Number of children	1.0342	0.31715	0.02556
β_6	Index of religiosity	−1.7083	0.25598	0.02063
β_7	Index of education	0.025752	0.01224	0.00099
β_8	Index of occupation	0.21332	0.01432	0.00115
β_9	Index of marital happiness	−2.2890	0.33774	0.02722
σ		8.3354	1.4021	0.11298

expect some bias. The standard deviations in the second last column are the extra variation in the SEM estimator over the standard error of the MLE listed in table 1. Although the standard deviations in table 2 are relatively small, they could easily be reduced further using a larger sample of simulations than the sample size.

Appendix

Proof of Theorem 2

If L_{11} is nonsingular, then so is Q_{11} by (13) so that (16) and (17) are valid. Using the second-order Taylor series expansion of $L(\theta; y)$,

$$L(\theta_{NR}) - L(\theta_0) = -\tfrac{1}{2}Q_1'(Q_{11}+Q_{12})^{-1}Q_1 + o\left(\|\theta_{NR}-\theta_0\|^2\right)$$

and

$$L(\theta_{EM}) - L(\theta_0) = -\tfrac{1}{2}Q_1'Q_{11}^{-1}(Q_{11}-Q_{12})Q_{11}^{-1}Q_1 + o\left(\|\theta_{EM}-\theta_0\|^2\right),$$

where all expressions in Q are evaluated at θ_0. Choose $\delta > 0$ so that $L_{11} = Q_{11} + Q_{12}$ is negative definite for all $\theta_0 \in \{\theta \mid \|\theta - \hat{\theta}\| < \delta\}$. Expression (13) implies that within this ball

$$Q_{12} - Q_{11} \quad \text{and} \quad Q_{11}^{-1}(Q_{11}-Q_{12})Q_{11}^{-1} - (Q_{11}+Q_{12})^{-1}$$

are positive definite matrices so that

$$0 < L(\theta_{\mathrm{EM}}) - L(\theta_0) + o\big(\|\theta_{\mathrm{EM}} - \theta_0\|^2\big)$$

$$< L(\theta_{\mathrm{NR}}) - L(\theta_0) + o\big(\|\theta_{\mathrm{NR}} - \theta_0\|^2\big).$$

According to (16) and (17), $O(\|\theta_{\mathrm{NR}} - \theta_0\|) = O(\|\theta_{\mathrm{EM}} - \theta_0\|) = O(\|\hat{\theta} - \theta_0\|)$. Therefore, as θ_0 approaches $\hat{\theta}$,

$$0 < \lim \frac{L(\theta_{\mathrm{EM}}) - L(\theta_0)}{\|\theta - \theta_0\|^2} < \lim \frac{L(\theta_{\mathrm{NR}}) - L(\theta_0)}{\|\theta - \theta_0\|^2}.$$

Therefore, there is an open neighborhood of the MLE $\hat{\theta}$ such that (18) is satisfied for all θ_0 in that neighborhood. ■

Convergence of the EM algorithm for truncated regression

In section 5 we refer to the truncated normal location model as an example of a missing-data problem to which we can extend the EM algorithm successfully. Here we consider that case. Let y be normally distributed $N(\mu, \sigma^2)$ and truncated below at zero. We choose y^* to be $N(\mu, \sigma^2)$ and say that y is equal to y^* whenever $y^* \geq 0$; otherwise no observation is made. The Q and H functions are

$$Q(\theta, \theta_0; y) = -\tfrac{1}{2}\Big\{N \log \sigma^2 + \sigma^{-2}\Big[\Sigma(y_n - \mu)^2 + N\sigma_0^2$$

$$-N\,\mathrm{E}\big(u_n^2 | \mathscr{B}, \theta_0\big)\Big]\Big\} + N(\mu - \mu_0)\mathrm{E}\big(u_n | \mathscr{B}, \theta_0\big),$$

$$H(\theta, \theta_0) = -\tfrac{1}{2}\sigma^{-2}\Big[N\sigma_0^2 - N\,\mathrm{E}\big(u_n^2 | \mathscr{B}, \theta_0\big)$$

$$+2N\sigma_0(\mu - \mu_0)\mathrm{E}\big(u_n | \mathscr{B}, \theta_0\big) + N \log\big[\Phi(\mu/\sigma)\big],$$

where $\mathscr{B} = \{u_n | u_n \geq -\mu_0\}$ and $u_n = y_n - \mu_0$. For all observations,

$$\mathrm{E}\big(u_n | \mathscr{B}, \theta_0\big) = \sigma_0 \phi(\mu_0/\sigma_0)/\Phi(\mu_0/\sigma_0) \equiv \lambda(\mu_0, \sigma_0),$$

$$\mathrm{E}\big(u_n^2 | \mathscr{B}, \theta_0\big) = \sigma_0^2 - \mu_0\sigma_0\lambda(\mu_0, \sigma_0)$$

are constant. H is concave in μ because the log-probability term is concave

in μ. The second partial derivative of H with respect to σ^{-1} is

$$\partial^2 H/\partial(1/\sigma)^2 = -N\{\sigma_0\mu_0\lambda(\mu_0,\sigma_0) + \mu^2\lambda(\mu,\sigma)[\mu/\sigma + \lambda(\mu,\sigma)]\}.$$

If μ_0 is positive, then this derivative is negative for all μ and σ. Therefore, the GEM property holds for μ, and for σ provided that μ_0 is positive. There are negative values of μ_0 that will make the H function convex for some values of σ. Hence, we cannot prove the GEM property for this case using the concavity of H.

Instead, we will examine the iteration scheme itself. For μ,

$$\mu_i = \bar{y} - \sigma_{i-1}\lambda_{i-1},$$

and for σ, provided convergence in μ ($\mu = \bar{y} - \sigma_{i-1}\lambda(\mu,\sigma_{i-1})$),

$$\sigma_i^2 = s^2 + \sigma_{i-1}\lambda_{i-1}\bar{y},$$

where $\lambda_i = \lambda(\mu_i,\sigma_i)$,

$$\bar{y} = \Sigma y_n/N \quad \text{and} \quad s^2 = \Sigma(y_n - \bar{y})^2/N.$$

For a constant σ, the iterations in μ form a monotonic Cauchy sequence which will converge to a fixed point. In addition, this fixed point is monotonically decreasing in σ because

$$\{\lambda(x) > -x, \lambda'(x) > -1\} \Rightarrow \{\lambda(x) < \alpha\cdot\lambda(x/\alpha) \text{ if } \alpha > 1\}.$$

For a constant μ, we can show that the iterations in σ also form a monotonic Cauchy sequence. Let

$$\bar{\sigma}^2 = s^2 + \bar{\sigma}\lambda(\mu/\bar{\sigma})\bar{y}$$

be the fixed point and note that

$$\sigma_0 \gtrless \bar{\sigma} \Rightarrow \sigma_1 \gtrless \bar{\sigma},$$

using the same inequalities. Therefore the EM step does not overshoot. Furthermore, eq. (4.29) establishes that, in general, the EM algorithm points toward an increase in the likelihood function. Since $\bar{\sigma}$ occurs at the maximum, we also have that

$$\sigma_0 \gtrless \bar{\sigma} \Rightarrow \sigma_0 \gtrless \sigma_1.$$

Therefore, the GEM property also holds for the iterations in σ. ■

The central problem in establishing the convergence of our extension of the EM algorithm to missing-data problems in the exponential family, as illustrated here, is proving that the algorithm does not overshoot. We conjecture that this result holds more generally in exponential models.

References

Berndt, E.K., B.H. Hall, R.E. Hall, and J.A. Hausman, 1974, Estimation and inference in nonlinear structural models, Annals of Economic and Social Measurement 3, 653–666.

Boyles, Russell A., 1983, On the convergence of the EM algorithm, Journal of the Royal Statistical Society B 45, 47–50.

Bresnahan, Timothy J., 1981, Departures from marginal-cost pricing in the American automobile industry: Estimates for 1977–1978, Journal of Econometrics 17, 201–227.

Dempster, A.P., N.M. Laird, and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39, 1–38.

Efron, B. and D.V. Hinkley, 1978, Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information, Biometrika 65, 457–487.

Fair, R.C., 1977, A note on the computation of the tobit estimator, Econometrica 45, 1723–1727.

Fair, R.C., 1978, A theory of extramarital affairs, Journal of Political Economy 86, 45–61.

Haberman, S.J., 1977, Discussion on A.P. Dempster, N.M. Laird, and D.B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39, 31–32.

Hartley, Michael J., 1977a, On the estimation of a general switching regression model via maximum likelihood methods, Discussion paper no. 415 (Department of Economics, State University of New York at Buffalo, NY).

Hartley, Michael J., 1977b, On the calculation of the maximum likelihood estimator for a model of markets in disequilibrium, Discussion paper no. 409 (Department of Economics, State University of New York at Buffalo, NY).

Heckman, James J., 1976, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, Annals of Economic and Social Measurement 5, 475–492.

Heckman, James J. and B. Singer, 1984, A method for minimizing the impact of distributional assumptions in econometric models for duration data, Econometrica 52, 271–320.

Jewell, Nicholas P., 1982, Mixtures of exponential distributions, Annals of Statistics 10, 479–484.

Johnson, N.L. and S. Kotz, 1970, Continuous univariate distributions – 1 (Wiley, New York, NY).

Kiefer, N.M., 1978, Discrete parameter variation: Efficient estimation of a switching regression model, Econometrica 46, 427–434.

Kiefer, N.M., 1980, A note on switching regressions and logistic discrimination, Econometrica 48, 1065–1069.

Lee, Lung-Fei, 1976, Estimation of limited dependent variable models by two-stage methods, Ph.D. dissertation (University of Rochester, Rochester, NY).

Louis, Thomas A., 1982, Finding the observed information matrix when using the EM algorithm, Journal of the Royal Statistical Society B 44, 226–233.

Maddala, G.S., 1983, Limited-dependent and qualitative variables in econometrics (Cambridge University Press, Cambridge).

McFadden, D., 1989, A method of simulated moments for estimation of discrete response models without numerical integration, Econometrica 57, 995–1026.

McKelvey, R. and W. Zavoina, 1975, A statistical model for the analysis of ordinal level dependent variables, Journal of Mathematical Sociology 4, 103–120.

Meilijson, I., 1989, A fast improvement to the EM algorithm on its own terms, Journal of the Royal Statistical Society B 51, 127–138.

Olsen, R.J., 1978, Note on the uniqueness of the maximum likelihood estimator for the Tobit model, Econometrica 46, 1211–1215.

Pakes, A. and D. Pollard, 1989, The asymptotic distribution of simulation experiments, Econometrica 57, 1027–1058.
Rosett, R.N. and F.D. Nelson, 1975, Estimation of a two-limit probit regression model, Econometrica 43, 141–146.
Rothenberg, T.J., 1971, Identification in parametric models, Econometrica 39, 577–591.
Rothenberg, T.J., 1981, Personal communication.
Rubin, D.B. and D.T. Thayer, 1982, EM algorithms for ML factor analysis, Psychometrica 47, 69–76.
Ruud, P.A., 1988, Extensions of estimation methods using the EM algorithm, Working paper no. 8899 (University of California at Berkeley, CA).
Sampford, M.R., 1953, Some inequalities on Mill's ratio and related functions, Annals of Mathematical Statistics 24, 130–132.
Shumway, Robert H. and D.S. Stoffer, 1982, An approach to time series smoothing and forecasting using the EM algorithm, Journal of Time Series Analysis 3, 253–264.
Tsur, Yacov, 1983, The formulation and estimation of supply models with discrete/continuous decisions under uncertainty, Ph.D. dissertation (University of California at Berkeley, CA).
Watson, Mark and Robert F. Engle, 1983, Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression models, Journal of Econometrics 23, 385–400.
Wu, C.F. Jeff, 1983, On the convergence properties of the EM algorithm, Annals of Statistics 11, 95–103.