Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm

Peter Arcidiacono*

Duke University

John Bailey Jones†

State University of New York at Albany

August 12, 2002

Abstract

A popular way to account for unobserved heterogeneity is to assume that the data are drawn from a finite mixture distribution. A barrier to using finite mixture models is that parameters that could previously be estimated in stages must now be estimated jointly: using mixture distributions destroys any additive separability of the log-likelihood function. We show, however, that an extension of the EM algorithm reintroduces additive separability, thus allowing one to estimate parameters sequentially during each maximization step. In establishing this result, we develop a broad class of estimators for mixture models. Returning to the mixture problem, we show that, relative to full information maximum likelihood, our sequential estimator can generate large computational savings with little loss of efficiency.

*psarcidi@econ.duke.edu

†jbjones@csc.albany.edu. We thank Donald Andrews, Arie Beresteanu, Mark Coppejans, Michael McCracken, Tom Mroz, Barbara Rossi, Wilbert van der Klaauw, and two anonymous referees for valuable comments.

Keywords: Unobserved heterogeneity, mixture distributions, EM algorithm, dynamic discrete choice. JEL C13, C61, D90

1 Introduction

One way to account for unobserved heterogeneity in data, and the related problem of self-selection, is to assume that the data are drawn from a finite mixture distribution. Under this approach, each observation is assumed to belong to one of several different "types," each of which has its own distribution. While the econometrician does not observe each observation's type, if her model is sufficiently structured she can infer it by applying Bayes' Theorem.

Models with finite mixtures have appeared in numerous applications.[1] In labor economics, Keane and Wolpin (1997) and Eckstein and Wolpin (1999) use mixtures to control for person-specific differences in models of dynamic discrete choice. Finite mixture models form the basis of Hamilton's (1989) influential regime-switching model of economic time series. A particularly important application has been to use finite mixture models as nonparametric approximations to more general mixture models. Important papers in this vein include Laird (1978), Lindsay (1983) and Heckman and Singer (1984). More recently, Cameron and Heckman (1998, 2001) use this sort of nonparametric maximum likelihood estimation to study the effect of family background on educational achievement. Mroz (1999) uses mixtures to control for endogeneity in a binary explanatory variable. He shows that "discrete factor approximations" to a continuous latent variable often outperform alternative estimators, especially when the unobservable components of the model have a non-normal distribution.

One drawback to using mixture models is that they can complicate the estimation process. In this paper we focus on a particular problem, namely the issue of sequential likelihood. Some complicated likelihood models can be feasibly estimated only in stages; a subset of the parameters is estimated using one portion of the likelihood function, with the remainder of the parameters estimated with the remainder of the likelihood function, using the parameters esti-

[1] Although we focus on economic applications, finite mixture models have been used widely in other fields as well. Titterington, et al. (1985) and McLachlan and Peel (2000) provide lists.

mated in the preceding step(s). While introducing a mixture distribution seemingly prevents one from proceeding sequentially, we show that if one extends the Expectation-Maximization (EM) algorithm, one can still estimate the likelihood function in steps.

In contrast to the EM algorithm, which is ultimately a search algorithm, our procedure does not yield full information maximum likelihood (FIML) estimates. Rather, our procedure introduces a broad class of estimators for mixture and switching models. In particular, a simple argument shows that any moment condition that holds across the unobserved "types" or "states" generates a moment condition that holds across the observed data.

In addition to providing general results, we construct a Monte Carlo exercise that shows the large savings in computational time from employing the EM algorithm with a sequential maximization step (ESM). Although the gains to using the method are problem-specific, we show reductions in computing time on the order of 20 for a relatively simple problem. More complicated problems should show even larger reductions. A further benefit of the ESM algorithm is that moving from a problem without unobserved heterogeneity to one with unobserved heterogeneity requires little change in code.

The next section reviews mixture distributions and the EM algorithm. Section 3 shows how the EM algorithm introduces an additive separability not previously present in mixture models. This allows for a sequential maximization step. Section 4 describes the asymptotics of our estimator, and shows how it can be generalized. Section 5 provides simulations showing that the ESM estimator performs as well as FIML and takes significantly less time to converge. Section 6 concludes.

2 Mixture Distributions and the EM Algorithm

The general relationship between mixtures and the EM algorithm has been covered in a number of sources, such as Everitt and Hand (1981), Titterington, et al. (1985), and Hamilton (1990). We provide a brief review.

Consider a panel data set of I individuals, where for each individual i we observe T realizations of the J-element vector x. Observations of x are independent across individuals,

although not necessarily across time. As a matter of notation, let the collection of x-vectors for agent i be denoted by the $J \cdot T$-element vector $\mathbf{x}_i = \left[x'_{i,1}, x'_{i,2}, ..., x'_{i,T}\right]'$.

Each individual belongs to one of K distinct types. While the econometrician knows K, he does not observe individuals' types. Let p_k denote the unconditional probability that an individual belongs to type k, with $\mathbf{p} = (p_1, p_2, ..., p_K)$ denoting the vector of these probabilities. Letting $f_k(\cdot)$ denote the density function for type k, and letting $\mathbf{\Theta} = (\theta_1, ..., \theta_M)$ denote a vector of parameters, the unconditional likelihood of $\mathbf{x}_i$ is

$$g(\mathbf{x}_i; \mathbf{\Theta}, \mathbf{p}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}_i; \mathbf{\Theta}).$$

It follows from Bayes' theorem that $\Pr(k|\, \mathbf{x}_i; \mathbf{\Theta}, \mathbf{p})$, the probability that agent i is of type k, conditional on having observed $\mathbf{x}_i$, is given by

$$\Pr(k|\, \mathbf{x}_i; \mathbf{\Theta}, \mathbf{p}) = \frac{p_k f_k(\mathbf{x}_i; \mathbf{\Theta})}{g(\mathbf{x}_i; \mathbf{\Theta}, \mathbf{p})}. \tag{1}$$

Let $\mathbf{S}_K$ denote the $K-1$-dimensional unit simplex. Using equation (1), it is straightforward to show that if one maximizes the sample log-likelihood, $L(\mathbf{\Theta}, \mathbf{p}) \equiv \sum_i \ln(g(\mathbf{x}_i; \mathbf{\Theta}, \mathbf{p}))$, subject to the restriction $\mathbf{p} \in \mathbf{S}_K$, the maximum likelihood estimate $\widehat{p}_k$ is given by

$$\widehat{p}_k = \frac{1}{I} \sum_{i=1}^{I} \Pr(k|\, \mathbf{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{p}}). \tag{2}$$

The maximum likelihood estimate $\widehat{\mathbf{\Theta}}$ must solve

$$\sum_{i=1}^{I} \sum_{k=1}^{K} \Pr(k|\, \mathbf{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{p}}) \frac{\partial \ln(f_k(\mathbf{x}_i; \widehat{\mathbf{\Theta}}))}{\partial \mathbf{\Theta}} = \mathbf{0}, \tag{3}$$

so that

$$\widehat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{I} \sum_{k=1}^{K} \Pr(k|\, \mathbf{x}_i; \widehat{\mathbf{\Theta}}, \widehat{\mathbf{p}}) \ln(f_k(\mathbf{x}_i; \mathbf{\Theta})). \tag{4}$$

In other words, $\widehat{\mathbf{\Theta}}$ maximizes the sample average of two different objects: (1) the log of the unconditionally-type-averaged likelihood ($\ln[\sum_k p_k f_k(\mathbf{x}_i)]$); and (2) the conditionally-type-averaged log-likelihood ($\sum_k \Pr(k|\, \mathbf{x}_i) \ln[f_k(\mathbf{x}_i)]$). The key insight of our paper is that while the first object does not support sequential estimation, the second one does.

Equations (1) through (4) suggest the following iterative algorithm, which is a special case of the EM algorithm developed by Dempster, Laird and Rubin (1977). Suppose that at the beginning of of iteration l, the operative value of $\mathbf{\Theta}$ is $\mathbf{\Theta}^l$ and the operative value of $\mathbf{p}$ is $\mathbf{p}^l$. In the "E" step, one uses equation (1) to find $\Pr\left(k|\,\mathbf{x}_i;\mathbf{\Theta}^l,\mathbf{p}^l\right)$. In the "M" step, one uses equations (2) and (4) to find $\mathbf{p}^{l+1}$ and $\mathbf{\Theta}^{l+1}$, respectively. One iterates until convergence.

3 The EM Algorithm with a Sequential M Step

Now divide the parameter vector $\mathbf{\Theta}$ into $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$. Clearly the solution to equation (4) can be found by maximizing across $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ simultaneously, or by iterating, using the most recent value of $\widehat{\mathbf{\Theta}}_1$ to update $\widehat{\mathbf{\Theta}}_2$ and then using this updated value of $\widehat{\mathbf{\Theta}}_2$ to recalculate $\widehat{\mathbf{\Theta}}_1$. For some applications, it is easier to proceed sequentially. Meng and Rubin (1993) call this approach the Expectation-Conditional Maximization (ECM) algorithm, and show that the ECM algorithm retains all of the convergence properties of the EM algorithm.[2]

A more interesting case occurs when the type-conditional likelihood function can be decomposed as

$$f_k(\mathbf{x}_i;\mathbf{\Theta}_1,\mathbf{\Theta}_2) = f_{1k}(\mathbf{x}_i;\mathbf{\Theta}_1)f_{2k}(\mathbf{x}_i;\mathbf{\Theta}_1,\mathbf{\Theta}_2),$$

and $f_{1k}(\mathbf{x}_i;\mathbf{\Theta}_1)$ can be written as a product of type-conditional likelihoods:

$$f_{1k}(\mathbf{x}_i;\mathbf{\Theta}_1) = \prod_{j=1}^{J} f_{1k}(\mathbf{x}_{i,j}|\,\mathbf{x}_{i,\sim j};\mathbf{\Theta}_1), \tag{5}$$

where $\mathbf{x}_{i,j}$ and $\mathbf{x}_{i,\sim j}$ are mutually exclusive subvectors of $\mathbf{x}_i$.

It proves instructive to consider the log-likelihood that arises when $K = 1$, i.e., there is only one type:

$$\begin{aligned} L(\mathbf{\Theta}) &= \sum_{i=1}^{I} \ln\left(f_1(\mathbf{x}_i;\mathbf{\Theta}_1)\right) + \sum_{i=1}^{I} \ln\left(f_2(\mathbf{x}_i;\mathbf{\Theta}_1,\mathbf{\Theta}_2)\right), \\ &\equiv L_1(\mathbf{\Theta}_1) + L_2(\mathbf{\Theta}_1,\mathbf{\Theta}_2). \end{aligned}$$

[2]As Ruud (1991) points out, one can update $\Pr(k|\,\mathbf{x}_i;\widehat{\mathbf{\Theta}},\widehat{\mathbf{p}})$ each time *either* $\widehat{\mathbf{\Theta}}_1$ or $\widehat{\mathbf{\Theta}}_2$ is updated. Meng and Rubin (1993) label this the "multi-cycle ECM" algorithm. Also see the discussion in McLachlan and Krishnan (1997).

In this case, consistent estimates of $\boldsymbol{\Theta}_1$ can be found by maximizing L_1, while consistent estimates of $\boldsymbol{\Theta}_2$ can be found from maximizing L_2, taking as given the estimates of $\boldsymbol{\Theta}_1$.[3] Note that this differs from the ECM approach in that we are not maximizing $f_k(.)$ in steps, but are instead sequentially maximizing two partial likelihoods. While this approach is less efficient than maximizing the log of $f(.)$, it is often much easier to implement, especially when L_2 is difficult to evaluate.

For example, Rust (1994) considers the maximum likelihood estimator for a Markov decision process:

$$\widehat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{I} \ln \left(\prod_{t=1}^{T} P(\, d_t^i \big|\, s_t^i; \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \pi(\, s_t^i \big|\, s_{t-1}^i, d_{t-1}^i; \boldsymbol{\Theta}_1) \right), \tag{6}$$

where d_t^i is agent i's decision vector at time t, and s_t^i is the vector of state variables that characterizes agent i's economic environment at time t. While $\pi(\, s_t^i | \, .)$ is straightforward to evaluate, $P(d_t^i | \, .)$ requires one to solve a dynamic programming problem. Rust finds that estimating $\boldsymbol{\Theta}_1$ as the maximizer of $\sum_i \sum_t \ln \left(\pi(\, s_t^i \big|\, s_{t-1}^i, d_{t-1}^i; \boldsymbol{\Theta}_1) \right)$ can greatly reduce the number of times $\sum_i \sum_t \ln \left(P(\, d_t^i \big|\, s_t^i; \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \right)$ must be evaluated, which in turn significantly lowers computational cost. Indeed, Rust and Phelan (1997) conclude that "[e]stimation is only feasible using a simpler two-stage estimation procedure[.]"

In the finite mixture case, the log-likelihood is

$$L(\boldsymbol{\Theta}, \mathbf{p}) = \sum_{i=1}^{I} \ln \left(\sum_{k=1}^{K} p_k f_{1k} \left(\mathbf{x}_i; \boldsymbol{\Theta}_1 \right) f_{2k} \left(\mathbf{x}_i; \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \right) \right),$$

which cannot be neatly decomposed into L_1 and L_2. This seemingly destroys the option of sequential estimation. But with the EM algorithm we work with equation (4) which can be written as

$$\begin{aligned} (\widehat{\boldsymbol{\Theta}}_1, \widehat{\boldsymbol{\Theta}}_2) \;\; = \;\; & \arg \max_{\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}} \sum_{i=1}^{I} \sum_{k=1}^{K} \Pr(\, k \big|\, \mathbf{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\mathbf{p}}) \ln \left(f_{1k}(\mathbf{x}_i; \boldsymbol{\Theta}_1) \right) + \\ & \sum_{i=1}^{I} \sum_{k=1}^{K} \Pr(\, k \big|\, \mathbf{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\mathbf{p}}) \ln \left(f_{2k}(\mathbf{x}_i; \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \right). \end{aligned}$$

[3] The asymptotic properties of these sorts of two-step estimators are discussed in Cox (1975) and Amemiya (1978), as well as in the next section.

Once again we can proceed sequentially, using the partial likelihood estimators

$$\widetilde{\boldsymbol{\Theta}}_1 = \arg\max_{\boldsymbol{\Theta}_1} \sum_{i=1}^{I}\sum_{k=1}^{K} \Pr(k|\,\mathbf{x}_i;\widetilde{\boldsymbol{\Theta}},\widetilde{\mathbf{p}}) \ln\left(f_{1k}(\mathbf{x}_i;\boldsymbol{\Theta}_1)\right), \tag{7}$$

$$\widetilde{\boldsymbol{\Theta}}_2 = \arg\max_{\boldsymbol{\Theta}_2} \sum_{i=1}^{I}\sum_{k=1}^{K} \Pr(k|\,\mathbf{x}_i;\widetilde{\boldsymbol{\Theta}},\widetilde{\mathbf{p}}) \ln\left(f_{2k}(\mathbf{x}_i;\widetilde{\boldsymbol{\Theta}}_1,\boldsymbol{\Theta}_2)\right). \tag{8}$$

Applying the EM algorithm in this way introduces an additive separability that allows $\boldsymbol{\Theta}$ to be estimated sequentially, with each stage using the estimates from the previous stage. Note that the derivative of $f_{2k}(.)$ with respect to $\boldsymbol{\Theta}_1$ never has to be calculated. This means that the estimates generated by equations (7) and (8) are less efficient than the FIML estimates, but potentially much easier to compute.

4 Asymptotic Behavior of the Sequential Estimator

As the review in Section 2 reveals, the EM algorithm is a method for finding standard FIML estimates. Our sequential estimator, on the other hand, is not equivalent to FIML. The asymptotic properties of our estimator can be shown instead by constructing moment conditions, to which standard GMM results can be applied. In the next section we derive these moment conditions. In the succeeding section, we discuss conditions that ensure the parameters of interest are identified. We finish our theoretical discussion by showing how our approach generates a wide class of estimators.

4.1 Moment Conditions

Let starred values denote population parameters. Note first that at the population level

$$(\boldsymbol{\Theta}^*,\mathbf{p}^*) = \arg\max_{\{\boldsymbol{\Theta},\mathbf{p}\in\mathbf{S}_K\}} E_{\mathbf{x},k}\left(\ln\left[p_k f_{1k}(\mathbf{x};\boldsymbol{\Theta}_1) f_{2k}(\mathbf{x};\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)\right]\right), \tag{9}$$

with the expectation taken over both k and $\mathbf{x}$. It then follows from the law of total probability that

$$(\boldsymbol{\Theta}^*,\mathbf{p}^*) = \arg\max_{\{\boldsymbol{\Theta},\mathbf{p}\in\mathbf{S}_K\}} E_{\mathbf{x}}\left(\sum_{k=1}^{K}\Pr(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*)\ln\left[p_k f_{1k}(\mathbf{x};\boldsymbol{\Theta}_1) f_{2k}(\mathbf{x};\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)\right]\right), \tag{10}$$

with the latter expectation taken over $\mathbf{x}$ alone. This result is the self-consistency property, which dates back to work by R.A. Fisher.[4]

It immediately follows from equation (10) that $\boldsymbol{\Theta}_2^*$ solves

$$\max_{\boldsymbol{\Theta}_2} E_{\mathbf{x}} \left(\sum_{k=1}^{K} \Pr(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*) \ln\left[f_{2k}(\mathbf{x};\boldsymbol{\Theta}_1^*,\boldsymbol{\Theta}_2)\right] \right),$$

the population analog to equation (8). The first-order condition for this problem is

$$E_{\mathbf{x}} \left(\sum_{k=1}^{K} \Pr\left(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*\right) \frac{\partial \ln\left(f_{2k}\left(\mathbf{x};\boldsymbol{\Theta}^*\right)\right)}{\partial \boldsymbol{\Theta}_2} \right) = \mathbf{0}.$$

The population analog to equation (2) can be constructed in a similar fashion.

Since $f_{1k}\left(\mathbf{x}_i;\boldsymbol{\Theta}_1\right)$ is a type-conditional likelihood in its own right—recall equation (5)—it must solve[5]

$$\max_{\boldsymbol{\Theta}_1} E_{\mathbf{x}} \left(\sum_{k=1}^{K} \Pr\left(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*\right) \ln\left[f_{1k}(\mathbf{x};\boldsymbol{\Theta}_1)\right] \right),$$

the population analog to equation (7). The associated first-order condition is

$$E_{\mathbf{x}} \left(\sum_{k=1}^{K} \Pr\left(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*\right) \frac{\partial \ln\left(f_{1k}(\mathbf{x};\boldsymbol{\Theta}_1^*)\right)}{\partial \boldsymbol{\Theta}_1} \right) = \mathbf{0}.$$

The population moment conditions for $\boldsymbol{\Theta}$ and $\mathbf{p}$ are thus:

$$E_{\mathbf{x}} \left(\begin{array}{c} \sum_{k=1}^{K} \Pr(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*) \frac{\partial \ln\left(f_{1k}(\mathbf{x};\boldsymbol{\Theta}_1^*)\right)}{\partial \boldsymbol{\Theta}_1} \\ \sum_{k=1}^{K} \Pr(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*) \frac{\partial \ln(f_{2k}(\mathbf{x};\boldsymbol{\Theta}^*))}{\partial \boldsymbol{\Theta}_2} \\ \Pr(1|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*) - p_1^* \\ \vdots \\ \Pr(K|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*) - p_K^* \end{array} \right) = \mathbf{0}, \tag{11}$$

with $\Pr(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*)$ given by equation (1). Then it follows from standard arguments (see Hansen, 1982, or Newey and McFadden, 1994) that, subject to the usual regularity conditions, $\widetilde{\boldsymbol{\Theta}}_1$, $\widetilde{\boldsymbol{\Theta}}_2$ and $\widehat{\mathbf{p}}$ are consistent and asymptotically normal, with the variance-covariance matrix given by the standard method-of-moments formula. Note that even though $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ can be estimated sequentially, finding standard errors requires evaluating all the moment conditions

[4]See the discussion in Efron (1982) and McLachlan and Krishnan (1997).

[5]Also see Cox's (1975) discussion of partial likelihood.

together.[6] Equation (11) also reveals that the sequential estimator will not be as efficient as FIML, for

$$\frac{\partial \ln\left(g(\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*)\right)}{\partial \boldsymbol{\Theta}_1} = \sum_{k=1}^{K} \Pr(k|\,\mathbf{x};\boldsymbol{\Theta}^*,\mathbf{p}^*)\left[\frac{\partial \ln\left(f_{1k}(\mathbf{x};\boldsymbol{\Theta}_1^*)\right)}{\partial \boldsymbol{\Theta}_1} + \frac{\partial \ln\left(f_{2k}(\mathbf{x};\boldsymbol{\Theta}^*)\right)}{\partial \boldsymbol{\Theta}_1}\right],$$

which means that the first element of the moment vector in equation (11) is not part of the score vector for the FIML function, even though the remaining elements are.

4.2 Asymptotic Identification

Consistency and asymptotic normality require that an estimator satisfy regularity conditions of the sort set forth by Newey and McFadden (1994). Of these the most important is asymptotic identification. One approach for achieving identification is to assume that the moment conditions given by equation (11) are satisfied only by the parameter vector $(\boldsymbol{\Theta}^*, \mathbf{p}^*)$. Given that mixture likelihoods are often not globally concave, we also consider an alternative approach. In particular, we assume that the expectation of the log-likelihood function is uniquely maximized at $(\boldsymbol{\Theta}^*, \mathbf{p}^*)$ and characterize the moment conditions listed in equation (11) as features of this optimum.[7]

Wu (1983) shows that the EM algorithm converges to flat points on the likelihood surface, so that the EM solution yielding the highest likelihood value can be taken as the maximum likelihood estimate. One can see this heuristically by considering equations (2) and (4). While our sequential estimator is not a reformulation of the FIML estimator, it can nonetheless be used in a similar way. In particular, one can apply the likelihood criterion when the sample analog to equation (11) has multiple solutions.[8] Although this does not yield FIML estimates—equation (11) is not the FIML score—using a likelihood tiebreaker ensures consistency. We provide a formal proof of consistency in the appendix, using arguments that apply

[6] Rust (1994) discusses this issue in some detail for the one-type case.

[7] In assuming uniqueness, we are imposing several normalizations. Titterington et al. (1985) discuss exact conditions for identifying finite mixture models.

[8] In choosing this way, one must take care to restrict oneself to stationary solutions. It is well known, for example, that one can drive the sample log-likelihood of a normal mixture to infinity by assuming that one of the observations belongs to its own zero-variance type.

to almost any GMM estimator.[9]

It is worth reiterating that even if the population likelihood function has a unique maximum, FIML estimation can require one to compare numerous local extrema on the sample likelihood surface. If the ESM algorithm yields multiple fixed points, it is likely to be the case that a gradient-based FIML search will yield multiple solutions as well. In either case, a likelihood tiebreaker will have to be applied. The difference is that in the searches *before* the tiebreaker is applied, the sequential estimator can be much less computationally demanding.

To this point, we have focused on how the ESM algorithm generates a sequential alternative to the FIML estimator. A different approach is to use the ESM algorithm to generate initial values for a FIML search. Using the ESM algorithm in this way allows one to enjoy some of the cost savings of sequential estimation without losing asymptotic efficiency. A particularly interesting possibility is to utilize the ESM algorithm as a search routine in nonparametric maximum likelihood, in a way similar to how Follmann and Lambert (1989) combine the EM and quasi-Newton algorithms. Such an approach extends the benefits of sequential estimation to cases where the number of types (K) is not known.[10]

As Rust (1994) points out, yet another way to recover asymptotic efficiency is to use the sequential estimator as the basis for a one-step estimator: starting with the sequential estimates, one can take one Gauss-Newton step with the full likelihood function.

4.3 Generalizations of the Sequential Estimator

Our approach extends in a very straightforward way to general moment conditions in mixture models. In the interest of brevity, we continue to work with finite mixtures, but extensions to general mixtures or regime-switching models are straightforward.

[9] An interesting result from this section is that to ensure consistency, one has to consider local as well as global minima of the GMM criterion function generated by equation (11).

[10] We are grateful to a referee for this suggestion. As described by Heckman and Singer (1984) and Follmann and Lambert, when K is unknown one proceeds by finding FIML estimates with successively larger values of K until, roughly speaking, the derivative of the likelihood function with respect to K is non-positive. A topic we do not explore here is whether ESM estimates can fully replace FIML estimates in this computationally intensive procedure, or can serve only as starting values.

As before, it follows from the law of total probability any function $\mathbf{h}(\mathbf{x}, k; \boldsymbol{\Theta})$ that satisfies

$$E_{\mathbf{x},k}\left(\mathbf{h}(\mathbf{x}, k; \boldsymbol{\Theta}^*)\right) = \mathbf{0},$$

also satisfies

$$E_{\mathbf{x}}\left(\sum_{k=1}^{K} \Pr(k|\,\mathbf{x}; \boldsymbol{\Theta}^*, \mathbf{p}^*)\mathbf{h}(\mathbf{x}, k; \boldsymbol{\Theta}^*)\right) = \mathbf{0}. \tag{12}$$

Equation (12) provides a basis for estimation. There is a long tradition of analyzing mixture distributions with classical method of moments estimators;[11] we have simply extended the classical approach to general moment conditions. As in the motivating case of sequential likelihood, some of these alternative conditions might be less computationally demanding than the likelihood equations. It is also straightforward to construct overidentification tests.

By way of example, consider the following linear regression model:

$$y_i = \mathbf{x}_i'\mathbf{b}_k^* + e_i,$$

where: $\mathbf{x}_i$ is an M-element random vector; $\mathbf{b}_k$ is a parameter vector; and e_i is a standard logistic random variable that is independent of $\mathbf{x}_i$. As before i indexes observation and k denotes observation i's unobserved type. Let $\boldsymbol{\Theta}$ denote the collection of $\mathbf{b}$'s. Note that

$$E_{y,\mathbf{x}}\left(\Pr(k|\,y, \mathbf{x}; \boldsymbol{\Theta}^*, \mathbf{p}^*)\mathbf{x}\left[y - \mathbf{x}'\mathbf{b}_k^*\right]\right) = \mathbf{0}, \quad k \in \{1, ..., K\}.$$

Following Kiefer (1980), under random sampling the sample analog to this equation can be found using weighted least squares, where the weighting matrix $\widehat{\mathbf{W}}_k$ is a diagonal matrix whose i-th element is $\sqrt{\Pr(k_i|\,y_i, \mathbf{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\mathbf{p}})}$. As before, one can proceed iteratively, estimating $\widehat{\mathbf{b}}_k$ with the sample matrices $\widehat{\mathbf{W}}_k\mathbf{X}$, $\widehat{\mathbf{W}}_k\mathbf{y}$, and using these estimates to update $\widehat{\mathbf{W}}_k$.

5 Simulations

Two questions remain. First, are there common cases where the sequential M step results in significant savings in computational time? Second, since the two-step estimator described above is not efficient, how much information is lost by using it? To address these issues,

[11] See Everitt and Hand (1981), and Titterington, et al. (1985).

we perform a Monte Carlo simulation with a dynamic discrete choice problem. Even for this relatively simple problem, the computational gains are quite large, with little loss of information.

5.1 The Model

The model we use in our Monte Carlo exercise is one of sequential decision-making because, as discussed above, sequential estimation works particularly well with models of dynamic choice. The model we simulate is similar in spirit to Cameron and Heckman (2001).[12] In each of three periods, individuals decide whether to continue their education. In the fourth period individuals receive earnings. Earnings depend on education, observable characteristics, a random shock and an individual's unobserved type. Different types have different labor market abilities, and have different preferences over education itself. Individuals face uncertainty over both the pecuniary and non-pecuniary returns to education. As time passes, individuals receive new information that allows them to reduce this uncertainty.[13]

In the absence of type-based differences, the likelihood function for education choices and earnings generated by this model resembles the likelihood function in equation (6) and can be estimated in a similar sequential fashion. This will yield consistent estimates of, among other things, the returns to college, γ_C. But with unobservable type-based differences, estimates of γ_C will be biased upwards (and inconsistent) unless the estimates account for type-based selection. The goal of the Monte Carlo exercise is to see whether the ESM algorithm can account for selection, by estimating the mixture model, more quickly and as accurately as FIML.

[12] The model also resembles Aricidiacono's (2002) model of application, college, and major choice.

[13] A detailed description of the model, the parameters of the data generating process, and the starting values for the optimization routines are in a simulation appendix and can be downloaded from http:\\www.econ.duke.edu\~psarcidi\simulation.pdf.

5.2 Simulation Results

All of the simulations were conducted in Matlab, using Matlab's **fminunc** optimization package. The number of individuals is fixed at 3000 and the number of types is two. Crucial to the calculation time is the number of points that are used to approximate the distribution of new information. An increase in the number of points leads to more complicated expectations and a larger computational burden. For the simulations, we approximate the distributions of unknown state variables with 10-point discrete distributions. This discretization is applied to two unknown state variables at $t = 1$ and one unknown state variable at $t = 2$.

The model is estimated 100 times using four different methods. First, we estimate the model with the complete data, where each individual's type is observed. Second, we estimate the model with incomplete data, where type is unobserved, and pretend that there is no selection problem. We then control for unobserved types by estimating the mixture model, first with FIML and then with the ESM algorithm. We do not report estimates for the EM algorithm itself, as it was substantially slower than FIML.

As we are primarily interested in how well the various approaches to estimating the mixture distribution mitigate the selection problem, we only report the coefficient on the return to college, γ_C.[14] The key feature of the model is that the estimates of γ_C are biased upwards from the population value of 0.2 (and inconsistent) unless the estimates account for selection based upon type. We also report the standard deviation of the estimated returns and the mean squared difference between the estimates and the true value of γ_C. To get a sense of speed, we record the number of floating point operations (FLOPs) the various algorithms took to converge.[15] We then report the ratio of FIML FLOPs to ESM FLOPs.

Table 1 presents the simulation results. As expected, not controlling for the selection problem yields estimates of γ_C that are too high relative to the complete data estimates. Using either FIML or ESM to estimate the mixture model yields estimates much closer to

[14] All of the approaches produced similar estimates for the other coefficients.

[15] Jamshidian and Jennrich (1993, 1997) use this measure of speed in their study of enhancements to the EM algorithm. An advantage of using FLOPs is that we were able to run simulations on multiple computers of varying clock speeds and still have a consistent measure of speed.

Table 1: Simulation Results

Simulation Results[†]				
	Complete	Incomplete	FIML	ESM
Mean($\hat{\gamma}_C$)	0.2078	0.2932	0.2255	0.2226
Standard Deviation($\hat{\gamma}_C$)	0.0330	0.0323	0.0496	0.0565
Mean Squared Error($\times 100$)[‡]	0.1141	0.9731	0.3082	0.3667
(FIML FLOPs)/(ESM FLOPs)				22.48

[†]Each simulation was conducted 100 times with 3000 observations. The underlying model is described in Appendix B. The distributions of unknown state variables were approximated with 10-point discrete distributions. Values for the data generating process are found in Appendix B.

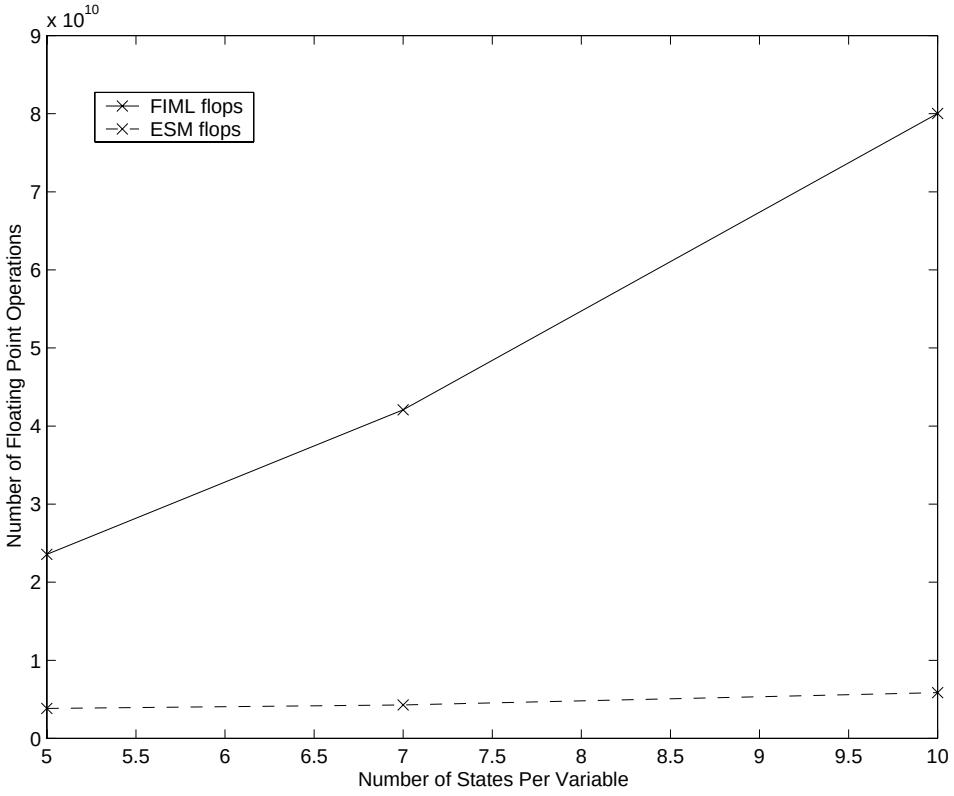
[‡] Mean squared error refers to the squared differences between estimates of γ_C and its true value of 0.2.

those found when individuals' types were observed. Moving from FIML to ESM increases the standard deviation for $\hat{\gamma}_C$ and the mean squared error, both by less than twenty percent. The last line in Table 1 shows that this relatively small loss of precision leads to large gains in speed: the ESM algorithm improves the rate of convergence by a factor of roughly twenty. To see the rate at which adding states affects computational time, we also perform the simulation with five and seven states for each of the discretized state variables. Figure 1 graphs the number of FLOPS for FIML and ESM. While the computational gains are large for five states (over ten times as fast), it is clear that the gains increase with the number of states.

5.3 Discussion

It is worth stressing that the ESM algorithm requires little researcher time: programming the algorithm can be very easy. In the algorithm's simplest form, all that one needs is to save the full density functions so that one can estimate the type probabilities by Bayes' rule. One otherwise uses the same estimators as in the non-mixture case, except that the data are weighted by the imputed type probabilities. Hence, adding decisions or state variables has very little effect on the time spent on programming the ESM algorithm. In general, the ESM algorithm is easier to program than FIML. Because the simulations behind Table 1 employ

Figure 1: Number of FLOPs as a Function of the Number of States



the simplest version of the ESM algorithm, the substantial savings in computational time come with savings in programming time as well.

There are, however, at least three reasons to believe that the estimates in Table 1 give lower bounds on the computational savings from the ESM algorithm. First, in our current optimization routine, the estimate of the Hessian at each stage is re-initialized at the beginning of each ESM maximization step. Hence, all the updating of the Hessians that occurs while maximizing the type-conditional log-likelihood functions is lost. Changing the optimization code to carry estimated Hessians across ESM iterations could substantially reduce convergence times.

Second, the convergence criteria we used at the maximization step did not depend upon how close the ESM algorithm was to converging. Precise maximization is not necessary when the ESM algorithm is far from the optimum. Setting the convergence criteria at the maximization step to be a function of the changes in the conditional probabilities and the

likelihoods should thus speed up convergence.

Third, some of the work in the statistics literature on accelerating the EM algorithm can be applied here. An iteration in the EM (or ESM) algorithm uses $\boldsymbol{\Theta}^l$ and $\mathbf{p}^l$ to find $\Pr\left(k|\,\mathbf{x}_i;\boldsymbol{\Theta}^l,\mathbf{p}^l\right)$, $\mathbf{p}^{l+1}$ and $\boldsymbol{\Theta}^{l+1}$. This can be described as

$$(\boldsymbol{\Theta}^{l+1},\mathbf{p}^{l+1})=\mathbf{G}(\boldsymbol{\Theta}^l,\mathbf{p}^l), \tag{13}$$

where $\mathbf{G}(.)$ is the vector-valued function given by an EM iteration. It is then easy to see that the EM estimates are fixed points in the nonlinear system given by equation (13). Given that the EM algorithm proceeds iteratively, there are potential speed gains if one treats the EM estimate as a zero of a system of non-linear equations, and uses more sophisticated solution routines to find these zeros. Jamshidian and Jennrich (1997) show that using quasi-Newton methods to solve equation (13) can accelerate convergence of the EM algorithm, sometimes dramatically.

6 Conclusion

This paper provides a simple way to add unobserved heterogeneity to models that, in the absence of such heterogeneity, could be estimated sequentially. In particular, if one assumes that the data are drawn from a finite mixture distribution, the EM algorithm contains a step where one maximizes an additively separable type-conditional log-likelihood function. Hence, one can control for unobserved heterogeneity even in problems where the parameters are most simply estimated in stages. Although our ESM algorithm does not yield FIML estimates, it is asymptotically well-behaved—in fact the ESM estimator introduces a broad class of GMM-type estimators. Simulation results show that the ESM algorithm performs very well with substantial computational savings and little loss of information.

Appendix: Consistency with Weak Identification

We begin with some notation. Assume that we have an i.i.d. sample of $\mathbf{x}$'s of size I. Let $\mathbf{s}=[\theta',\mathbf{p}']'\in\mathbf{S}\subset\mathcal{R}^{M+K}$ denote a parameter vector, with $\mathbf{s}^*$ denoting the population value of $\mathbf{s}$

and $\widehat{\mathbf{s}}$ denoting a sample estimate. Let $Q(\mathbf{s})$ denote the negative of a GMM criterion function, such as the one behind the sequential estimator, and let $Q_I(\mathbf{s})$ denote the sample analog of $Q(.)$. Similarly, let $L(\mathbf{s}) \equiv E\left(\ln(g(\mathbf{x};\mathbf{s}))\right)$ and $L_I(\mathbf{s})$ denote the population and sample expectations of the log-likelihood function. Let $NM(Q,\mathbf{S})$ denote the consistency conditions used in Newey and McFadden's (NM, 1994) Theorem 2.1, when applied to the function $Q(.)$ and the parameter space $\mathbf{S}$. The conditions are: (i) $Q(\mathbf{s})$ is uniquely maximized at $\mathbf{s}^*$; (ii) $\mathbf{S}$ is compact; (iii) $Q(\mathbf{s})$ is continuous; and (iv) $Q_I(\mathbf{s})$ converges uniformly in probability to $Q(\mathbf{s})$.

We will consider $Q(\mathbf{s}_l)$ to be a local maximum if there exists a closed ball $\mathbf{B}_\epsilon(\mathbf{s}_l)$, $\epsilon > 0$, in $\mathbf{S}$ such that $Q(\mathbf{s}_l)$ is a maximum over $\mathbf{B}_\epsilon(\mathbf{s}_l)$. Let $\mathbf{S}_\mathcal{L}$ denote the set of *local* maximizers of $Q(\mathbf{s})$ over the entire space $\mathbf{S}$, and let $\widehat{\mathbf{S}}_\mathcal{L}$ denote the analagous set generated by the sample analog $Q_I(\mathbf{s})$. Let $\mathbf{S}_m$ denote a closed subset of $\mathbf{S}$. Let $\mathbf{s}_m^*$ denote the population maximizer of $Q(\mathbf{s})$ over $\mathbf{S}_m$, and let $\widehat{\mathbf{s}}_m$ denote the sample maximizer of $Q_I(\mathbf{s})$ over $\mathbf{S}_m$. In the proof below, $\mathbf{s}_m^*$ will be the local maximizer of $Q(\mathbf{s})$ that equals the likelihood parameter vector $\mathbf{s}^*$.

Recall that the motivating problem is a lack of identification: even if it were restricted to *global* maximizers, the set $\mathbf{S}_\mathcal{L}$ would have multiple elements. The approach suggested in the text was to pick $\widehat{\mathbf{s}}$ as the element of $\widehat{\mathbf{S}}_\mathcal{L}$ that maximizes $L_I(\mathbf{s})$; we term this the "tiebreaker" estimator. We now show consistency.

Theorem 1

Suppose that: (i) conditions $NM(Q,\mathbf{S}_m)$ hold (local GMM regularity); (ii) $\mathbf{s}_m^*$ lies in the interior of $\mathbf{S}_m \subseteq \mathbf{S}$ (interiority); (iii) conditions $NM(L,\mathbf{S})$ hold (global MLE regularity); (iv) $\ln(g(\mathbf{x};\mathbf{s}_m^*))$ and $\mathbf{S}_m$ satisfy the conditions of Newey and McFadden's Lemma 4.3 (local MLE regularity); (v) $\mathbf{s}_m^* = \mathbf{s}^*$, the maximizer of $L(\mathbf{s})$ (cross-identification); and (vi) $\widehat{\mathbf{s}} = \arg\max_{\{\mathbf{s}\in\widehat{\mathbf{S}}_\mathcal{L}\}} L_I(\mathbf{s})$ (tiebreaking estimate). Then $\widehat{\mathbf{s}} \xrightarrow{\text{p}} \mathbf{s}^*$.

Proof

The proof proceeds in 2 steps: (1) $L_I(\widehat{\mathbf{s}}_m) \xrightarrow{\text{p}} L(\mathbf{s}_m^*)$; and (2) convergence of $L_I(\widehat{\mathbf{s}}_m)$ implies convergence of $\widehat{\mathbf{s}}$. In the interest of brevity, we assume that all measurability conditions are satisfied. (See also the discussion of NM's Theorem 2.1.)

To get step 1, note that by condition (i) and NM's Theorem 2.1, $\widehat{\mathbf{s}}_m \xrightarrow{\text{p}} \mathbf{s}_m^*$. Then it follows from condition (iv) and NM's Lemma 4.3 that $L_I(\widehat{\mathbf{s}}_m) \xrightarrow{\text{p}} L(\mathbf{s}_m^*)$.

It remains to show convergence of $\widehat{\mathbf{s}}$. It follows from consistency of $\widehat{\mathbf{s}}_m$ and condition (ii) that with probability approaching 1 (w.p.a.1) $\widehat{\mathbf{s}}_m$ is in the interior of $\mathbf{S}$ and thus a local maximizer, so that w.p.a.1 $\widehat{\mathbf{s}}_m \in \widehat{\mathbf{S}}_{\mathcal{L}}$. We now proceed as in NM's Theorem 2.1. $\forall \epsilon > 0$, we have w.p.a.1: (1) $L(\widehat{\mathbf{s}}) > L_I(\widehat{\mathbf{s}}) - \epsilon/3$ (from condition (iii)); (2) $L_I(\widehat{\mathbf{s}}) > L_I(\widehat{\mathbf{s}}_m) - \epsilon/3$ (by condition (vi) and $\widehat{\mathbf{s}}_m \in \widehat{\mathbf{S}}_{\mathcal{L}}$); (3) $L_I(\widehat{\mathbf{s}}_m) > L(\mathbf{s}^*_m) - \epsilon/3$ (from step 1). Together these conditions imply that w.p.a.1 $L(\widehat{\mathbf{s}}) > L(\mathbf{s}^*_m) - \epsilon$. But by condition (v) $\mathbf{s}^*_m = \mathbf{s}^*$. It then follows from condition (iii) and arguments in NM's Theorem 2.1 that $\widehat{\mathbf{s}} \xrightarrow{\text{p}} \mathbf{s}^*$. Q.E.D.

Unless one of the local GMM maximizers is also the maximum likelihood estimator, it is essential to include local as well as global maximizers in the set $\widehat{\mathbf{S}}_{\mathcal{L}}$. This can be illustrated with a simple example. Suppose that $\mathbf{S}$ can partitioned into two disjoint compact subsets, $\mathbf{S}_1$ and $\mathbf{S}_2$, and that in addition to satisfying the conditions for Theorem 1, the maximizers of each subset, $\mathbf{s}^*_1$ and $\mathbf{s}^*_2$, are global maximizers of $Q(\mathbf{s})$ over $\mathbf{S}$. Suppose further that $\mathbf{s}^*_1$ is also the MLE maximizer $\mathbf{s}^*$. Finally, suppose that over $\mathbf{S}_1$, $Q_I(\mathbf{s}) = Q(\mathbf{s}) - 1/I$, while over $\mathbf{S}_2$, $Q_I(\mathbf{s}) = Q(\mathbf{s}) + 1/I$. It immediately follows that $\widehat{\mathbf{s}}_1 = \mathbf{s}^*_1$ and $\widehat{\mathbf{s}}_2 = \mathbf{s}^*_2$, and the proof of Theorem 1 goes through. But $Q_I(\widehat{\mathbf{s}}_1) < Q_I(\widehat{\mathbf{s}}_2)$, so that a search over global maxima would exclude $\widehat{\mathbf{s}}_1 = \mathbf{s}^*$.

The conditions for the proof apply naturally to the sequential estimator developed in the main text. $Q(\mathbf{s})$ is the negative of inner product of the expectation vector in equation (11), and Q_I is its sample analog. Condition (v) (cross-identification) follows from the construction of equation (11). Since any solution to equation (11) will be a zero of $Q(\mathbf{s})$, the sequential estimator is a local maximizer of $Q(\mathbf{s})$. One potential difficulty is that, as noted by Wu (1983), some mixture problems lack a compact parameter space.

References

[1] Amemiya, Takeshi, 1978, "On a Two-step Estimation of a Multivariate Logit Model," Journal of Econometrics, 8, pp.13-21.

[2] Arcidiacono, Peter, 2002, "Affirmative Action in Higher Education: How Do Admissions and Financial Aid Rules Affect Future Earnings?" unpublished manuscript.

[3] Cameron, Stephen and James J. Heckman, 1998, "Life Cycle Schooling and Dynamic Selection bias: Models and Evidence for Five Cohorts of American Males," Journal of Political Economy, 106, pp. 262-333.

[4] Cameron, Stephen and James J. Heckman, 2001, "The Dynamics of Educational Attainment for Black, Hispanic, and White Males," Journal of Political Economy, 109, pp. 455-499.

[5] Cox, D.R., 1975, "Partial Likelihood," Biometrika, 62, pp. 269-275.

[6] 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, pp. 1-38.

[7] Eckstein, Zvi and Kenneth Wolpin, 1999, "Why Youths Drop Our of High School: The Impact of Preferences, Opportunities and Abilities," Econometrica, 67, pp 1295-1339.

[8] Efron, Bradley, 1982, "Maximum Likelihood and Decision Theory," Annals of Statistics, 10, pp. 323-339.

[9] Everitt, B.S. and D.J. Hand, 1981, Finite Mixture Distributions, London: Chapman & Hall.

[10] Follmann, Dean A., and Diane Lambert, 1989, "Generalized Logistic Regression by Nonparametric Mixing," Journal of the American Statistical Association, 84, pp. 295-300.

[11] Hamilton, James D., 1989, "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," Econometrica, 57, pp. 357-385.

[12] Hamilton, James D., 1990, "Analysis of Time Series Subject to Changes in Regime," Journal of Econometrics, 45, pp. 39-70.

[13] Hansen, Lars, 1982, "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 50, pp. 1029-1054.

[14] Heckman, J. and B. Singer, 1984, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," Econometrica, 52, pp. 271-320.

[15] Jamshidian, Mortaza, and Robert Jennrich, 1993, "Conjugate Gradient Acceleration of the EM Algorithm," Journal of the American Statistical Association, 88, pp. 221-228.

[16] Jamshidian, Mortaza, and Robert Jennrich, 1997, "Acceleration of the EM Algorithm by using Quasi-Newton Methods," Journal of the Royal Statistical Society, B 59, pp. 569-587.

[17] Keane, Michael and Kenneth Wolpin, 1997, "The Career Decisions of Young Men," Journal of Political Economy, 105, pp. 473-522.

[18] Kiefer, Nicholas, 1980, "A Note on Switching Regressions and Logistic Discrimination," Econometrica, 48, pp. 1065-1069.

[19] Laird, Nan, 1978, "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," Journal of the American Statistical Association, 73, pp. 805-811.

[20] Lindsey, Bruce, 1989, "The Geometry of Mixture Likelihoods: A General Theory" Annals of Statistics, 11, pp. 86-94.

[21] McLachlan, Geoffrey J., and Thriyambakam Krishnan, 1997, The EM Algorithm and Extensions, New York: John Wiley & Sons.

[22] McLachlan, Geoffrey J., and David Peel, 2000, Finite Mixture Models, New York: John Wiley & Sons.

[23] Meng, Xiao-Li and Donald B. Rubin, 1993, "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," Biometrika, 80, pp. 267-278.

[24] Mroz, Thomas A., 1999, "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome," Journal of Econometrics, 92, pp. 233-274.

[25] Newey, Whitney K. and Daniel McFadden, 1994, "Large Sample Estimation and Hypothesis Testing," in R.F. Engle and D.L. McFadden, Eds., Handbook of Econometrics, Volume IV, Amsterdam: Elsevier Science B.V.

[26] Rust, John, 1987, "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," Econometrica, 55, pp. 999-1033.

[27] Rust, John, 1994, "Structural Estimation of Markov Decision Processes," in R.F. Engle and D.L. McFadden, Eds., Handbook of Econometrics, Volume IV, Amsterdam: Elsevier Science B.V.

[28] Rust, John, and Christopher Phelan, 1997, "How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Markets," Econometrica, 65, pp. 781-831.

[29] Ruud, Paul A., 1991, "Extensions of Estimation Methods Using the EM Algorithm," Journal of Econometrics, 49, pp. 305-341.

[30] Titterington, D.M., A.F.M. Smith and U.E. Makov, 1985, Statistical Analysis of Finite Mixture Distributions, New York: John Wiley & Sons.

[31] Wu, C.F., 1983, "On the Convergence Properties of the EM Algorithm," Annals of Statistics, 11, pp. 95-103.