Comment on

"Iterative and Recursive Estimation in Structural Non-Adaptive Models"

by Sergio Pastorello, Valentin Patilea, and Eric Renault

by

Robert P. Sherman

California Institute of Technology

This paper studies observed variables modeled as known functions of unobserved random variables and unknown parameters, and proposes an iterative procedure for estimating the model parameters. The procedure is an Expectation-Maximization, or EM-type procedure in the sense that each iteration involves an expectation of a sample criterion function over unobserved variables given observed variables and current parameter values (the E step) followed by optimization of the criterion function (the M step). As the authors indicate, such indirect procedures can be effective in situations where direct estimation is difficult, as in certain latent regression models and dynamic equilibrium models involving latent state variables. Some asymptotic theory is developed and the procedure is applied to estimate parameters of a term structure model of U.S. monthly zero-coupon interest rates from January 1970 to February 1991. The authors also develop a related recursive procedure which can yield additional computational benefits, at the cost of a loss in statistical efficiency. Some asymptotic theory is developed, but the procedure is not applied to data.

This comment compares and contrasts the asymptotic theory developed by the authors for their iterative procedure with that developed in Dominitz and Sherman (2003) (hereafter DS03) for iterative estimation procedures. This work has developed independently and simultaneously.

As background to this comment, it is useful to note that there already exists a substantial literature on convergence properties of the EM algorithm and its many descendants (see, for example,

McLachlan and Krishnan, 1997). Often in this literature, observed data is conditioned on, and convergence means numerical convergence of sample iterates to a sample fixed point (usually, a type of maximum likelihood estimator) as the number of iterations tends to infinity. Similarly, a rate of convergence refers to how fast sample iterates converge to a sample fixed point.

By contrast, both this paper and DS03 study probabilistic convergence of iterative estimation procedures. Observed data is not conditioned on, and the objective is to establish standard modes of convergence (e.g., rates of convergence, convergence in distribution) of sample iterates to a population fixed point (the parameter of interest) as the number of iterations and the sample size simultaneously tend to infinity. This probabilistic approach is required to do asymptotic inference on the parameter of interest using the sample iterates. In addition, this approach can lead to conclusions not predicted by an analysis of numerical convergence of sample iterates to a sample fixed point. For example, as the sample size increases, one would expect that fewer iterations would be required for sample iterates to converge to within a fixed tolerance of a sample fixed point. However, as discussed in more detail below, in order to achieve a rate of convergence in probability of the sample iterates to the population fixed point, the number of iterations must increase with the sample size in order to control a certain bias term.

We now discuss some of the similarities and differences between the asymptotic theory developed by the authors and that developed in DS03. In order to do so, some notation is required.

Let θ_0 denote a parameter of interest. In the authors' application, $\theta_0 = (k_0, c_0, \sigma_0, \lambda_0)$, where k_0, c_0, and σ_0 are unknown factor dynamics parameters (σ_0 is the volatility parameter) and λ_0 is the unknown risk premium parameter. Let $\hat{Q}(\theta \mid \phi)$ denote a sample criterion function formed in the E step of the EM-type procedure. We let hats suggest dependence on random quantities in the model as well as on the sample size, n. Here, ϕ denotes a current guess at θ_0. (For simplicity,

we do not introduce notation allowing for nuisance parameters.) In the M step of the procedure, $\hat{Q}(\theta \mid \phi)$ is optimized over θ to produce the next iterate, $\hat{\theta}(\phi)$. For example, if the optimization is a maximization, then $\hat{\theta}(\phi) = \underset{\theta}{\text{argmax}}\, \hat{Q}(\theta \mid \phi)$. Let $Q(\theta \mid \phi)$ and $\theta(\phi)$ denote corresponding population analogues. (Typically, $Q(\theta \mid \phi)$ is a uniform probability limit of $\hat{Q}(\theta \mid \phi)$.) Finally, define the sample gradient and sample hessian functions $\hat{G}(\theta \mid \phi) = \frac{\partial}{\partial \theta}\hat{Q}(\theta \mid \phi)$ and $\hat{H}(\theta \mid \phi) = \frac{\partial^2}{\partial \theta^2}\hat{Q}(\theta \mid \phi)$.

Let $\hat{\theta}^0$ denote a starting value for the iterative sequence. For $i \geq 1$, define $\hat{\theta}^i = \hat{\theta}(\hat{\theta}^{i-1})$. The authors develop conditions implying consistency (Proposition 4.4) and asymptotic normality (Proposition 4.8) of the $\hat{\theta}^i$ sequence as i and n simultaneously tend to infinity. DS03 develop conditions implying consistency, rates of convergence, and convergence in distribution of the $\hat{\theta}^i$ sequence.

The principal difference between the authors' approach and the approach in DS03 is the level at which the corresponding sufficient conditions are developed. The authors develop conditions at the level of the sample criterion function $\hat{Q}(\theta \mid \phi)$, whereas DS03 develop conditions on the level of the sample mapping $\hat{\theta}(\phi)$ used to define the sequence of iterates.

The former approach is natural in the context of applications where each iteration involves an optimization of an sample criterion function where the corresponding sample mapping does not have a closed form. This is the context of interest to the authors. Their main applications involve either likelihood-based or GMM-based sample criterion functions with corresponding sample mappings without closed forms.

The latter approach is more general in that it does not require that sample iterates be generated through optimization. Moreover, it covers the applications considered by the authors by taking $\hat{\theta}(\phi) = \underset{\theta}{\text{argmax}}\, \hat{Q}(\theta \mid \phi) = \theta(\phi) - [\hat{H}(\theta^*(\phi) \mid \phi)]^{-1}\hat{G}(\theta(\phi) \mid \phi)$, where $\theta^*(\phi)$ lies between $\hat{\theta}(\phi)$ and $\theta(\phi)$. (The second representation for $\hat{\theta}(\phi)$ is derived from a Taylor expansion of the sample gradient

function $\hat{G}(\hat{\theta}(\phi) \mid \phi)$ about $\theta(\phi)$, after noting that $\hat{G}(\hat{\theta}(\phi) \mid \phi) = 0$.) This latter approach is more natural in applications where $\hat{\theta}(\phi)$ has a closed form, as in the iterative least squares and Newton-Raphson applications considered in DS03. It is also more natural in applications where the presence of nuisance parameters leads to more than one optimization per iteration, as in iterative estimation of the parameters of $AR(p)$ models, where each iteration involves an optimization over regression parameters followed by an optimization over the p autoregression parameters. Similarly, in iterative estimation of the parameters of parametric and semiparametric censored regression models, each iteration involves an optimization over regression parameters followed by an optimization over the parameter denoting the standard deviation of the error term in the regression.

The objective of asymptotic theory for iterative estimators is to determine conditions under which the sequence of sample iterates $\hat{\theta}^i$ converges under various modes to the parameter of interest θ_0. The DS03 approach develops conditions on the sample mapping $\hat{\theta}(\phi)$ that defines the sequence of iterates, and so is more direct than the approach of the authors. This direct approach makes it easier to see precisely what drives the asymptotic behavior of the $\hat{\theta}^i$ sequence.

To see this, recall that $\hat{\theta}^0$ is the starting point of the sequence of sample iterates $\hat{\theta}^i$, $i \geq 1$. Let us define a population sequence with the same starting point. That is, let $\theta^0 = \hat{\theta}^0$ and, for $i \geq 1$, define population iterates $\theta^i = \theta(\theta^{i-1})$. If the population mapping $\theta(\phi)$ is a contraction mapping with fixed point θ_0, then a standard fixed point theorem guarantees that $\theta^i \to \theta_0$ as $i \to \infty$. If, in addition, $\hat{\theta}(\phi)$ is uniformly close to $\theta(\phi)$ in some stochastic sense, one would expect the sample iterates to be close to their population counterparts in the same sense, and so converge to θ_0 in this sense.

This simple intuition is formalized in Theorem 1 of DS03. This theorem says that if (i) $\theta(\phi)$ is an asymptotic contraction mapping with fixed point θ_0 and (ii) $\hat{\theta}(\phi)$ converges uniformly in

probability to $\theta(\phi)$ at rate n^δ, $\delta > 0$, then $\hat{\theta}^i$ converges in probability to θ_0 at rate n^δ as both i and n tend to infinity in a suitable way. The contraction mapping condition and the uniform convergence condition are the needle and thread, so to speak, used to stitch together the sample and population sequences.

The proof of Theorem 1 in DS03 is instructive. It is based on a decomposition of the difference between $\hat{\theta}^i$ and θ_0 into a stochastic term and a bias term:

$$|\hat{\theta}^i - \theta_0| \leq |\hat{\theta}^i - \theta^i| + |\theta^i - \theta_0| \,.$$

Condition (ii) controls the stochastic term through a recursive argument while condition (i) controls the bias term. It follows immediately that if the difference $\hat{\theta}^i - \theta_0$ is multiplied by n^δ, then to prevent the bias term from diverging, the iteration number i must increase at an asymptotic rate no slower than $-\delta \log n / \log c$, where $c \in [0, 1)$ is the modulus of contraction of the mapping $\theta(\phi)$. If the modulus of contraction is not known, then $i \geq n^\alpha$ for any $\alpha > 0$ will suffice. In any event, as mentioned earlier, the number of iterations i must increase as a function of the sample size n in the sense just described in order that $\hat{\theta}^i$ converge in probability to θ_0 at rate n^δ.

In order to establish the asymptotic distribution of the $\hat{\theta}^i$ sequence, DS03 require that the sample mapping $\hat{\theta}(\phi)$ also be an asymptotic contraction mapping. Their Lemma 2 provides three simple, checkable primitive conditions for this to hold: (i) $\theta(\phi)$ is an asymptotic contraction mapping with fixed point θ_0 (ii) $\hat{\theta}(\phi)$ converges uniformly in probability to $\theta(\phi)$ and (iii) $\frac{\partial}{\partial \phi}\hat{\theta}(\phi)$ converges uniformly in probability to $\frac{\partial}{\partial \phi}\theta(\phi)$. The proof is elementary.

Theorem 3 in DS03 provides sufficient conditions for establishing the limiting distribution of the $\hat{\theta}^i$ sequence. The key requirements are (i) $\hat{\theta}(\phi)$ is as asymptotic contraction mapping and (ii)

convergence in distribution of the infeasible estimator $\hat{\theta}(\theta_0)$. The latter condition corresponds to Assumption 4.7 in the authors' paper.

In sum, developing conditions at the level of the sample mapping $\hat{\theta}(\phi)$ can lead to a simpler, more transparent asymptotic theory of iterative estimation. As is clear from the discussion above, the centerpiece of this theory is the requirement that the population mapping $\theta(\phi)$ be an asymptotic contraction mapping. In order for this theory to be confidently applied to do asymptotic inference, this contraction mapping condition must either be formally established or at the very least, informally checked for the application at hand. Note that the authors also require a condition like this as part of the theory they develop (Assumption 4.3). It would be very useful to know whether this condition can be formally established or informally checked for the application that the authors consider in Section 6.

As a final note, it is instructive to compare standard sufficient conditions for consistency and convergence in distribution of an optimization estimator with standard sufficient conditions for consistency and convergence in distribution of an iterative estimator.

Start with consistency (or rates of consistency). Standard conditions for consistency of an optimization estimator are (i) global uniform convergence of the sample objective function to a population objective function in an appropriate stochastic sense and (ii) a strong optimization condition on the population objective function that guarantees identification of the parameter of interest (e.g., Amemiya 1985, Theorem 4.1.1). (Here, global uniform convergence means uniform convergence over the entire parameter space). As discussed above, standard conditions for consistency of an iterative estimator are (i$'$) global uniform convergence of the sample mapping $\hat{\theta}(\phi)$ to a population mapping $\theta(\phi)$ in an appropriate stochastic sense and (ii$'$) an asymptotic contraction mapping condition on $\theta(\phi)$ to guarantee identification. Thus, consistency of an iterative estimator

requires global uniform convergence one level deeper than the level of global convergence required for consistency of an optimization estimator. Similarly, the identification condition for an iterative estimator is stronger than the corresponding condition for an optimization estimator. These stronger conditions are needed to string together a sequence of estimators.

Consider convergence in distribution. For an optimization estimator, in addition to the standard consistency conditions (i) and (ii) just mentioned, one requires (iii) local uniform convergence of the sample Hessian function to its population counterpart in an appropriate stochastic sense and (iv) convergence in distribution of the sample gradient function evaluated at the parameter of interest. (Here, local uniform convergence means convergence uniform over shrinking neighborhoods of the parameter of interest.) For an iterative estimator, in addition to the standard consistency conditions (i′) and (ii′), one requires (iii′) global uniform convergence of $\frac{\partial}{\partial\phi}\hat{\theta}(\phi)$ to $\frac{\partial}{\partial\phi}\theta(\phi)$ in an appropriate stochastic sense and (iv′) convergence in distribution of the infeasible estimator $\hat{\theta}(\theta_0)$. As with consistency, convergence in distribution of an iterative estimator requires analogous, but stronger, conditions than convergence in distribution of an optimization estimator.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics.* Harvard Univ. Press, Cambridge, Mass.

Dominitz, J. and R. P. Sherman (2003): "Some Convergence Theory for Iterative Estimation Procedures," under review, *Econometric Theory.*

McLachlan, G. J. and T. Krishnan (1997): *The EM Algorithm and Extensions.* New York: Wiley.