

Efficient Sequential Extremum Estimation and a Comparison with Maximization by Parts

David Frazier

February 6, 2012

Abstract

This research consider the problem of efficiently estimating a parameter of interest when the model is complicated by a vector of nuisance parameters. If the model is nonadaptive we must often resort to full information estimation to gain an efficient estimator for the parameter of interest. In certain cases full information estimation can be computationally intensive and lead to poor finite sample properties, Fan, Pastorello and Renault(2012)[9]. To avoid such complications Fan, Pastorello and Renault(2012)[9](FPR hereafter) derive algorithms, known as maximization by parts(MBP), which yield iterative estimators for the parameter of interest that converge to the full information estimates as the number of iterations goes to infinity. However, these iterative estimators are only applicable when a set of technical conditions are satisfied.

As an alternative to the MBP algorithms we derive consistent and efficient sequential extremum estimators for the parameter of interest. Unlike MBP these sequential estimators only rely on a standard set of regularity conditions, which are common across extremum estimators. To compare the computational cost of the sequential and iterative estimators we derive Newton-Raphson(NR) updating rules for the sequential estimators and compare these to the known NR updating rules for the iterative estimators. We show that within the specific case of separable log-likelihood functions the sequential estimators are strictly preferable as they achieve consistency and efficiency in two steps whereas the iterative estimators are only consistent after two steps. We demonstrate the applicability of this method by applying the sequential estimators to the stochastic volatility model of Taylor(1994)[26] and the affine term structure models of Dai and Singleton(2000)[7]

Applying the sequential estimation methodology to the generalized method of moments(GMM) leads to a dual representation of the sequential GMM estimator as a minimum chi-squared estimator. Thus, if we use GMM as a unifying estimation framework the sequential estimators have a dual representation as minimum chi-squared estimators.

1 Introduction and Literature Review

The goal of this analysis is to derive a unified framework for sequential estimation of the parameter of interest within nonadaptive models. Many economic models are characterized by the issue of nonadaptivity. Perhaps the best known example of a nonadaptive model comes from simple linear regression with two regressors.

$$\begin{aligned}y_t &= X_t\beta_0 + \epsilon \\X_t &= (X_{1t}, X_{2t}) \\ \beta_0 &= (\theta_0, \nu_0)' \\ E[y_t - X_{1t}\theta_0 - X_{2t}\nu_0 \mid X_{1t}, X_{2t}] &= 0\end{aligned}$$

with X_{1t}, X_{2t} and θ_0, ν_0 all scalars. Furthermore, assume θ_0 is the parameter of interest and ν_0 is a nuisance parameter. In this case we may estimate θ_0 by least squares. The sample objective function is given by

$$Q_T[\theta, \nu] = \frac{1}{T} \sum_{t=1}^T (y_t - X_{1t}\theta - X_{2t}\nu)^2.$$

If we are to base the estimator of θ_0 on the first order conditions of $Q_T[\theta, \nu]$, the resulting estimator would depend on the nuisance parameter ν . This is the general issue of nonadaptivity.

However, in this instance the Frisch Waugh theorem delivers a means of dealing with this dependence. Using the Frisch Waugh theorem we can derive an equation for ν that can be used within the calculation of our estimator for θ_0 . This equation is given by,

$$\nu(\theta) = \left(\frac{1}{T} \sum_{t=1}^T X_{2t}X_{2t}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T y_t - X_{1t}\theta \right).$$

Our objective function now becomes,

$$Q_T[\theta, \nu(\theta)] = \frac{1}{T} \sum_{t=1}^T (y_t - X_{1t}\theta - X_{2t}\nu(\theta))^2.$$

Maximizing $Q_T[\theta, \nu(\theta)]$ with respect to θ leads to a simple closed form solution and an estimator of θ_0 which is efficient and consistent.

Unfortunately, linear regression is too restrictive to be useful in many economic settings. In many economic models gaining estimators for the parameter of interest requires the optimization of a sample criterion function.

Many different types of estimators are defined as the solution of an optimization problem based on sample objective function $Q_T[\theta, \nu]$. Such estimators are generally referred

to as extremum estimators. The objective function, $Q_T[\theta, \nu]$, is dependent on the p dimensional parameter of interest θ and a $r - p$ dimensional nuisance parameter ν . In certain cases, such as structural nonadaptive models(see Pastorello, Patilea and Renault(2003)(PPR hereafter)[22]) we can characterize the parameter vector ν as a known function of the parameters θ . Just as in the linear regression case we denote this function as $\nu(\theta)$ ¹. The sample objective function can then be stated generally as $Q_T[\theta, \nu(\theta)]$. Objective functions with this structure arise in many examples within economics and finance², leading examples come from empirical estimation of asset pricing models.

If we use the relationship $\nu(\theta)$ within estimation the parameter of interest becomes a p -dimensional unknown parameter, $\theta_0 \in \Theta \subset \mathfrak{R}^p$, where θ_0 is the argument maximizer of the corresponding population objective function. The extremum estimator of θ_0 , denoted $\hat{\theta}_T$, is defined as:

$$\hat{\theta}_T = \arg \max_{\theta \in \Theta} Q_T[\theta, \nu(\theta)] \quad (1.1)$$

for some known function $\nu(\theta)$ and a given criterion function

$$Q_T[\theta, \nu(\theta)] = Q_{(T)}[\theta, \nu, (X_t)_{1 \leq t \leq T}],$$

associated with observations $(X_t)_{1 \leq t \leq T}$.

Often the objective function is numerically cumbersome to maximize with respect to the second occurrence of θ or this occurrence can be the source of some singularity within the objective function, Fan, Pastorello and Renault(2012)(FPR hereafter)[9]. In this case obtaining consistent and efficient parameter estimates by solving (1.1) may be difficult, if not impossible³.

In general this extremum estimator satisfies the following first order conditions

$$\left[\frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \theta'} + \frac{\partial \nu(\theta)'}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu'} \right] = 0. \quad (1.2)$$

Solving the first order conditions with conventional Gauss-Newton type algorithms can be computationally cumbersome due to the second term within equation (1.2). Recently, methods have been devised to avoid maximization with respect to the second occurrence of θ .

1.1 Potential Solutions

FPR derive iterative algorithms, collectively referred to as maximization by parts(MBP), for extremum estimators by solving different version of the first order conditions 1.2. The algorithms of FPR contain as special cases the iterative estimator of PPR and Song, Fan

¹Note that this characterization is always possible given that we may concentrate out the parameters ν to achieve such a relationship.

²PPR and Following Fan, Pastorello and Renault(2011), the assumption that the function $\nu(\cdot)$ is known and not dependent on the sample is not restrictive since we can incorporate this dependence into the sample objective function $Q_T[\theta, \nu(\theta)]$

³Such cases can be encountered within the affine term structure models of Dai and Singleton(2000)

and Kalfebeish(2005)(SFK hereafter)[24]. These estimators avoid directly maximizing the more cumbersome portions of the first order conditions, namely $\nu(\theta)$, by iteratively solving equation (1.2). This strategy proves useful when there exists singularities within the Hessian matrix or when computationally cumbersome portions of the objective function can be attributed to the parameters $\nu(\cdot)$ ⁴.

However, there is no such thing as a free lunch. Even if the iterative estimators begin from an initially consistent starting point, certain technical conditions, known formally as information dominance conditions, must hold for these estimators to be efficient. These conditions require that, under the true data generating process, the portions of the Hessian not used within estimation must be negligible when compared to the portions of the Hessian which are used within estimation. In practice, these conditions are impossible to verify. To date, there is no evidence to show how severely these estimators are affected when the information dominance conditions are not satisfied. While the iterative estimators are computationally attractive the imposition of unverifiable technical assumptions may not be palatable for some researchers.

As an alternative, this research derives efficient and consistent sequential extremum estimators which rely only on a standard set of regularity conditions employed in extremum estimation. These new sequential estimators extend existing sequential estimators in two important directions. First, unlike existing sequential estimators the sequential estimators derived within this research are applicable when the nuisance parameters are related to the parameter of interest through a structural relationship within the underlying economic model and when the nuisance parameters have been “profiled” or “concentrated” out of the objective function, see Amemiya(1985). Secondly, we derive general Newton-Raphson(NR) updating rules for the sequential estimators to gage their computational burden. These NR updating rules can then be employed in the creation of computationally light numerical algorithms which yield estimators that are asymptotically equivalent to the full information estimators. Using these NR updating rules we are able to give a general comparison between the sequential estimators and the iterative estimators of Fan, Pastorello and Renault(2012)[9]⁵.

When the objective function is of an additively separable form, we show that, the iterative estimators achieve consistency after two steps. The sequential estimators, on the other hand, will deliver an efficient and consistent estimator in two steps. To achieve efficiency the iterative estimators must iterate a large number of times⁶. Lastly, we show that many of the sequential estimators have a dual representation as minimum chi-squared estimators.

⁴In most cases singularities within the objective function will render many Gauss-Newton type maximization algorithms useless.

⁵It will be made more precise later that the iterative estimators are an alternative method for dealing with the presence of nuisance parameters.

⁶Technically they must iterate $k > T^\delta$ times, where δ is some positive constant less than one.

1.2 Review of Literature

There is a long standing history of sequential estimators in economics and econometrics⁷. The earliest work on sequential estimators, focused on adaptive estimation and comes from Harvey(1976)[11] and Amemiya(1978)[25]. The general article detailing conditions under which adaptive estimation can take place in likelihood settings comes from Pagan(1986)[21]. For GMM and minimum distance estimators Newey and McFadden(1994)[19], chapter 5, detail conditions for adaptive estimation of parameters. Further references for adaptive sequential estimation are detailed within Appendix 8.

This research differs from these earlier works in that, we provide efficient estimates not by deriving conditions under which adaptive estimation can occur but by altering the objective function in such a way as to recover efficiency for the parameter of interest. This approach is non-standard and has received little attention within the general literature on sequential estimation. The main references in this vein of research are Trognon and Gouriéroux(1990)(TG hereafter)[27] in an extremum estimator setting, Gouriéroux, Monfort and Renault(1996)(GMR hereafter)[10] and Crepon et al.(1997)[6] in a GMM setting. In essence this research takes elements from each of these papers and combines them to derive a general method which is capable of solving not only the specific problems considered within these references but also more general problems. Furthermore, we show that in nonadaptive models the sequential estimators can have better finite sample properties than their full information counterparts^{8,9}.

The remainder of the paper is organized as follows. Section two derives the efficient sequential estimators and proves the asymptotic properties of these estimators. Section three compares these methods with existing iterative estimators. Within this section we derive Newton-Raphson updating rules for the sequential estimators and employ these rules to compare the computational cost of the sequential and iterative estimators in a general setting. For a more specific comparison we analyze both estimators when the objective function is given by a separable log-likelihood function. Section four shows that we can recast the sequential extremum estimators as minimum chi-squared estimators. Section five contains two applications for the sequential estimators and determines their finite sample behavior in these applications. Lastly, section six concludes. All proofs are relegated to the appendix.

2 Efficient Sequential Estimation

If the objective function is given by $Q_T[\theta, \nu(\theta)]$, the resulting first order conditions are

$$\left[\frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \theta'} + \frac{\partial \nu(\theta)'}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu'} \right] = 0.$$

⁷Many examples have been included within Appendix 8.

⁸A similar result was found in Hoffman(1991)[13] for adaptive sequential estimation

⁹We concentrate this comparison on small samples since asymptotically the sequential estimators are equivalent to the full information estimators.

Numerically solving these first order conditions can be difficult using standard Gauss-Newton type algorithms. These difficulties are often due to the inclusion of the second term within the first order conditions. Given this fact, we can alleviate the computational difficulties of solving (1.1) by using a preliminary consistent estimator $\tilde{\theta}_T$ and solving

$$\max_{\theta \in \Theta} Q_T[\theta, \nu(\tilde{\theta}_T)].$$

However, the resulting estimator will be inefficient. To see this fact consider the score of $Q_T[\theta, \nu(\tilde{\theta}_T)]$ with respect to θ ,

$$\frac{\partial Q_T(\theta, \nu(\tilde{\theta}_T))}{\partial \theta} = 0.$$

Clearly any estimator based on these conditions will be inefficient since we no longer use the information contained in the second term within the first order conditions. Even if such a strategy yields a consistent estimator the asymptotic variance will be inflated since we no longer include the terms

$$\lim_{T \rightarrow \infty} \left\{ \frac{\partial}{\partial \theta} \left[\frac{\partial \nu(\theta)'}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu'} \right] \right\} \Big|_{\theta=\theta_0}, \quad (2.1)$$

within the Hessian matrix.

An additional complication of implementing such a strategy arises from the fact that the sampling distribution of the second stage estimator

$$\hat{\theta}_T = \arg \max_{\theta} Q_T[\theta, \nu(\tilde{\theta}_T)]$$

is dependent on the sampling distribution of $\tilde{\theta}_T$ ¹⁰. If we wish to derive a consistent and efficient sequential estimator we must find a way to regain the information lost in the Hessian matrix, given by equation 2.1, and alleviate the dependence on the initial estimator $\tilde{\theta}_T$.

To circumvent these problems, PPR, SFK and FPR derive algorithms which iterate on various portions of the first order conditions (1.2). Under technical conditions, these algorithms yield an efficient estimator as the number of iterations goes to infinity. Instead of resorting to iterative estimators, we devise an efficient sequential estimation method to solve this problem.

To motivate our efficient sequential estimation method we ask the question: can $Q_T[\theta, \nu(\tilde{\theta}_T)]$ be modified so that we receive an estimator of θ which is as efficient as $\hat{\theta}_T$ and which is less computationally taxing to derive? To answer this question in the affirmative the sequential estimator must alleviate the two issues mentioned previously¹¹. To see how this is possible, consider the second order expansion of the objective function with respect to the second

¹⁰This fact comes from the nonadaptivity of the model.

¹¹These issues were the loss of terms within the Hessian matrix and the dependence on the sampling distribution of the initial estimator.

occurrence of θ evaluated at the point $\tilde{\theta}_T$:

$$Q_T[\theta, \nu(\theta)] \simeq Q_T[\theta, \nu(\tilde{\theta}_T)] + (\theta - \tilde{\theta}_T)' \frac{\partial Q_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \nu'} \frac{\partial \nu(\tilde{\theta}_T)'}{\partial \theta} + \frac{1}{2}(\theta - \tilde{\theta}_T)' \left[\sum_{i=1}^{\dim(\nu)} \frac{\partial Q_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \nu_i} \frac{\partial^2 \nu_i(\tilde{\theta}_T)}{\partial \theta \partial \theta'} + \frac{\partial \nu(\tilde{\theta}_T)'}{\partial \theta'} \frac{\partial^2 Q_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'} \right] (\theta - \tilde{\theta}_T)$$

Using this expression as a basis for our objective function will allow us to regain the portions of the Hessian matrix that were lost by implementing the sequential estimation procedure. However, we can not determine simply by looking at this expression if this objective function will lead to an estimator which is not dependent on the sampling distribution of $\tilde{\theta}_T$.

The result below shows that maximizing a form of the above expansion with respect to θ will allow the sampling distribution for the resulting estimator to be independent of the preliminary estimator $\tilde{\theta}_T$. To state this result define,

$$\Phi_T[\theta, \nu(\tilde{\theta}_T)] = Q_T[\theta, \nu(\tilde{\theta}_T)] + (\theta - \tilde{\theta}_T)' \frac{\partial Q_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \nu'} \frac{\partial \nu(\tilde{\theta}_T)'}{\partial \theta} - \frac{1}{2}(\theta - \tilde{\theta}_T)' T B_T(\theta) (\theta - \tilde{\theta}_T) \quad (2.2)$$

where

$$p \lim_{T \rightarrow \infty} \left[B_T(\theta_0) + \frac{1}{T} \left(\sum_{i=1}^{\dim(\nu)} \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu_i} \frac{\partial^2 \nu_i(\theta)}{\partial \theta \partial \theta'} + \frac{\partial \nu(\theta_0)'}{\partial \theta} \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \right) \right] = 0.$$

With these definitions we can state the following result:

Theorem 1. *Under normal regularity conditions for extremum estimators, if $\sqrt{T}(\tilde{\theta}_T - \theta_0) \rightarrow \mathcal{N}(0, S)$ and if $\Phi_T[\theta, \nu(\tilde{\theta}_T)]$ is given by (2.2), then*

$$\hat{\theta}_T = \arg \max_{\theta} Q_T[\theta, \nu(\theta)]$$

and

$$\theta_T^* = \arg \max_{\theta} \Phi_T[\theta, \nu(\tilde{\theta}_T)]$$

are asymptotically equivalent. That is $\sqrt{T}(\theta_T^* - \hat{\theta}_T) = o_p(1)$

It is important to note that under our assumptions $B_T(\tilde{\theta}_T)$ is a consistent estimator for

$$p \lim_{T \rightarrow \infty} \frac{1}{T} \left(\sum_{i=1}^{\dim(\nu)} \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu_i} \frac{\partial^2 \nu_i(\theta)}{\partial \theta \partial \theta'} + \frac{\partial \nu(\theta_0)'}{\partial \theta} \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \right).$$

With this fact $B_T(\theta)$ can be replaced by $B_T(\tilde{\theta}_T)$ in the previous result.

This demonstrates that we can gain simple efficient sequential estimators by altering the objective function in a specific way. A second important point is that the sequential estimator only relies on the standard assumptions used in extremum estimation.

3 Comparison With Existing Methods: Maximization by Parts(MBP)

To gain efficient and consistent estimators the MBP algorithms exploit a set of technical conditions, known collectively as information dominance conditions. Only when the information dominance conditions are satisfied¹² and after a large number of iterations will the iterative estimators be consistent and efficient. Later in this section, within the confines of a specific example, we will give a detailed description of these information dominance conditions. For a general discussion of information dominance conditions the reader is referred to SFK and FPR.

The next section conducts a general comparison between the sequential and iterative estimators. To facilitate such a comparison we state NR updating rules for each estimator. These updating rules yield equivalent estimators for each method, while ensuring that the estimators have a common structure. As a specific example we compare the sequential and iterative estimators when the objective function is of an additively separable form. In this case we can conclude that the sequential estimator should be strictly preferable to the iterative estimators on the grounds of computation and assumptions.

3.1 General Comparison

Since both methods achieve efficiency, the main difference between the two methods should be one of computation and assumptions. Besides the standard assumptions for extremum estimators the iterative estimators require auxiliary technical assumptions that may be impossible to verify in practice. We will detail these assumptions in the specific case of a separable objective function in the next subsection.

The computational issues are compared by analyzing the NR updating rules for the iterative and sequential estimators. To give a general comparison of these estimators we detail the first two MBP algorithms.

Define the score equation $\frac{\partial L_T(\theta, \nu(\theta))}{\partial \theta}$ as,

$$\frac{\partial L_T(\theta, \nu(\theta))}{\partial \theta} = \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} + \frac{\partial \nu(\theta)'}{\partial \theta} \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \nu'}$$

Algorithm I

Step 1: We start the algorithm from an initially consistent estimator denoted as $\hat{\theta}^{(0)}$

Step k: Let $\hat{\theta}^k$ solve ($k = 1, 2, 3, \dots$) :

$$\frac{\partial Q_T(\theta, \nu(\hat{\theta}^{(k-1)}))'}{\partial \theta} = - \frac{\partial \nu(\hat{\theta}^{(k-1)})'}{\partial \theta} \frac{\partial Q_T(\hat{\theta}^{(k-1)}, \nu(\hat{\theta}^{(k-1)}))'}{\partial \nu'} \tag{3.1}$$

¹²These conditions must be satisfied at the true parameter value and hence the econometrician is uncertain if and when these conditions are ever truly satisfied.

Define

$$\Sigma_T(\hat{\theta}^{(k-1)}) = \frac{\partial^2 Q_T(\hat{\theta}^{(k-1)}, \nu(\hat{\theta}^{(k-1)}))}{\partial \theta \partial \theta'}.$$

We then have the alternative NR updating rule,

Step k': Let $\hat{\theta}^k$ solve ($k = 1, 2, 3, \dots$) :

$$\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} - \left[\Sigma_T(\hat{\theta}^{(k-1)}) \right]^{-1} \left[\frac{\partial L_T(\hat{\theta}^{(k-1)})}{\partial \theta} \right]. \quad (3.2)$$

Algorithm II

Step 1: We start the algorithm from an initially consistent estimator denoted as $\hat{\theta}^{(0)}$

Step k: Let $\hat{\theta}^k$ solve ($k = 1, 2, 3, \dots$) :

$$\frac{\partial Q_T(\theta, \nu(\theta))'}{\partial \theta} = - \frac{\partial \nu(\hat{\theta}^{(k-1)})'}{\partial \theta} \frac{\partial Q_T(\hat{\theta}^{(k-1)}, \nu(\hat{\theta}^{(k-1)}))'}{\partial \nu}. \quad (3.3)$$

Let

$$H_T(\hat{\theta}^{(k-1)}) = \frac{\partial^2 Q_T(\hat{\theta}^{(k-1)}, \nu(\hat{\theta}^{(k-1)}))}{\partial \theta \partial \nu'} \frac{\partial \nu(\hat{\theta}^{(k-1)})}{\partial \theta'}.$$

We then have the alternative NR updating rule,

Step k': Let $\hat{\theta}^k$ solve ($k = 1, 2, 3, \dots$) :

$$\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} - \left[\Sigma_T(\hat{\theta}^{(k-1)}) + H_T(\hat{\theta}^{(k-1)}) \right]^{-1} \left[\frac{\partial L_T(\hat{\theta}^{(k-1)})}{\partial \theta} \right] \quad (3.4)$$

It is simple to see that the only difference between the two algorithms is the number of free parameters that must be updated at each step.

To facilitate a comparison between the iterative and sequential estimators, we can state a NR updating rule for the sequential estimator. To this end, define:

$$\frac{\partial \bar{L}_T(\theta, \nu(\hat{\theta}^{(0)}))}{\partial \theta} = \frac{\partial L_T(\theta, \nu(\hat{\theta}^{(0)}))}{\partial \theta} + \frac{\partial \nu(\hat{\theta}^{(0)})'}{\partial \theta} \frac{\partial^2 Q_T(\hat{\theta}^{(0)}, \nu(\hat{\theta}^{(0)}))}{\partial \nu \partial \theta'} (\theta - \hat{\theta}^{(0)}) - T \hat{B}_T(\theta - \hat{\theta}^{(0)}). \quad (3.5)$$

The NR updating rule for the sequential estimator is then given by,

NR: Sequential Estimator

$$\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} - \left[D^2 Q_T(\hat{\theta}^{(0)}, \nu(\hat{\theta}^{(0)})) \right]^{-1} \frac{\partial \bar{L}_T(\hat{\theta}^{(k-1)}, \nu(\hat{\theta}^{(0)}))}{\partial \theta} \quad (3.6)$$

where $D^2 Q_T$ represents the full Hessian matrix.

We see that updating rule (3.6) makes use of the entire Hessian matrix whereas the iterative estimator only requires updating portions of the Hessian matrix¹³. While the updating rule for the sequential estimator does evaluate the entire Hessian matrix, the evaluation of the Hessian must only be done once.

It is also important to note that while the sequential estimator may seem more intensive to update, given the extra two terms in equation 3.5, in actuality it is not much more difficult to update than the iterative estimators. This fact can be seen by noting that the additional terms in equation 3.5 are both the same linear function(up to a constant) and therefore we are only required to evaluate one additional linear function.

3.2 Separable Likelihood Functions

We now compare both estimators when the objective function $Q_T[\theta, \nu(\theta)]$ is an additively separable log-likelihood function:

$$Q_T[\theta, \nu(\theta)] = Q_{1T}(\theta_1) + Q_{2T}(\nu(\theta_1), \theta_2), \quad (3.7)$$

$\theta = (\theta'_1, \theta'_2)'$. In this context the iterative and sequential estimators have a simplified representation. These simplified representations will allow us to analyze the inefficiencies of the iterative estimators, and at the same time illuminate why the sequential estimators achieve efficiency.

3.2.1 Sequential Estimators

When the objective function takes the form of equation 3.7 the full information first order conditions are given by:

$$\frac{\partial Q_{1T}(\theta_1)}{\partial \theta_1} + \frac{\partial \nu(\theta_1)'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\theta_1), \theta_2)}{\partial \nu} = 0 \quad (3.8)$$

$$\frac{\partial Q_{2T}(\nu(\theta_1), \theta_2)}{\partial \theta_2} = 0. \quad (3.9)$$

Within the context of separable likelihood functions it is often assumed that the equations

$$\begin{aligned} \frac{\partial Q_{1T}(\theta_1)}{\partial \theta_1} &= 0 \\ \frac{\partial Q_T(\nu(\theta_1), \theta_2)}{\partial \theta_2} &= 0 \end{aligned}$$

yield a consistent estimator for $\theta_0 = (\theta'_{10}, \theta'_{20})'$. We can then define the following consistent estimators

$$\tilde{\theta}_{1T} = \arg \max_{\theta_1} Q_{1T}(\theta_1) \quad (3.10)$$

$$\tilde{\theta}_{2T} = \arg \max_{\theta_2} Q_{2T}(\nu(\tilde{\theta}_{1T}), \theta_2). \quad (3.11)$$

¹³The decreased computational burden of the iterative estimators comes from the fact that the entire Hessian matrix is not updated within each iteration.

When this assumption is true we may view the second set of equations in 3.8 as additional moment conditions which are used to gain efficiency. In this sense, equations 3.8 simply details the optimal way of combining these moment conditions. The sequential estimator can then be thought of as combining the efficient moment conditions in a computationally light way.

We now state an algorithm for deriving the sequential estimator when the objective function is a separable log-likelihood. Define,

$$\hat{B}_T = \frac{\partial^2 Q_{1T}(\tilde{\theta}_{1T})}{\partial \theta_1 \partial \theta_1'} + \frac{\partial \nu(\tilde{\theta}_{1T})}{\partial \theta_1} \frac{\partial^2 Q_{2T}(\nu(\tilde{\theta}_{1T}), \tilde{\theta}_{2T})}{\partial \nu \partial \nu} \frac{\partial \nu(\tilde{\theta}_{1T})'}{\partial \theta_1} + \sum_{i=1}^{\dim(\nu)} \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \tilde{\theta}_{2T})}{\partial \nu_i} \frac{\partial^2 \nu_i(\tilde{\theta}_{1T})}{\partial \theta_1 \partial \theta_1'}. \quad (3.12)$$

$$\Pi(\tilde{\theta}_{1T}, \theta_2) = Q_{2T}(\nu(\tilde{\theta}_{1T}), \theta_2) - \frac{1}{2} \left[\frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \theta_2)}{\partial \nu_i} \frac{\partial \nu(\tilde{\theta}_{1T})}{\partial \theta_1} \right] \hat{B}_T^{-1} \left[\frac{\partial \nu(\tilde{\theta}_{1T})'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \theta_2)}{\partial \nu} \right] \quad (3.13)$$

Sequential Algorithm

Step 1: Compute $\tilde{\theta}_{1T}$ and $\tilde{\theta}_{2T}$ from equations 3.10, 3.11.

Step 2: Compute $\theta_{1T}^*, \theta_{2T}^*$ from the following equations

$$\theta_{2T}^* = \arg \max_{\theta_2} \Pi(\tilde{\theta}_{1T}, \theta_2) \quad (3.14)$$

$$\theta_{1T}^* = \tilde{\theta}_{1T} - \hat{B}_T^{-1} \left[\frac{\partial Q_{1T}(\tilde{\theta}_{1T})}{\partial \theta_1} + \frac{\partial \nu(\tilde{\theta}_{1T})'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \theta_{2T}^*)}{\partial \nu} \right] \quad (3.15)$$

where \hat{B}_T and $\Pi(\tilde{\theta}_{1T}, \theta_2)$ were defined in equations 3.12 and 3.13 respectively.

The estimator defined by equation 3.14, and thus the entire vector θ_T^* , can be stated using a NR updating rule. We can then easily compare these updating rules with the NR rules for iterative estimators given in the next section.

Define,

$$\begin{aligned} \psi_{11}(\tilde{\theta}_T) &= \hat{B}_T & \psi_{12}(\tilde{\theta}_T) &= \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \theta_2)}{\partial \theta_2 \partial \nu} \frac{\partial \nu(\tilde{\theta}_{1T})}{\partial \theta_1} \\ \psi_{21}(\tilde{\theta}_T) &= \psi_{12}(\tilde{\theta}_T)' & \psi_{22}(\tilde{\theta}_T) &= \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \tilde{\theta}_{2T})}{\partial \theta_2 \partial \theta_2} \\ \psi_T &= \begin{pmatrix} \psi_{11}(\tilde{\theta}_T) & \psi_{12}(\tilde{\theta}_T) \\ \psi_{21}(\tilde{\theta}_T) & \psi_{22}(\tilde{\theta}_T) \end{pmatrix} & (\tilde{\theta}_T) &= \begin{pmatrix} \tilde{\theta}_{1T} \\ \tilde{\theta}_{2T} \end{pmatrix}. \end{aligned}$$

The NR updating rule for the sequential estimator θ_T^* is given by

$$\theta_{1T}^* = \tilde{\theta}_{1T} - \psi_{11}^{-1}(\tilde{\theta}_T) \left[\frac{\partial Q_{1T}(\tilde{\theta}_{1T})}{\partial \theta_1} + \frac{\partial \nu(\tilde{\theta}_{1T})'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \tilde{\theta}_{2T})}{\partial \nu} \right] \quad (3.16)$$

$$\theta_{2T}^* = \tilde{\theta}_{2T} - \left[\psi_{22}(\tilde{\theta}_T) - \psi_{21}(\tilde{\theta}_T) \psi_{11}^{-1}(\tilde{\theta}_T) \psi_{12}(\tilde{\theta}_T) \right]^{-1} \frac{\partial Q_{2T}(\nu(\tilde{\theta}_{1T}), \tilde{\theta}_{2T})}{\partial \theta_2}. \quad (3.17)$$

3.2.2 Iterative Estimators

SFK provide examples where $Q_{2T}(\nu(\theta_1), \theta_2)$ may be difficult to optimize. As an alternative to optimizing this piece of the objective function SFK propose the maximization by parts(MBP) algorithm. In the case of separable objective functions the algorithms of FPR all collapse down to the algorithm of SFK.

Starting from the initially consistent estimates $\tilde{\theta}_{1T}$, $\tilde{\theta}_{2T}$ the MBP algorithms iteratively solve equations 3.8 and 3.9. The k^{th} step of the algorithm solves:

$$\frac{\partial Q_{1T}(\theta_1)}{\partial \theta_1} + \frac{\partial \nu(\hat{\theta}_1^{(k-1)})'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\hat{\theta}_1^{(k-1)}), \hat{\theta}_2^{(k-1)})}{\partial \nu} = 0 \quad (3.18)$$

$$\frac{\partial Q_{2T}(\nu(\hat{\theta}_1^{(k-1)}), \theta_2)}{\partial \theta_2} = 0. \quad (3.19)$$

SFK showed that, under an information dominance condition, the MBP algorithm will yield the full MLE upon convergence. To detail this condition define

$$\Sigma = \lim_{T \rightarrow \infty} \begin{pmatrix} -\frac{\partial^2 Q_{1T}(\theta_{10})}{\partial \theta_1 \partial \theta_1} & 0 \\ 0 & -\frac{\partial^2 Q_{2T}(\nu(\theta_{10}), \theta_{20})}{\partial \theta_2 \partial \theta_2} \end{pmatrix} \quad P = \lim_{T \rightarrow \infty} \begin{pmatrix} \psi_{11}(\theta_0) - \frac{\partial^2 Q_{1T}(\theta_{10})}{\partial \theta_1 \partial \theta_1}(\theta_0) & \psi_{12}(\theta_0) \\ \psi_{21}(\theta_0) & 0 \end{pmatrix}$$

$$\Gamma = \Sigma^{-1} P$$

SFK show that the information dominance condition is given by $\|\Gamma\| < 1$, which will be satisfied if $\Gamma^k \rightarrow 0$ as $k \rightarrow \infty$

Lia and Quaquish(2005) in their comments on SFK suggested implementing a Newton-Raphson version of equations 3.18 and 3.19. These Newton-Raphson updating steps are given by

$$\hat{\theta}_{1T}^{(k)} = \hat{\theta}_{1T}^{(k-1)} - \left[\frac{\partial^2 Q_{1T}(\hat{\theta}_{1T}^{(k-1)})}{\partial \theta_1 \partial \theta_1'} \right]^{-1} \left[\frac{\partial Q_{1T}(\hat{\theta}_{1T}^{(k-1)})}{\partial \theta_1} + \frac{\partial \nu(\hat{\theta}_{1T}^{(k-1)})'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\hat{\theta}_{1T}^{(k-1)}), \hat{\theta}_{2T}^{(k-1)})}{\partial \nu} \right] \quad (3.20)$$

$$\hat{\theta}_{2T}^{(k)} = \hat{\theta}_{2T}^{(k-1)} - \left[\frac{\partial^2 Q_{2T}(\nu(\hat{\theta}_{1T}^{(k-1)}), \hat{\theta}_{2T}^{(k-1)})}{\partial \theta_2 \partial \theta_2'} \right]^{-1} \frac{\partial Q_{2T}(\nu(\hat{\theta}_{1T}^{(k-1)}), \hat{\theta}_{2T}^{(k-1)})}{\partial \theta_2} \quad (3.21)$$

To allow for a simple comparison with equations 3.16 and 3.17 we use the earlier stated notation (the ψ^s) and re-write the Newton-Raphson updating steps 3.20 and 3.21 as

$$\hat{\theta}_{1T}^{(k)} = \hat{\theta}_{1T}^{(k-1)} - \left[\frac{\partial^2 Q_{1T}(\tilde{\theta}_{1T})}{\partial \theta_1 \partial \theta_1'} \right]^{-1} \left[\frac{\partial Q_{1T}(\hat{\theta}_{1T}^{(k-1)})}{\partial \theta_1} + \frac{\partial \nu(\hat{\theta}_{1T}^{(k-1)})'}{\partial \theta_1} \frac{\partial Q_{2T}(\nu(\hat{\theta}_{1T}^{(k-1)}), \hat{\theta}_{2T}^{(k-1)})}{\partial \nu} \right] \quad (3.22)$$

$$\hat{\theta}_{2T}^{(k)} = \hat{\theta}_{2T}^{(k-1)} - \left[\psi_{22}(\tilde{\theta}_T) \right]^{-1} \frac{\partial Q_{2T}(\nu(\hat{\theta}_{1T}^{(k-1)}), \hat{\theta}_{2T}^{(k-1)})}{\partial \theta_2}. \quad (3.23)$$

Comparing equations 3.16, 3.17 and 3.22, 3.23 it is clear that both estimators will deliver a consistent estimator in the same number of steps. However, comparing the NR updating rules it is simple to see that the sequential estimator is also efficient after the second step. The MBP algorithm on the other hand only gains efficiency as the number of iterations goes to infinity. Furthermore, efficiency is only guaranteed so long as $\|\Gamma\| < 1$.

4 Sequential Estimators and Minimum Chi-squared Estimators

It is well documented that many types of extremum estimators fit within a GMM estimation framework, examples include; maximum likelihood, euclidean empirical likelihood and non-linear least squares¹⁴. This section shows that applying **Theorem 1** to the case of GMM estimators allows us to recast the sequential GMM estimator as a minimum chi-squared estimator. Given that most extremum estimators can be represented as GMM estimators we can conclude that most sequential extremum estimators have a dual representation as minimum chi-squared estimators. To discuss how this dual representation is possible we first discuss the sequential GMM estimator.

Let θ_0 and $\nu(\theta_0)$ satisfy the assumptions given in the general setup. Assume we have observable moment conditions satisfying

$$E(\phi(X_t, \theta, \nu(\theta))) = 0 \iff \theta = \theta_0.$$

The expectation is taken with respect to the true distribution of $(X_t)_{1 \leq t \leq T}$. For a given weighting matrix W_T define the GMM objective function $Q_T[\theta, \nu(\theta)]$ as,

$$Q_T[\theta, \nu(\theta)] = -\bar{\phi}_T(\theta, \nu(\theta))' W_T \bar{\phi}_T(\theta, \nu(\theta))$$

where $\bar{\phi}_T(\theta, \nu(\theta)) = T^{-1} \sum_{i=1}^T \phi(X_i, \theta, \nu(\theta))$. The first order conditions of this estimator are given by

$$\left[\frac{\bar{\phi}_T(\theta, \nu(\theta))}{\partial \theta'} + \frac{\bar{\phi}_T(\theta, \nu(\theta))}{\partial \nu'} \frac{\partial \nu(\theta)}{\partial \theta'} \right]' W_T \bar{\phi}_T(\theta, \nu(\theta)) = 0. \quad (4.1)$$

¹⁴In many ways the same analysis devised here for extremum estimators could have in fact be carried out using minimum distance estimators

The first order conditions for the GMM estimator are complicated by the second term within (4.1). FPR document cases where the inclusion of the second term within equation 4.1 can lead to computational complexities. Given the potential for computational complexities, and possibly poor finite sample properties, we can employ the sequential extremum estimation strategy to derive a consistent and efficient estimator which is not complicated by the second term within equation (4.1).

Assume we possess an initially consistent estimator $\tilde{\theta}_T$. To derive the sequential GMM estimator, define the altered estimating equations:

$$\tilde{\phi}_T(\theta, \nu(\tilde{\theta}_T)) = \frac{1}{T} \sum_{i=1}^T \phi(X_i, \theta, \nu(\tilde{\theta}_T)) - \frac{1}{T} \sum_{i=1}^T \frac{\partial \phi(X_i, \tilde{\theta}_T, \nu(\tilde{\theta}_T))}{\partial \nu} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta} (\tilde{\theta}_T - \theta). \quad (4.2)$$

The objective function for the sequential estimator is then given by

$$\Phi_T[\theta, \nu(\tilde{\theta}_T)] = -\tilde{\phi}_T(\theta, \nu(\tilde{\theta}_T))' W_T \tilde{\phi}_T(\theta, \nu(\tilde{\theta}_T)). \quad (4.3)$$

If W_T is a consistent estimator for the inverse of the asymptotic variance covariance matrix of $\sqrt{T}\tilde{\phi}_T(\theta_0, \nu(\theta_0))$, denoted by W^{-1} , the estimator defined as

$$\theta_T^* = \arg \max_{\theta} \Phi_T[\theta, \nu(\tilde{\theta}_T)] \quad (4.4)$$

is the GMM version of the efficient sequential estimator. Given the specific structure of the GMM objective function we can recast the sequential GMM estimator as a minimum chi-squared estimator.

4.1 Classical Minimum Distance(CDM) and Minimum Chi-squared Estimators.

In CMD estimation we assume that there exists a vector of reduced form parameters γ , structural parameters θ , and a known continuously differentiable function $h(\cdot)$, that uniquely satisfies:

$$h(\theta_0) = \gamma_0. \quad (4.5)$$

The function $h(\cdot)$ is said to map the structural parameters θ into the reduced form parameters γ . CMD entails first estimating γ_0 by $\hat{\gamma}$ and then using these parameters to solve the following optimization problem in θ :

$$\min_{\theta} (\hat{\gamma} - h(\theta))' M (\hat{\gamma} - h(\theta)), \quad (4.6)$$

for some positive definite weighting matrix M . If the choice of M leads to efficient estimators, this procedure is known as minimum chi-squared estimation.

Given the quadratic nature of the sequential GMM optimization problem we can build a parallel between the GMM estimator defined in equation 4.4 and the CMD estimator defined by equation 4.6. To do so, consider the estimating equations given in 4.2. Expanding equation 4.2 we have

$$\frac{1}{T} \sum_{i=1}^T \phi(X_i, \theta, \nu(\tilde{\theta}_T)) + \frac{1}{T} \sum_{i=1}^T \frac{\partial \phi(X_i, \tilde{\theta}_T, \nu(\tilde{\theta}_T))}{\partial \nu} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'}(\theta) - \frac{1}{T} \sum_{i=1}^T \frac{\partial \phi(X_i, \tilde{\theta}_T, \nu(\tilde{\theta}_T))}{\partial \nu} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'}(\tilde{\theta}_T).$$

If we define

$$\hat{\gamma} = -\frac{1}{T} \sum_{i=1}^T \frac{\partial \phi(X_i, \tilde{\theta}_T, \nu(\tilde{\theta}_T))}{\partial \nu} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'}(\tilde{\theta}_T),$$

$$h(\theta) = -\frac{1}{T} \sum_{i=1}^T \phi(X_i, \theta, \nu(\tilde{\theta}_T)) + \frac{1}{T} \sum_{i=1}^T \frac{\partial \phi(X_i, \tilde{\theta}_T, \nu(\tilde{\theta}_T))}{\partial \nu} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'} \theta,$$

we may restate equation 4.2 as

$$\tilde{\phi}(\theta, \nu(\tilde{\theta}_T)) = \hat{\gamma} - h(\theta).$$

Under the assumption

$$E(\phi(X_t, \theta_0, \nu(\theta_0))) = 0$$

we have $\gamma_0 - h(\theta_0) = 0$, with probability approaching one¹⁵

Given these definitions we see that the sequential GMM estimator given in equation 4.3 can be restated as:

$$\theta_T^* = \arg \min_{\theta} (\hat{\gamma} - h(\theta))' W_T (\hat{\gamma} - h(\theta)). \quad (4.7)$$

Many extremum estimators can be represented as GMM estimators. Given that the expression in equation 4.2 is always valid¹⁶, many of the sequential estimators defined by **Theorem 1** can therefore be recast as minimum chi-squared estimators.

5 Applications of Sequential Estimators

This section details two applications outlining the usefulness of the sequential methodology. The first application considers maximum likelihood estimation of affine term structure models. The second application compares the finite sample properties of the sequential GMM estimator and two-step GMM estimator in the stochastic volatility model of Taylor(1994)[26]¹⁷.

5.1 Affine Term Structure Models.

The econometric analysis of continuous time finance models has received a great deal of attention over the last three decades, Phillips and Yu(2009)[23]. If the model is correctly specified maximum likelihood estimation(MLE) is the preferred choice for parameter estimation.

¹⁵This result holds so long as we can apply a weak law of large numbers. With stricter assumptions on the random variables and functions this result would hold almost surely.

¹⁶Give the standard set of regularity conditions for extremum estimators.

¹⁷Multiple authors have commented on the poor finite sample performance of the GMM estimator in this setting, see references with.

However, due to the discrete sampling nature of the data, we are often unable to explicitly compute likelihood functions associated with financial models specified by continuous-time diffusion equations, Ait-Sahalia(1999)[3].

For many univariate and multivariate affine term structure models the likelihood functions associated with these models are not known in closed form. Fortunately, as shown in Ait-Sahalia(2002)[4], Ait-Sahalia(2008)[1], and Ait-Sahalia and Kimmel(2010)(ASK hereafter)[5] we can employ closed-form discrete approximations to the true likelihood even when the true likelihood is not known in closed form. These approximations allow us to gain consistent and efficient estimators without resorting to semi-parametric or simulation based estimators. Unfortunately, these likelihood approximations suffer from a potential drawback. In certain models, such as the example within this section, the likelihood approximations can become divergent, PPR.

Fortunately, the sequential estimation methodology can be employed to alleviate these divergence issues. We illustrate this concept in the context of closed-form likelihood approximations to the canonical affine term structure models of Dai and Singleton(2000)[7]. The next two sections develop the general canonical affine term structure model and define the general likelihood function. The last three sections derive the sequential estimator and carry out a Monte Carlo experiment to determine the finite sample properties of this method.

5.1.1 Setup and Canonical Representation

In multivariate term structure models the risk-less interest rate r_t is a linear deterministic function of a vector of N state variables X_t . This relationship is stated as:

$$r_t = \delta_0 + \delta_1' X_t, \quad (5.1)$$

δ_0 is an unknown scalar and δ_1 is a $N \times 1$ vector of unknown constants. To avoid arbitrage opportunities the price at time t of a zero coupon bond which matures at time T is given by

$$P(t, \tau) = E^Q \left[\exp \left(- \int_t^\tau r_s ds \right) \middle| X_t = x \right], \quad (5.2)$$

where τ is the time to maturity and Q is an equivalent martingale measure. Under the measure Q the state vector X_t depends on parameters θ and is assumed to follow an “affine diffusion,”

$$dX_t = \mu(X_t; \theta) dt + \sigma(X_t; \theta) \sqrt{X_t} dW_t^Q. \quad (5.3)$$

X_t, μ are $N \times 1$ vectors, σ is a $N \times N$ matrix and W_t^Q is a standard Brownian motion under the martingale measure Q . We assume the market price of risk Λ_t , is of the Dai and Singleton(2000)[7] form and is given by $\Lambda_t = \sqrt{S_t} \lambda$. S_t is a diagonal matrix with i_{th} diagonal element

$$[S_t]_{ii} = \alpha_i + \beta_i' X_t.$$

λ is a $N \times 1$ vector of unknown constants.

Under these assumptions the dynamics of the process X_t under the actual probability measure \mathcal{P} are

$$dX_t = (\tilde{A} + \tilde{B}X_t)dt + \Sigma\sqrt{S_T}dW_t^{\mathcal{P}}. \quad (5.4)$$

The matrix \tilde{A} is a $N \times 1$ vector, \tilde{B} , Σ are $N \times N$ matrices and $W_t^{\mathcal{P}}$ is a standard Brownian motion under \mathcal{P} . This is a slight departure from the canonical representation of Dai and Singleton(2000). This representation is used to make the comparison between our method and those of Ait-Sahalia and Kimmel(2010) easier.

Letting the matrix \mathcal{B} denote the $N \times N$ matrix whose i_{th} column is the vector β_i , the model parameters are given by $(\tilde{A}, \tilde{B}, \Sigma, \alpha, \mathcal{B})$. As stated in Dai and Singleton(2000), certain admissibility restrictions must be placed on these parameters to ensure the existence of the process X_t . Following Dai and Singleton(2000), let $m = rank(\mathcal{B})$ index the degree of dependence of the conditional variances on the number of state variables. With this index, almost all term structure models can be classified into one of $N + 1$ subfamilies¹⁸. The models which satisfy the admissibility conditions found in Dai and Singleton(2000) or ASK are denoted by $A_m(N)$.

In affine term structure models Duffie and Kahn(1996)[8] have shown that bond prices, given by equation (5.2), have an exponential form

$$P(t, \tau) = exp(A(\tau) - B(\tau)'X_t). \quad (5.5)$$

Equation (5.5) ensures that bond yields, $Z_t(\theta, \tau) = -ln(P(t, \tau))$, are affine functions of the state vector X_t and matrices $A(\tau), B(\tau)$:

$$Z_t(\theta, \tau) = -A(\tau) + B(\tau)'X_t.$$

5.1.2 Estimation: Exact Maximum Likelihood

Estimation of the model parameters via exact maximum likelihood requires two sets of bond yields. The first set consists of N zero coupon bond yields observed without error and maturing at dates $\tau_N = (\tau_1, \dots, \tau_N)$. The second set consists of H zero coupon bond yields measured with observation error and maturing at date $\tau_{N+H} = (\tau_{N+1}, \dots, \tau_{N+H})$. Together these equations imply,

$$\begin{aligned} Z_t(\theta, \tau_N) &= -A(\tau_N) + B(\tau_N)'X_t \\ V_t(\theta, \tau_{N+H}) &= -A(\tau_{N+H}) + B(\tau_{N+H})'X_t + \epsilon_t. \end{aligned} \quad (5.6)$$

We assume the $H \times 1$ dimensional observation error ϵ_t is *i.i.d* $\mathcal{N}(0, \Omega)$.

The procedure for evaluating the full likelihood of yields consists of four steps.

(1) Yields: In the first step we extract the value of the state variable by inverting the yield equation to receive the state vector X_T :

$$X_t = B(\tau_N)^{-1} (Z_t(\theta, \tau_N) + A(\tau_N)).$$

¹⁸These sub-families are based on the corresponding value of m

(2) Latent Likelihood: Evaluate the joint likelihood of the series of state vectors using the values derived in the first step. Letting x be the forward value of the state vector and x_0 the backward value. The likelihood of the state vector is then given by

$$\mathcal{L}_x(x, \Delta|x_0; \theta) = \mathcal{L}_x [B(\tau_N)^{-1} (z(\theta, \tau_N) + A(\tau_N)), \Delta | B(\tau_N)^{-1} (z_0(\theta, \tau_N) + A(\tau_N)); \theta].$$

Δ is the discrete sampling interval on which the stochastic process is observed.

(3) Change of Variable: Since the likelihood is now a function of yields, z , we must multiply the joint likelihood by a jacobian determinant.

(4) Total Likelihood: For the yields estimated with error we calculate the likelihood $\mathcal{L}_e(x|x_0; \theta, \eta)$. η are the parameters contained within the variance co-variance matrix Ω . Multiplying the two likelihoods and taking logs yields our objective function

$$Q_n(\theta, \eta) = \ell_x(x, \Delta|x_0; \theta) + \ell_e(x|x_0; \theta, \eta).$$

$\ell_x(\cdot)$ depends on the model but the portion ℓ_e is the same regardless of the model. ℓ_e is given by

$$\begin{aligned} \ell_e(x|x_0; \theta, \eta) = & -\frac{n(H)}{2} \log(2\pi) - \frac{n}{2} \log(\det(\Omega)) \\ & - \frac{1}{2} \sum_{t=1}^n (V_t + A(\tau_{N+H}) - B(\tau_{N+H})' X_t)' \Omega^{-1} (V_t + A(\tau_{N+H}) - B(\tau_{N+H})' X_t), \end{aligned} \quad (5.7)$$

where X_t is given by the values calculated in the first stage.

5.1.3 Sequential Likelihood Approximations

For many of the canonical $A_m(N)$ models the log-likelihood function of the state vector, $\ell_x(\cdot)$, is not known in closed form. To circumvent this issue we employ closed-form likelihood approximations to the log-likelihood of the state vector, Ait-Sahalia(2002), Ait-Sahalia(2008) and ASK. If during the numerical procedure either the forward or backward values of the state vector are zero or close to zero, the likelihood will diverge to plus or minus infinity¹⁹. To some extent Ait-Sahalia and Kimmel(2010) acknowledged this fact. The authors show in the confines of the $A_1(1)$ model, that certain parameter combinations can cause identification failure or near identification failure.

To alleviate the divergence issues associated with approximated likelihood functions we can replace the portions of the likelihood which cause divergence with consistent estimators. We may then maximize the augmented log-likelihood function or use the NR updating rules to determine the efficient parameter estimates

Approximations.

¹⁹This is true for the $A_1(1)$ model.

In this context the closed-form likelihood approximation, assuming all the necessary conditions of ASK are met, is given by

$$\begin{aligned} \ell_x^{(k)}(x, \Delta | x_0; \theta, \nu(\theta)) &= -\frac{m}{2} \ln(2\pi\Delta) - \frac{1}{2} \ln(\det(D_v(x; \theta))) + C_x^{(-1)}(x, \Delta | x_0; \theta, \nu(\theta)) \\ &\quad + \sum_{j=0}^k C_x^j(x, \Delta | x_0; \theta, \nu(\theta)) \frac{\Delta^k}{k!} \end{aligned} \quad (5.8)$$

where $D_v(x; \theta) = \sigma(x; \theta)\sigma(x; \theta)'$ and k refers to the number of terms used in the approximation. The terms $C_x^{(j)}$ represent the coefficients from the likelihood expansions derived in ASK. For the specific form of these coefficients the reader is referred to ASK.

The portions of the coefficients that can cause singularities²⁰ are denoted by $\nu(\theta)$.

If we possess a consistent first stage estimator, $\tilde{\theta}_n$, we can derive the sequential likelihood approximation, $\tilde{\ell}_x^{(k)}(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n))$. This approximation is of the same form as (5.8) with a few alterations,

$$\begin{aligned} \tilde{\ell}_x^{(k)}(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n)) &= -\frac{m}{2} \ln(2\pi\Delta) - \frac{1}{2} \ln(\det(D_v(x; \theta))) + \tilde{C}_x^{(-1)}(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n)) \\ &\quad + \sum_{j=0}^k \tilde{C}_x^j(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n)) \frac{\Delta^k}{k!} - \frac{1}{2} (\theta - \tilde{\theta}_n)' B(\tilde{\theta}_n) (\theta - \tilde{\theta}_n). \end{aligned} \quad (5.9)$$

The sequential approximation coefficients are given by²¹

$$\tilde{C}_x^{(j)}(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n)) = C_x^{(j)}(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n)) + \frac{\partial C_x^{(j)}(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n))}{\partial \nu} \frac{\partial \nu(\tilde{\theta}_n)}{\partial \theta'} (\theta - \tilde{\theta}_n).$$

The matrix $B(\tilde{\theta}_n)$ within equation (5.9) is given by

$$\begin{aligned} B(\tilde{\theta}_n) &= - \sum_{j=0}^k \frac{\partial^2 C_x^{(j)}(x, \Delta | x_0; \tilde{\theta}_n, \nu(\tilde{\theta}_n))}{\partial \nu \partial \nu} \left(\frac{\partial \nu(\tilde{\theta}_n)}{\partial \theta'} \right)^2 \frac{\Delta^j}{j!} - \frac{\partial^2 C_x^{(-1)}(x, \Delta | x_0; \tilde{\theta}_n, \nu(\tilde{\theta}_n))}{\partial \nu \partial \nu} \left(\frac{\partial \nu(\tilde{\theta}_n)}{\partial \theta'} \right)^2 \\ &\quad - \sum_{j=0}^k \frac{\partial^2 C_x^{(j)}(x, \Delta | x_0; \tilde{\theta}_n, \nu(\tilde{\theta}_n))}{\partial \nu} \left(\frac{\partial^2 \nu(\tilde{\theta}_n)}{\partial \theta' \partial \theta} \right) \frac{\Delta^j}{j!} - \frac{\partial^2 C_x^{(-1)}(x, \Delta | x_0; \tilde{\theta}_n, \nu(\tilde{\theta}_n))}{\partial \nu} \left(\frac{\partial^2 \nu(\tilde{\theta}_n)}{\partial \theta' \partial \theta} \right) \end{aligned}$$

The sequential log-likelihood is then the addition of equations (5.9) and (5.7)

$$Q_n(\theta, \nu(\tilde{\theta}_n), \eta) = \tilde{\ell}_x(x, \Delta | x_0; \theta, \nu(\tilde{\theta}_n)) + \ell_e(x, \Delta | x_0; \theta, \eta).$$

A Monte Carlo study explores the differences between our sequential estimator and the original estimators of ASK.

²⁰The values of the state vectors in the denominators

²¹In many cases, such as the $A_1(1)$ model, $\nu(\theta)$ can be expressed as a specific function of $x(\theta)$ and $x_0(\theta)$.

5.1.4 Monte Carlo Setup

To determine the applicability of the sequential method we carry out a Monte Carlo study using the $A_1(1)$ canonical model. For this model the dynamics under the measure \mathcal{P} are given by

$$dX_t = (a_1 + b_1 X_t)dt + \sqrt{X_t}dW_t^{\mathcal{P}},$$

where W_t is a standard Brownian motion under the measure \mathcal{P} . The time varying price of risk is $\Lambda_t = \lambda\sqrt{X_t}$. The dynamics under the risk-neutral measure \mathcal{Q} are given by

$$dX_t = (a_1 + (b_1 - \lambda)X_t)dt + \sqrt{X_t}dW_t^{\mathcal{Q}}.$$

The risk free interest rate is deterministic and modeled as

$$r_t = \delta_0 + \delta_1 X_t.$$

We estimate this model using the original likelihood approximation of ASK and the sequential likelihood approximation derived in the previous sub-section²².

We simulate 1000 data series of 501, 1001, 2001, 5001 and 10001 weekly observations ($\Delta = 1/52$) from the above model. This yields $n = 500, 1000, 2000, 5000$ and 10000 pairs of discrete transitions for the process X_t . Each path is simulated by an Euler discretization using 30 intervals per week, 29 of the thirty observations are discarded. We generate $N + H = 3$ yields with maturities $\tau = [.5, 1, 2]$. We assume that the yields on the $H = 2$ longest maturities, (τ_{N+H}) , contain observation errors. The observations errors are given by $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I_2)$.

$\hat{\theta}^{(k)}$ denotes the ASK estimator of order k and $\tilde{\theta}^{(k)}$ denotes the sequential ASK estimator of order k . For this simulation we set the number of approximation terms to 3 ($k = 3$). For both methods we report; Monte Carlo mean squared error(MSE), parameter averages and average standard errors²³. The below tables detail the results.

Results: Averages, MSE.

The results for the Monte Carlo Averages and MSE are given below.

$k=3, n=500$	$(\hat{\theta}^k)$	$(\tilde{\theta}^k)$	$MSE(\hat{\theta}^k)$	$MSE(\tilde{\theta}^k)$	Std.error($\hat{\theta}^k$)	Std.Error($\tilde{\theta}^k$)
$a_1 = 1.00$	1.0659	1.0273	.0157	.0108	.1867	.0983
$b_1 = -.60$	-.7363	-.6113	.0202	.0023	.1772	.0820
$\delta_0 = .0050$.0058	.0062	.0000	.0000	.0399	.0006
$\delta_1 = .0225$.0231	.0222	.0000	.0000	.0558	.0025
$\lambda = -.150$	-.1780	-.1509	.0010	.0002	.0324	.0188

²²The specific coefficients for this model can be found in the appendix(not yet typed up).

²³The variance of the estimators is calculated using the robust variance covariance definition for quasi-maximum likelihood estimators.

$k=3, n=1000$	$(\hat{\theta}^k)$	$(\tilde{\theta}^k)$	$MSE(\hat{\theta}^k)$	$MSE(\tilde{\theta}^k)$	Std.error($\hat{\theta}^k$)	Std.Error($\tilde{\theta}^k$)
$a_1 = 1.00$	1.0342	1.0178	.0109	.0088	.1372	.0277
$b_1 = -.60$	-.7353	-.5894	.0189	.0019	.1380	.0209
$\delta_0 = .0050$.0058	.0060	.0000	.0000	.0011	.0002
$\delta_1 = .0225$.0225	.0217	.0000	.0000	.0040	.0006
$\lambda = -.150$	-.1761	-.1488	.0007	.0001	.0309	.0051
$k=3, n=2000$	$(\hat{\theta}^k)$	$(\tilde{\theta}^k)$	$MSE(\hat{\theta}^k)$	$MSE(\tilde{\theta}^k)$	Std.error($\hat{\theta}^k$)	Std.Error($\tilde{\theta}^k$)
$a_1 = 1.00$	1.0075	.9974	.0073	.0100	.1215	.0169
$b_1 = -.60$	-.7348	-.5904	.0184	.0001	.0856	.0068
$\delta_0 = .0050$.0058	.0059	.0000	.0000	.0007	.0001
$\delta_1 = .0225$.0219	.0221	.0000	.0000	.0024	.0002
$\lambda = -.150$	-.1755	-.1560	.0007	.0001	.0192	.0017
$k=3, n=5000$	$(\hat{\theta}^k)$	$(\tilde{\theta}^k)$	$MSE(\hat{\theta}^k)$	$MSE(\tilde{\theta}^k)$	Std.error($\hat{\theta}^k$)	Std.Error($\tilde{\theta}^k$)
$a_1 = 1.00$.9793	1.0017	.0027	.0049	.0681	.0067
$b_1 = -.60$	-.7356	-.5910	.0185	.0001	.0716	.0023
$\delta_0 = .0050$.0058	.0059	.0000	.0000	.0006	.0000
$\delta_1 = .0225$.0214	.0223	.0000	.0000	.0020	.0001
$\lambda = -.150$	-.1755	-.1579	.0007	.0001	.0161	.0006
$k=3, n=10000$	$(\hat{\theta}^k)$	$(\tilde{\theta}^k)$	$MSE(\hat{\theta}^k)$	$MSE(\tilde{\theta}^k)$	Std.error($\hat{\theta}^k$)	Std.Error($\tilde{\theta}^k$)
$a_1 = 1.00$.9723	1.0015	.0013	.0017	.0450	.0056
$b_1 = -.60$	-.7362	-.5917	.0186	.0001	.0285	.0011
$\delta_0 = .0050$.0058	.0059	.0000	.0000	.0002	.0000
$\delta_1 = .0225$.0211	.0225	.0000	.0000	.0007	.0001
$\lambda = -.150$	-.1752	-.1602	.0006	.0002	.0063	.0002

Analyzing the results we see that both estimators seem to converge towards the true value fairly quickly, with the sequential estimator converging somewhat faster. However, the estimated standard errors for these estimators are markedly different from one another. This difference may reflect the speed at which the estimators achieve their asymptotic representation.

To determine which of the estimators reaches its asymptotic representation faster, the next table gives the Monte Carlo standard errors (*MC S.E.*) over 1000 replications using 10,000 observations. The next two columns give Monte Carlo standard errors as multiples of the average standard errors.

$k=3$	$MC\ S.E.(\hat{\theta}^{(k)})$	$MC\ S.E.(\tilde{\theta}^{(k)})$	$Mult.(\hat{\theta}^{(k)})$	$Mult.(\tilde{\theta}^{(k)})$
a_1	.0007	.0013	64.16	3.01
b_1	.0002	$6.34e - 005$	128.72	16.56
δ_0	$5.53e - 005$	$3.66e - 006$	39.77	.1700
δ_1	$1.99e - 005$	$2.44e - 005$	36.95	1.10
λ	$7.85e - 005$.0002	80.85	1.12

The above table shows that the standard errors for the sequential estimators are much closer to the Monte Carlo standard errors than the estimator of ASK. This difference would suggest that the sequential method does a better job of approximating the empirical distribution of the estimators than its full information counterpart. This result is not altogether unexpected as a similar result was given by Hoffman(1991)[13] for sequential estimation of rational expectation models²⁴.

5.2 Monte Carlo: GMM Estimation of a Stochastic Volatility Model

To illustrate the merits of the two-stag estimation method within GMM we estimate the log-normal stochastic volatility(SV) model of Taylor(1994)[26] using sequential GMM. The specific SV model used in our example is the log-normal SV model of .

5.2.1 Log-Normal Stochastic Volatility

The log-normal SV model we study is structured as

$$y_t = \sigma_t Z_t \tag{5.10}$$

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \sigma_u u_t \tag{5.11}$$

where (Z_t, u_t) is i.i.d. $\mathcal{N}(0, I_2)$ and the parameter vector is $\theta = (\omega, \beta, \sigma_u)$. (5.10) describes the process mean and (5.11) describes the volatility dynamics. As demonstrated by Taylor(1994) and Melino and Turnbull(1990)[16], log normality allows for closed form solutions of the moment conditions as well as “fat tailed” data, Anderson and Sorenson(1996)(AS

²⁴The specific examples considered within Hoffman(1991) allow for adaptive parameter estimation whereas adaptive estimators are not possible in this setting.

hereafter)[2]. The moments which are the focus of estimation are given by

$$\begin{aligned} E|y_t| - (2/\pi)^{1/2} \exp\left(\frac{\omega}{2(1-\beta)} + \frac{\sigma_u^2}{8(1-\beta^2)}\right) &= 0 \\ E(y_t^2) - \exp\left(\frac{\omega}{(1-\beta)} + \frac{\sigma_u^2}{2(1-\beta^2)}\right) &= 0 \\ E|y_t^r y_{t-j}^k| - (2/\pi) E(\sigma_t^r \sigma_{t-j}^k) &= 0 \\ E(y_t^2 y_{t-j}^2) - E(\sigma_t^2 \sigma_{t-j}^2) &= 0 \end{aligned}$$

where $j = 1, \dots, 5$ and for $r, k = 1, 2$ we have

$$E(\sigma_t^r \sigma_{t-j}^k) = E(\sigma_t^r) E(\sigma_{t-j}^k) \exp\left(rk\beta^j \frac{\sigma_u^2}{8(1-\beta^2)}\right)$$

and

$$E(\sigma_t^r) = \exp\left(r \frac{\omega}{2(1-\beta)} + r^2 \frac{\sigma_u^2}{8(1-\beta^2)}\right).$$

5.2.2 Estimation and Monte Carlo Setup

Estimation of this model by GMM is generally numerically cumbersome and yields poor results, AS. If the estimate of β is sufficiently close to 1, any quadratic objective function using these moment conditions will become unstable. AS have this to say on the subject: “For lower sample sizes our estimation algorithm was frequently unable to locate a minimum for the criterion function within the parameter space,...., as $\hat{\beta}$ became approximately 1, the iterations would crash as the weighting matrix became singular or the criterion function diverged to infinity.” For certain Monte Carlo specifications the authors find that over a third of the replications did not converge.

Estimation of the log-normal SV model by GMM with relatively few observations or with few moments often leads to poor parameter estimates, Jacquier, Polson and Rossi(1994)[14], (JPR hereafter). We show through Monte Carlo simulations that the sequential methods outperform efficient GMM when the number of observations and moment conditions are relatively small.

5.2.3 Sequential Estimation: Setup

Assume we wish to estimate the parameter θ from the sample moment conditions

$$\bar{\phi}_T(\theta) = \frac{1}{T} \sum_{i=1}^T \phi(\theta, y_i),$$

where $\phi(\cdot, \cdot)$ is a vector of functions. For complicated functions $\phi(\cdot, \cdot)$ efficient GMM estimators for θ can be computational burdensome and yield poor results. Through implementing a

sequential method we can decrease the computational burden associated with efficient GMM estimation of θ .

For the log-normal SV model, the explosive behavior of the objective function stems from the estimation of the parameter β . The sequential estimator alleviates the explosive behavior by using consistent estimators for the more ill behaved occurrences of β . The sequential GMM estimator is based on the altered moment conditions,

$$\tilde{\phi}_T(\theta, \tilde{\theta}_T) = \bar{\phi}_T(\omega, \tilde{\beta}_T, \sigma_u) - C_T(\tilde{\beta}_T - \beta). \quad (5.12)$$

The matrix C_T satisfies

$$C_T \rightarrow_p E \left(\frac{\partial \phi(\theta_0, y_t)}{\partial \beta} \right),$$

and $\tilde{\theta}_T$ is a consistent estimator for the parameter θ_0 . In this case the first stage estimates needed to construct C_T can be derived from an initial GMM or method of moments estimator.

To obtain efficient sequential estimators we can minimize either of the below objective functions,

$$J_1(\theta) = \tilde{\phi}_T(\theta, \tilde{\theta}_T)' V_T(\theta)^{-1} \tilde{\phi}_T(\theta, \tilde{\theta}_T)$$

$$J_2(\theta) = \tilde{\phi}_T(\theta, \tilde{\theta}_T)' W_T(\tilde{\theta}_T)^{-1} \tilde{\phi}_T(\theta, \tilde{\theta}_T).$$

The matrix $W_T(\tilde{\theta}_T)$ converges to $E[\phi(\theta_0)\phi(\theta_0)']$ and the matrix $V_T(\hat{\theta})$ converges to $E[\phi(\theta_0)\phi(\theta_0)']$ for $\hat{\theta} \rightarrow_p \theta_0$.

Imprecise estimation of the weighting matrix W_T can lead to numerical instability, AS. To deal with this instability we use the sequential continuously updating GMM estimator (2S-CUGMM). The internally computed weighting matrix $V_T(\theta)$, is estimated from the HAC class of estimators given by Newey and West(1987)[20]. Namely,

$$V_T(\theta) = \sum_{j=-(T-1)}^{T-1} k(j) \hat{M}_T(\theta, j).$$

$k(j)$ is a kernel function dependent on bandwidth h_T and $\hat{M}_T(\theta, j)$ is a covariance estimator at lag j defined by

$$\hat{M}_T(\theta, j) = \frac{1}{T} \sum_{t=j+1}^T (\phi(\theta, y_t) - \bar{\phi}_T(\theta)) \phi(\theta, y_{t-j})'.$$

5.2.4 Monte Carlo Setup

We employ the same setup Monte Carlo setup as in JPR, AS, Takada(2009) and Laurini and Hotla(2010). However, unlike the above authors, our goal is to show that the sequential GMM methods work well with relatively few observations and few moment conditions.

We consider parameter values, justified in JPR, given by $(\omega, \beta, \sigma_u) = (-.736, .9, .363)$. Estimation is carried out across samples of size $T = 500, 1000$. Model parameters are estimated by the sequential CUGMM estimator (2S-CUGMM) using collections of 3, 4 and 5 moments. We conduct 5,000 replications for each sample size and moment combination.

We use the same moment configurations as AS for 3 and 5 moments. For the case of 4 moments we use the original 3 moment configuration of AS as well as the additional moment $E(y_t^2 y_{t-1}^2)$. We then have the following moment collections

$$\begin{aligned} m_1 &= (E|y_t|, E(y_t^2), E(|y_t y_{t-1}|)), \\ m_2 &= (E|y_t|, E|y_t y_{t-1}|, E(y_t^2), E(y_t^2 y_{t-1}^2)), \\ m_3 &= (E|y_t|, E|y_t y_{t-2}|, E(y_t^2), E(y_t^4), E(y_t^2 y_{t-1}^2)). \end{aligned}$$

To facilitate direct comparison with the baseline case of AS we employ a Bartlett kernel estimator for the weighting matrix $V_T(\theta)$, with bandwidth $h_T = 10$.

Tables 1 and 2 detail the simulation results across the different moment and sample size combinations. It is immediately apparent that we no longer have the non-convergence problem that plagued JPR and AS²⁵. Comparing the resulting sequential estimators with those of AS for 3 and 5 moments we see that the sequential methods are superior in terms of RMSE. The RMSEs for the sequential method with a sample size of 1000 and 5 moments are comparable to those of AS with sample sizes of 10,000 and 14 moments.

T=500	m_1	m_2	m_3	$m_1 - AS$	$m_3 - AS$
Mean ω	-.2596	-.8829	-.8018	-1.951	-1.636
RMSE ω	.5015	.1431	.0622	1.854	1.763
Mean β	.8418	.8667	.9006	.736	.786
RMSE β	.1129	.0360	.0179	.250	.209
Mean σ_u	.2911	.3627	.3604	.503	.413
RMSE σ_u	.1160	.0337	.0572	.237	.197
Un-convergent	0	0	0	519	528

T=1000	m_1	m_2	m_3	$m_1 - AS$	$m_3 - AS$
Mean ω	-.3498	-.8818	-.8006	-1.475	-1.135
RMSE ω	.4347	.1419	.0606	1.259	.913
Mean β	.8221	.8724	.9033	.800	.847
RMSE β	.2117	.0319	.0103	.170	.123
Mean σ_u	.3603	.3581	.3475	.458	.372
RMSE σ_u	.1707	.0214	.0339	.201	.142
Un-convergent	0	0	0	422	298

²⁵A contributing factor for this result might be differences in starting values and maximization routines.

6 Conclusion

This article has shown that in situations where an initially consistent estimator exists, the sequential estimator can be used to decrease computational costs while still receiving consistent and efficient estimators. This article has also shown that we can gain estimators which are as computationally friendly as those of closely related iterative estimators, without the need to impose further assumptions which may or may not be satisfied. Furthermore, as evidenced by the Monte Carlo results, when portions of the objective function are singular we can use the sequential methodology to gain estimators with good statistical properties. Future research will focus on extending the sequential methodology to hypothesis testing within a GMM framework. We will also explore applications of the sequential methodology to variables problem of microeconometrics and finance.

References

- [1] Yacine Ait-Sahalia. Closed-form likelihood expansions for multivariate diffusions. NBER Working Papers 8956, National Bureau of Economic Research, Inc, May 2002.
- [2] Torben G. Andersen and Bent E. Srensen. Gmm estimation of a stochastic volatility model: A monte carlo study. *Journal of Business and Economic Statistics*, 14(3):pp. 328–352, 1996.
- [3] Yacine At-Sahalia. Transition densities for interest rate and other nonlinear diffusions. *The Journal of Finance*, 54(4):1361–1395, 1999.
- [4] Yacine At-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002.
- [5] Yacine At-Sahalia and Robert L. Kimmel. Estimating affine multifactor term structure models using closed-form likelihood expansions. *Journal of Financial Economics*, 98(1):113–144, October 2010.
- [6] Bruno Crepon, Francis Kramarz, and Alain Trognon. Parameters of interest, nuisance parameters and orthogonality conditions an application to autoregressive error component models. *Journal of Econometrics*, 82(1):135 – 156, 1997.
- [7] Qiang Dai and Kenneth J. Singleton. Specification analysis of affine term structure models. *The Journal of Finance*, 55(5):pp. 1943–1978, 2000.
- [8] Darrell Duffie and Rui Kan. A yield-factor model of interest rates. *Mathematical Finance*, 6(4):379–406, 1996.
- [9] Pastorello Sergio Fan, Yanqin and Eric Renault. Maximization by parts in extremum estimation. 2012.

- [10] Christian Gouriroux, Alain Monfort, and Eric Renault. Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. *Journal of Statistical Planning and Inference*, 50(1):37 – 63, 1996. Econometric Methodology, Part III.
- [11] A. C. Harvey. Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, 44(3):pp. 461–465, 1976.
- [12] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):pp. 153–161, 1979.
- [13] Dennis L. Hoffman. Two-step and related estimators in contemporary rational-expectations models: An analysis of small-sample properties. *Journal of Business and Economic Statistics*, 9(1):pp. 51–61, 1991.
- [14] Eric Jacquier, Nicholas G. Polson, and Peter E. Rossi. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, 12(4):pp. 371–389, 1994.
- [15] Lung-Fei Lee, G. S. Maddala, and R. P. Trost. Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity. *Econometrica*, 48(2):pp. 491–503, 1980.
- [16] Angelo Melino and Stuart M. Turnbull. Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45(1-2):239–265, 1990.
- [17] Kevin M. Murphy and Robert H. Topel. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 20(1):88–97, 2002.
- [18] Whitney K. Newey. Two-step series estimation of sample selection models. *Econometrics Journal*, 12(s1):S217–S229, 2009.
- [19] Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier, 1994.
- [20] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):pp. 703–708, 1987.
- [21] Adrian Pagan. Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4):pp. 517–538, 1986.
- [22] Sergio Pastorello, Valentin Patilea, and Eric Renault. Iterative and recursive estimation in structural nonadaptive models. *Journal of Business and Economic Statistics*, 21(4):pp. 449–482, 2003.

- [23] Peter C. B. Phillips and Jun Yu. Maximum likelihood and gaussian estimation of continuous time models in finance. In Thomas Mikosch, Jens-Peter Krei, Richard A. Davis, and Torben Gustav Andersen, editors, *Handbook of Financial Time Series*, pages 497–530. Springer Berlin Heidelberg, 2009.
- [24] Peter X.-K. Song, Yanqin Fan, John D. Kalbfleisch, Jiming Jiang, Thomas A. Louis, J. G. Liao, Bahjat F. Qaqish, and David Ruppert. Maximization by parts in likelihood inference [with comments, rejoinder]. *Journal of the American Statistical Association*, 100(472):pp. 1145–1167, 2005.
- [25] Takeshi and Amemiya. On a two-step estimation of a multivariate logit model. *Journal of Econometrics*, 8(1):13 – 21, 1978.
- [26] Stephen J. Taylor. Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, 4(2):183–204, 1994.
- [27] Alain Trognon and Christina Grourieroux. A note on the efficiency of two-step estimation methods. pages 236–248, 1990.
- [28] Francis Vella and Marno Verbeek. Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics*, 90(2):239 – 263, 1999.

7 Appendix

7.1 Proof: Theorem 1

Proof. The first piece of the proof will revolve around the Taylor series expansion of the first order conditions. These first order conditions are given by,

$$\frac{\partial \Phi_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \theta} = 0.$$

This yields,

$$0 = \frac{\partial Q_T(\theta_T^*, \nu(\tilde{\theta}_T))}{\partial \theta'} + \frac{\partial Q_T(\theta_T^*, \nu(\tilde{\theta}_T))}{\partial \nu'} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'} + \frac{\partial Q_T(\theta_T^*, \nu(\tilde{\theta}_T))}{\partial \theta \partial \nu'} \frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'} (\theta_T^* - \tilde{\theta}_T) - TB_T(\theta_T^*)(\theta_T^* - \tilde{\theta}_T) - \frac{1}{2}(\theta_T^* - \tilde{\theta}_T)T \frac{\partial B_T(\theta_T^*)}{\partial \theta'} (\theta_T^* - \tilde{\theta}_T). \quad (7.1)$$

A first order Taylor series expansion of (7.1) around (θ_0, θ_0) gives

$$\begin{aligned}
0 \approx & \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta'} + \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} (\theta_T^* - \tilde{\theta}_T) - T B_T(\theta_0) (\theta_T^* - \tilde{\theta}_T) \\
& + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta \partial \theta'} (\theta_T^* - \theta_0) + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} (\theta_T^* - \theta_0) + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} (\tilde{\theta}_T - \theta_0) \\
& + \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial^2 \nu(\theta_0)}{\partial \theta \partial \theta'} (\tilde{\theta}_T - \theta_0) + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} (\tilde{\theta}_T - \theta_0).
\end{aligned} \tag{7.2}$$

In this setting the full Hessian matrix is denoted by $D^2 Q_T$ and is given by,

$$\begin{aligned}
D^2 Q_T[\theta_0, \nu(\theta_0)] = & \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} + \\
& \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} + \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial^2 \nu(\theta_0)}{\partial \theta \partial \theta'} + \\
& \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'}
\end{aligned}$$

We now show that the second order terms within the expansion (7.2) are the same as in the above Hessian. The trick to showing this is to recall that

$$p \lim_{T \rightarrow \infty} \left[B_T(\theta_0) + \frac{1}{T} \left(\frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial^2 \nu(\theta_0)}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \right) \right] = 0.$$

We then have that

$$\begin{aligned}
\left[-B_T(\theta_0) (\theta_T^* - \tilde{\theta}_T) + \frac{1}{T} \left(\frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial^2 \nu(\theta_0)}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \right) (\tilde{\theta}_T - \theta_0) \right] = \\
\left[\frac{1}{T} \left(\frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial^2 \nu(\theta_0)}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu \partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \right) (\theta_T^* - \theta_0) + o_p(1) \right]
\end{aligned} \tag{7.3}$$

Plugging this relationship into the first order expansion (7.2) and multiplying by $1/\sqrt{T}$ we may cancel and re-arrange terms to receive

$$0 = \frac{1}{\sqrt{T}} \left(\frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \theta'} + \frac{\partial Q_T[\theta_0, \nu(\theta_0)]}{\partial \nu'} \frac{\partial \nu(\theta_0)}{\partial \theta'} \right) + \frac{1}{T} [D^2 Q_T[\theta_0, \nu(\theta_0)]] \sqrt{T} (\theta_T^* - \theta_0) \tag{7.4}$$

This result allows us to see that the expansion for $(\theta_T^* - \theta_0)$, which is given by (7.4), will be the same expansion as the full information estimator given by (1.1). \square

7.2 More Comments on Likelihood Monte Carlo

- (1) It should be noted that any gradient based maximization algorithms, such as `fminunc` in Matlab, will not be able to maximize the approximated likelihood function. Inevitably employing any type of gradient algorithm will result in infinite values of the objective function forcing the algorithm to terminate.
- (2) The starting values of the two estimators were not the same. The original estimator of ASK were initialized at the true parameter values whereas the sequential estimators were initialized at the same values except for the parameter a_1 , which was initialized lower than the true value.
- (3) The first stage estimators have been specified to be very close to the true values. This was done since we are trying to discern estimation properties in the best case scenario. However, a few numerical experiments have shown that so long as the first stage estimators are fairly close to the true values this does not affect the accuracy of the sequential estimators.

8 Sequential Estimators

This section gives a short synopsis of some of the more general studies dealing and employing with sequential estimation. These are separated into adaptive estimators and nonadaptive estimators.

8.1 Adaptive Estimation

8.1.1 General Sequential Estimators

- (1) Harvey(1976)[11]. Generally accepted as the first application of sequential estimation within econometrics. Two-step estimation methods are derived for models with heteroskedasticity. These estimators are compared with full MLE.
- (2) Amemiya(1978)[25]. Details adaptive sequential estimation of the multinomial logit model.
- (3) Lee and Maddala(1980)[15]. Analyzes the inefficiencies introduced by two-step estimation in the confines of tobit and probit models for simultaneous equation models with selectivity.
- (4) Pagan(1986)[21]. Details conditions for sequential adaptive estimation of parameters within MLE.
- (5) Hoffman(1991)[13]. Applies the estimators derived within Pagan(1986) to rational expectation models. This study is one of the few studies to detail simulation results for two-step methods. These results show that in certain cases two-step/sequential estimators exhibit better performance than full information estimators.

- (6) Newey and McFadden(1994)[19]. Chapter 5 details conditions under which two-step estimators can be adaptively estimated.

8.1.2 Select Notable Applications of Sequential Estimation

- (1) Heckman(1979)[12]. This paper derives the famous sequential “Heck-it” estimators for sample selection methods.
- (2) Vella and Verbeek(1999)[28]. The authors apply the sequential estimation method to panel models with censored variables and selection bias.
- (3) Murphey and Roberts(2002)[17]. This paper deals with imputed regressions generated from preliminary estimators.
- (4) Newey(2009)[18]. The author derives two-step series estimators for sample selection models where the selection equation contains infinite dimensional nuisance parameters.

8.2 Nonadaptive Estimation

- (1) Trognon and Gourieroux(1990)[27]. This paper details a general method for deriving efficient estimators within nonadaptive models estimated using extremum estimators.
- (2) Gourieroux, Monfort and Renault(1996)[10]. The authors detail a sequential estimation strategy where the nuisance parameters is a second occurrence of the parameter of interest which complicates estimation.
- (3) Creopn, et al.(1997)[6]. The authors presume there exists a specific relationship between the nuisance parameters and the parameter of interest. These relationships are used in a general way to obtain sequential estimators for the parameter of interest and then the nuisance parameters.
- (4) Newey and McFadden(1994)[19]. The authors give a general discussion of sequential estimators. Conditions are detailed under which the asymptotic variance of the sequential estimators will be larger or smaller than the variance which does not account for the preliminary estimators.