3
Large-Sample Properties of Extremum Estimators

EXTREMUM ESTIMATORS ARE estimators obtained by either maximizing or minimizing a criterion function over the admissible parameter space. In this chapter we introduce more formally the concept of an extremum estimator and discuss the large-sample properties of these estimators.[1] After briefly setting up notation and describing the probability environment within which we discuss estimation, we describe regularity conditions under which an estimator converges almost surely to its population counterpart.

We then turn to the large-sample distributions of extremum estimators. Throughout this discussion we maintain the assumption that θ_T is a consistent estimator of θ_0 and focus on properties of the distribution of θ_T as T gets large. Whereas discussions of consistency are often criterion-function specific, the large-sample analyses of most of the extremum estimators we will use subsequently can be treated concurrently. We formally define a family of estimators that encompasses the first-order conditions of the ML, standard GMM, and LLS estimators as special cases. Then, after we present a quite general central limit theorem, we establish the asymptotic normality of these estimators. Finally, we examine the relative asymptotic efficiencies of the GMM, LSS, and ML estimators and interpret their asymptotic efficiencies in terms of the restrictions on the joint distribution of the data used in estimation.

3.1. Basic Probability Model

Notationally, we let Ω denote the sample space, $\mathcal{F}$ the set of events about which we want to make probability statements (a "σ-algebra" of events), and

[1] The perspective on the large-sample properties of extremum estimators taken in this chapter has been shaped by my discussions and collaborations with Lars Hansen over the past 25 years. In particular, the approach to establishing consistency and asymptotic normality in Sections 3.2–3.4 follows that of Hansen (1982b, 2005).

Pr the probability measure.[2] Thus, we denote the probability space by $(\Omega, \mathcal{F}, \Pr)$. Similarly, we let $\mathcal{B}^K$ denote the Borel algebra of events in $\mathbb{R}^K$, which is the smallest σ-algebra containing all open and closed rectangles in $\mathbb{R}^K$. A K-dimensional vector random variable X is a function from the sample space Ω to $\mathbb{R}^K$ with the property that for each $B \in \mathcal{B}^K$, $\{\omega : X(\omega) \in B\} \in \mathcal{F}$. Each random variable X induces a probability space $(\mathbb{R}^K, \mathcal{B}^K, \mu_X)$ by the correspondence $\mu_X(B) = \Pr\{\omega : X(\omega) \in B\}$, for all $B \in \mathcal{B}^K$.

Two notions of convergence of sequences of random variables that we use extensively are as follows.

Definition 3.1. *The sequence of random variables $\{X_T\}$ is said to converge almost surely (a.s.) to the random variable X if and only if there exists a null set[3] $\mathcal{N}$ such that*

$$\forall \omega \in \Omega \setminus \mathcal{N} : \quad \lim_{T \to \infty} X_T(\omega) = X(\omega). \tag{3.1}$$

Definition 3.2. *The sequence of random variables $\{X_T\}$ is said to converge in probability to X if and only if, for every $\epsilon > 0$, we have*

$$\lim_{T \to \infty} \Pr\{|X_T - X| > \epsilon\} = 0. \tag{3.2}$$

When the Tth element of the sequence is the estimator θ_T for sample size T and the limit is the population parameter vector of interest θ_0, then we call the estimator θ_T *consistent* for θ_0.

Definition 3.3. *A sequence of estimators $\{\theta_T\}$ is said to be strongly (weakly) consistent for a constant parameter vector θ_0 if and only if θ_T converges almost surely (in probability) to θ_0 as $T \to \infty$.*

There are many different sets of sufficient conditions on the structure of asset pricing models and the probability models generating uncertainty for extremum estimators to be consistent. In this chapter we follow closely the approach in Hansen (1982b), which assumes that the underlying random vector of interest, z_t, is a stationary and ergodic time series. Chapters 9 and 10 discuss how stochastic trends have been accommodated in DAPMs.

We let $\mathbb{R}^\infty$ denote the space consisting of all infinite sequences $x = (x_1, x_2, \ldots)$ of real numbers (lower case x indicates $x \in \mathbb{R}$). A T-dimensional rectangle is of the form $\{x \in \mathbb{R}^\infty : x_1 \in I_1, x_2 \in I_2, \ldots, x_T \in I_T\}$, where $I_1, \ldots, I_T$ are finite or infinite intervals in $\mathbb{R}$. If $\mathcal{B}^\infty$ denotes the smallest

[2] The topics discussed in this section are covered in more depth in most intermediate statistics books. See Chung (1974) and Billingsley (1979).

[3] A null set $\mathcal{N}$ for P is a set with the property that $\Pr\{\mathcal{N}\} = 0$.

σ-algebra of subsets of $\mathbb{R}^\infty$ containing all finite dimensional rectangles, then $X = (X_1, X_2, \ldots)$ is a measurable mapping from Ω to $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ (here the X's are random variables).

Definition 3.4. *A process $\{X_t\}$ is called stationary if, for every k, the process $\{X_t\}_{t=k}^\infty$ has the same distribution as $\{X_t\}_{t=1}^\infty$; that is,*

$$P\{(X_1, X_2, \ldots) \in \mathcal{B}^\infty\} = P\{(X_{k+1}, X_{k+2} \ldots) \in \mathcal{B}^\infty\}. \tag{3.3}$$

In practical terms, a stationary process is one such that the functional forms of the joint distributions of collections $(X_k, X_{k-1}, \ldots, X_{k-\ell})$ do not change over time. An important property of a stationary process is that the process $\{Y_k\}$ defined by $Y_k = f(X_k, X_{k+1}, \ldots,)$ is also stationary for any f that is measurable relative to $\mathcal{B}^\infty$.

The assumption that $\{X_t\}$ is stationary is not sufficient to ensure that sample averages of the process converge to EX_1, a requirement that underlies our large-sample analysis of estimators. (Here we use EX_1, because all X_t have the same mean.) The reason is that the sample we observe is the realization $(X_1(\omega_0), X_2(\omega_0), \ldots)$ associated with a single ω_0 in the sample space Ω. If we are to learn about the distribution of the time series $\{X_t\}$ from this realization, then, as we move along the series $\{X_t(\omega_0)\}$, it must be as if we are observing realizations of $X_t(\omega)$ for fixed t as ω ranges over Ω.

To make this idea more precise,[4] suppose there is an event $A \in \mathcal{F}$ with the property that one can find a $B \in \mathcal{B}^\infty$ such that for every $t > 1$, $A = \{\omega : (X_t(\omega), X_{t+1}(\omega), \ldots) \in B\}$. Such an event A is called *invariant* because, for $\omega_0 \in A$, the information provided by $\{X_t(\omega_0), X_{t+1}(\omega_0), \ldots\}$ as t increases is essentially unchanged with t. On the other hand, if such a B does not exist, then

$$A = \{\omega : (X_1(\omega), X_2(\omega), \ldots) \in B\} \neq \{\omega : (X_t(\omega), X_{t+1}(\omega), \ldots) \in B\}, \tag{3.4}$$

for some $t > 1$, and $\{X_t(\omega), X_{t+1}(\omega), \ldots\}$ conveys information about a different event in $\mathcal{F}$ (different part of Ω).

Definition 3.5. *A stationary process is ergodic if every invariant event has probability zero or one.*

If the process is ergodic, then a single realization conveys sufficient information about Ω for a strong law of large numbers (SLLN) to hold.

[4] For further discussion of stationary and ergodic stochastic processes see, e.g., Breiman (1968).

Theorem 3.1. *If $X_1 X_2, \ldots,$ is a stationary and ergodic process and $E|X_1| < \infty$, then*

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to EX_1 \text{ a.s.} \tag{3.5}$$

One can relax the assumption of stationarity, thereby allowing the marginal distributions of z_t to change over time, and still obtain a SLLN. However, this is typically accomplished by replacing the relatively weak requirements implicit in the assumption of stationarity on the dependence between z_t and z_{t-s}, for $s \neq 0$, with stronger assumptions (see, e.g., Gallant and White, 1988).

Two considerations motivate our focus on the case of stationary and ergodic time series. First, in dynamic asset pricing models, the pricing relations are typically the solutions to a dynamic optimization problem by investors or a replication argument based on no-arbitrage opportunities. As we will see more formally in later chapters, both of these arguments involve optimal forecasts of future variables, and these optimal forecasting problems are typically solved under the assumption of stationary time series.[5] Indeed, these forecasting problems will generally not lend themselves to tractable solutions in the absence of stationarity. Second, the assumption that a time series is stationary does not preclude variation over time in the *conditional* distributions of z_t conditioned on its own history. In particular, the time variation in conditional means and variances that is often the focus of financial econometric modeling is easily accommodated within the framework of stationary and ergodic time series.

Of course, neither of these considerations rules out the possibility that the real world is one in which time series are in fact nonstationary. At a conceptual level, the economic argument for nonstationarity often comes down to the need to include additional conditioning variables. For example, the case of a change in operating procedures by a monetary authority, as we experienced in the United States in the early 1980s, could be handled by conditioning on variables that determine a monetary authority's operating procedures. However, many of the changes in a pricing environment that would lead us to be concerned about stationarity happen infrequently. Therefore, we do not have repeated observations on the changes that concern us the most. The pragmatic solution to this problem has often been to judiciously choose the sample period so that the state vector z_t in an asset pricing model can reasonably be assumed to be stationary. With these considerations in mind, we proceed under the formal assumption of stationary time series.

[5] An important exception is the case of nonstationarity induced by stochastic trends.

3.2. Consistency: General Considerations

Let $Q_T(\vec{z}_T, \theta)$ denote the function to be minimized by choice of the K-vector θ of unknown parameters within an admissible parameter space $\Theta \subset \mathbb{R}^K$, and let $Q_0(\theta)$ be its population counterpart. Throughout this chapter, it will be assumed that $Q_0(\theta)$ is uniquely minimized at θ_0, the model parameters that generate the data.

We begin by presenting a set of quite general sufficient conditions for θ_T to be a consistent estimator of θ_0. The discussion of these conditions is intended to illustrate the essential features of a probability model that lead to *strong* consistency (θ_T converges almost surely to θ_0). Without further assumptions, however, the general conditions proposed are not easily verified in practice. Therefore, we proceed to examine a more primitive set of conditions that imply the conditions of our initial consistency theorem.

One critical assumption underlying consistency is the uniform convergence of sample criterion functions to their population counterparts as T gets large. Following are definitions of two notions of uniform convergence.

Definition 3.6. *Let $g_T(\theta)$ be a nonnegative sequence of random variables depending on the parameter θ. Consider the two modes of uniform convergence of $g_T(\theta)$ to 0:*

$$P\left[\lim_{T\to\infty} \sup_{\theta\in\Theta} g_T(\theta) = 0\right] = 1, \tag{3.6}$$

$$\lim_{T\to\infty} P\left[\sup_{\theta\in\Theta} g_T(\theta) < \epsilon\right] = 1 \quad \textit{for any } \epsilon > 0. \tag{3.7}$$

If (3.6) holds, then $g_T(\theta)$ is said to converge to 0 almost surely uniformly in $\theta \in \Theta$. If (3.7) holds, then $g_T(\theta)$ is said to converge to 0 in probability uniformly in $\theta \in \Theta$.

The following theorem presents a useful set of sufficient conditions for θ_T to converge almost surely to θ_0.

Theorem 3.2. *Suppose*

(i) Θ is compact.

(ii) The nonnegative sample criterion function $Q_T(\vec{z}_T, \theta)$ is continuous in $\theta \in \Theta$ and is a measurable function of $\vec{z}_T$ for all θ.

(iii) $Q_T(\vec{z}_T, \theta)$ converges to a non-stochastic function $Q_0(\theta)$ almost surely uniformly in $\theta \in \Theta$ as $T \to \infty$; and $Q_0(\theta)$ attains a unique minimum at θ_0.

Define θ_T as a value of θ that satisfies

$$Q_T(\vec{z}_T, \theta_T) = \min_{\theta\in\Theta} Q_T(\vec{z}_T, \theta). \tag{3.8}$$

Then θ_T converges almost surely to θ_0.[6]

[6] In situations where θ_T is not unique, if we let Γ_T denote the set of minimizers, we can show that $\delta_T(\omega) = \sup\{|\theta_T - \theta_0| : \theta_T \in \Gamma_T\}$ converges almost surely to 0 as $T \to \infty$.

Proof (Theorem 3.2). *Define the function*

$$\rho(\epsilon) = \inf\{Q_0(\theta) - Q_0(\theta_0), \quad \textit{for } |\theta - \theta_0| \geq \epsilon\}. \tag{3.9}$$

As long as $\epsilon > 0$, *Assumptions (i)–(iii) guarantee that* $\rho(\epsilon) > 0$. *(Continuity of* Q_0 *follows from our assumptions.) Assumption (iii) implies that there exists a set* Λ *with* $P(\Lambda) = 1$ *and a positive, finite function* $T(\omega, \epsilon)$, *such that*

$$\rho_T(\omega) \equiv \sup_{\theta \in \Theta} |Q_T(\omega, \theta) - Q_0(\theta)| < \rho(\epsilon)/2, \tag{3.10}$$

for all $\omega \in \Lambda$, $\epsilon > 0$, *and* $T \geq T(\omega, \epsilon)$. *This inequality guarantees that for all* $\omega \in \Lambda$, $\epsilon > 0$, *and* $T \geq T(\omega, \epsilon)$,

$$\begin{aligned} Q_0(\theta_T) - Q_0(\theta_0) &= Q_0(\theta_T) - Q_T(\omega, \theta_T) + Q_T(\omega, \theta_T) \\ &\quad - Q_T(\omega, \theta_0) + Q_T(\omega, \theta_0) - Q_0(\theta_0) \\ &\leq Q_0(\theta_T) - Q_T(\omega, \theta_T) + Q_T(\omega, \theta_0) - Q_0(\theta_0) \\ &\leq |Q_0(\theta_T) - Q_T(\omega, \theta_T)| + |Q_T(\omega, \theta_0) - Q_0(\theta_0)| \\ &\leq 2\rho_T(\omega) < \rho(\epsilon), \end{aligned} \tag{3.11}$$

which implies that $|\theta_T - \theta_0| < \epsilon$ *for all* $\omega \in \Lambda$, $\epsilon > 0$, *and* $T \geq T(\omega, \epsilon)$.

The assumptions of Theorem 3.2 are quite general. In particular, the z_t's need not be identically distributed or independent. However, this generality is of little practical value unless the assumptions of the theorem can be verified in actual applications. In practice, this amounts to verifying Assumption (iii). The regularity conditions imposed in the econometrics literature to assure that (iii) holds typically depend on the specification of Q_T and Q_0 and, thus, are often criterion function specific. We present a set of sufficient conditions to establish the almost sure uniform convergence of the sample mean

$$G_T(\vec{z}_T, \theta) = \frac{1}{T} \sum_{t=1}^{T} g(z_t, \theta) \tag{3.12}$$

to its population counterpart $G_0(\theta) = E[g(z_t, \theta)]$. This result then is used to establish the uniform convergence of Q_T to Q_0 for the cases of ML and GMM estimators for stationary processes.

To motivate the regularity conditions we impose on the time series $\{z_t\}$ and the function g, it is instructive to examine how far the assumption that

$\{z_t\}$ is stationary and ergodic takes us toward fulfilling the assumptions of Theorem 3.2. Therefore, we begin by assuming:

Assumption 3.1. *$\{z_t : t \geq 1\}$ is a stationary and ergodic stochastic process.*

As discussed in Chapter 2, the sample and population criterion functions for LLP are

$$Q_0(\delta) = E\Big[\big(y_t - x_t'\delta\big)^2\Big], \quad Q_T(\delta) = \frac{1}{T}\sum_{t=1}^{T}\big(y_t - x_t'\delta\big)^2, \quad \delta \in \mathbb{R}^K. \tag{3.13}$$

For the LLP problem, $Q_0(\delta)$ is assured of having a unique minimizer δ_0 if the second-moment matrix $E[x_t x_t']$ has full rank. Thus, with this additional assumption, the second part of Condition (iii) of Theorem 3.2 is satisfied. Furthermore, under the assumption of ergodicity,

$$\frac{1}{T}\sum_{t=1}^{T} x_t x_t' \to E\big[x_t x_t'\big] \quad \text{and} \quad \frac{1}{T}\sum_{t=1}^{T} x_t y_t \to E[x_t y_t] \quad \text{a.s.} \tag{3.14}$$

It follows immediately that $\delta_T \to \delta_0$ a.s.

Though unnecessary in this case, we can also establish the strong consistency of δ_T for δ_0 from the observation that $Q_T(\delta) \to Q_0(\delta)$ a.s., for all $\delta \in \mathbb{R}^K$. From Figure 3.1 it is seen that the criterion functions are quadratic and eventually overlap (for large T), so the minimizers of $Q_T(\delta)$ and $Q_0(\delta)$ must eventually coincide. We conclude that the strong consistency of estimators in LLP problems is essentially implied by the assumption that $\{z_t\}$ is stationary and ergodic (and the rank condition on $E[x_t x_t']$).

More generally, the assumptions of ergodicity of $\{z_t\}$ and the continuity of $Q_T(\vec{z}_T, \theta)$ in its second argument do not imply the strong consistency of the minimizer θ_T of the criterion function $Q_T(\theta)$. The reason is that ergodicity guarantees only pointwise convergence, and the behavior in the "tails" of some nonlinear criterion functions may be problematic. To illustrate this

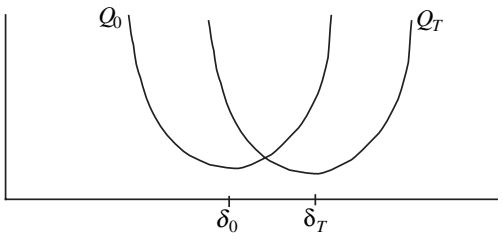

Figure 3.1. *Sample and population criterion functions for a least-squares projection.*

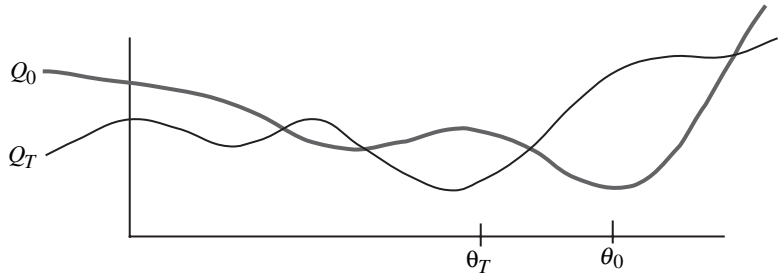

Figure 3.2. *Well-behaved Q_0, Q_T.*

point, Figure 3.2 depicts a relatively well-behaved function Q_T that implies the convergence of θ_T to θ_0. In contrast, although the function $Q_T(\theta)$ in Figure 3.3 can be constructed to converge pointwise to $Q_0(\theta)$, θ_0 and θ_T may grow increasingly far apart as T increases if the dip moves further out to the right as T grows. This potential problem is ruled out by the assumptions that $\{Q_T : T \geq 1\}$ converges almost surely uniformly in θ to a function Q_0 and that θ_0 is the unique minimizer of Q_0.

Even uniform convergence of Q_T to Q_0 combined with stationarity and ergodicity are not sufficient to ensure that θ_T converges to θ_0, however. To see why, consider the situation in Figure 3.4. If $Q_0(\theta)$ asymptotes to the minimum of $Q_0(\theta)$ over $\mathbb{R}$ (but does not achieve this minimum) in the left tail, then $Q_T(\theta_T)$ can get arbitrarily close to $Q_0(\theta_0)$, even though θ_T and θ_0 are growing infinitely far apart. To rule this case out, we need to impose a restriction on the behavior of Q_0 in the "tails." This can be accomplished either by imposing restrictions on the admissible parameter space Θ or by restricting Q_0 directly. For example, if it is required that

$$\inf\{Q_0(\theta) - Q_0(\theta_0) : \theta \in \Theta,\ |\theta - \theta_0| > \rho\} > 0, \tag{3.15}$$

then $Q_0(\theta)$ cannot asymptote to $Q_0(\theta_0)$, for θ far away from θ_0, and convergence of θ_T to θ_0 is ensured. This condition is satisfied by the least-squares

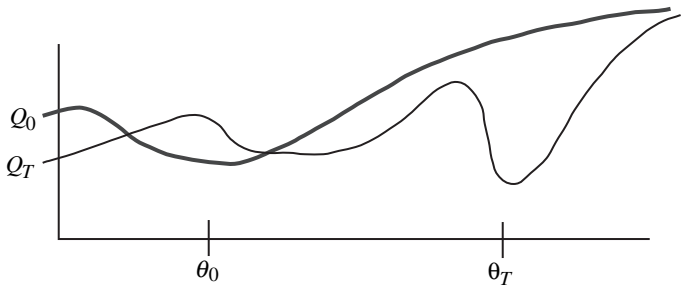

Figure 3.3. *Poorly behaved Q_T.*

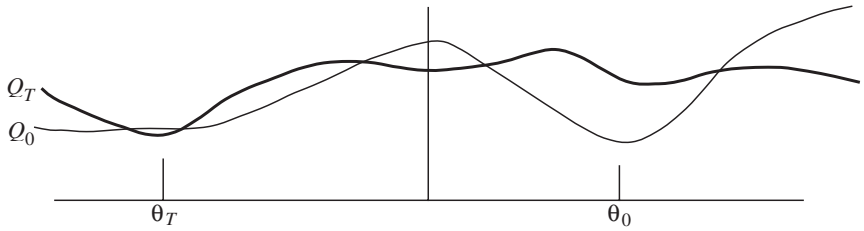

Figure 3.4. *Q_T converging to asymptoting Q_0.*

criterion function for linear models. For nonlinear models, potentially undesirable behavior in the tails is typically ruled out by assuming that Θ is compact (the tails are "chopped off").

With these observations as background, we next provide a primitive set of assumptions that assure the strong consistency of θ_T for θ_0. As noted in Chapter 1, most of the criterion functions we will examine can be expressed as sample means of functions $g(z_t, \theta)$, or are simple functions of such sample means (e.g., a quadratic form). Accordingly, we first present sufficient conditions (beyond Assumption 3.1) for the convergence of

$$G_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(z_t, \theta) \tag{3.16}$$

to $E[g(z_t, \theta)]$ almost surely, uniformly in $\theta \in \Theta$. Our first assumption rules out bad behavior in the tails and the second states that the function $g(z_t, \theta)$ has a finite mean for all θ:

Assumption 3.2. *Θ is a compact metric space.*

Assumption 3.3. *The function $g(\cdot, \theta)$ is Borel measurable for each θ in Θ; $Eg(z_t, \theta)$ exists and is finite for all θ in Θ.*

We will also need a stronger notion of continuity of $g(z_t, \theta)$. Let[7]

$$\epsilon_t(\theta, \delta) = \sup\{|g(z_t, \theta) - g(z_t, \alpha)| \text{ for all } \alpha \text{ in } \Theta \text{ with } |\alpha - \theta| < \delta\}. \tag{3.17}$$

Definition 3.7. *The random function $g(z_t, \theta)$ is first-moment continuous at θ if* $\lim_{\delta \downarrow 0} E[\epsilon_t(\theta, \delta)] = 0$.

[7] Assumption 3.2 guarantees that Θ has a countable dense subset. Hence, under Assumptions 3.2 and 3.3, the function $\epsilon_t(\theta, \delta)$ is Borel measurable (it can be represented as the almost sure supremum of a countable collection of Borel measurable functions).

First-moment continuity of $g(z_t, \theta)$ is a joint property of the function g and the random vector z_t. Under Assumptions 3.1–3.3, if $g(z_t, \theta)$ is first-moment continuous at θ, then $g(z_t, \theta)$ is first-moment continuous for every $t \geq 1$.

Assumption 3.4. *The random function $g(z_t, \theta)$ is first-moment continuous at all $\theta \in \Theta$.*

The measure of distance between G_T and $E[g(z_t, \cdot)]$ we are concerned with is

$$\rho_T = \sup_{\theta \in \Theta} \left| G_T(\theta) - Eg(z_t, \theta) \right|. \tag{3.18}$$

Using the compactness of Θ and the continuity of $g_t(\cdot)$, it can be shown that $\{\rho_T : T \geq 1\}$ converges almost surely to zero. The proof proceeds as follows: Let $\{\theta_i : i \geq 1\}$ be a countable dense subset of Θ. The distance between $G_T(\theta)$ and $Eg(z_t, \theta)$ satisfies the following inequality:

$$\begin{aligned} \left| G_T(\theta) - Eg(z_t, \theta) \right| \leq{}& \left| G_T(\theta) - G_T(\theta_i) \right| \\ &+ \left| G_T(\theta_i) - Eg(z_t, \theta_i) \right| + \left| Eg(z_t, \theta_i) - Eg(z_t, \theta) \right|. \end{aligned} \tag{3.19}$$

For all $\theta \in \Theta$, the first term on the right-hand side of (3.19) can be made arbitrarily small by choosing θ_i such that $|\theta_i - \theta|$ is small (because the θ_i are a dense subset of Θ) and then using ergodicity and the uniform continuity of $g(z_t, \theta)$ (uniform continuity follows from Assumptions 3.2 and 3.4). The second term can be made arbitrarily small for large enough T by ergodicity. Finally, the last term can be made small by exploiting the uniform continuity of g. The following theorem summarizes this result, a formal proof of which is provided in Hansen (2005).

Theorem 3.3 (Hansen, 1982b). *Suppose Assumptions 3.1–3.4 are satisfied. Then $\{\rho_T : T \geq 1\}$ in (3.18) converges almost surely to zero.*

3.3. Consistency of Extremum Estimators

Equipped with Theorem 3.3, the strong consistency of the extremum estimators discussed in Chapter 2 can be established.

3.3.1. Maximum Likelihood Estimators

Suppose that the functional form of the density function of y_t conditioned on $\vec{y}_{t-1}$, $f(y_t|\vec{y}_{t-1}^J; \beta)$, is known for all t. Let $Q_0(\beta) = E[\log f(y_t|\vec{y}_{t-1}^J; \beta)]$

denote the population criterion function and suppose that β_0, the parameter vector of the data-generating process for y_t, is a maximizer of $Q_0(\beta)$.

To show the uniqueness of β_0 as a maximizer of $Q_0(\beta)$, required by Condition (iii) of Theorem 3.2, we use Jensen's inequality to obtain

$$E\left[\log \frac{f\left(y_t|\vec{y}_{t-1}^{\,J};\beta\right)}{f\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)}\right] < \log E\left[\frac{f\left(y_t|\vec{y}_{t-1}^{\,J};\beta\right)}{f\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)}\right], \quad \beta \neq \beta_0. \tag{3.20}$$

The right-hand side of (3.20) is zero (by the law of iterated expectations) because

$$\int_{-\infty}^{\infty} \frac{f\left(y_t|\vec{y}_{t-1}^{\,J};\beta\right)}{f\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)} f\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right) dy = 1. \tag{3.21}$$

Therefore,

$$E\left[\log f\left(y_t|\vec{y}_{t-1}^{\,J};\beta\right)\right] < E\left[\log f\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)\right], \quad \text{if} \quad \beta \neq \beta_0 \tag{3.22}$$

and β_0 is the unique solution to (2.6).

The approximate sample log-likelihood function is

$$l_T(\beta) = \frac{1}{T}\sum_{t=J+1}^{T} \log f\left(y_t|\vec{y}_{t-1}^{\,J};\beta\right). \tag{3.23}$$

Thus, setting $z_t' \equiv (y_t', \vec{y}_{t-1}^{\,J\prime})$ and

$$g(z_t,\beta) = \log f\left(y_t|\vec{y}_{t-1}^{\,J};\beta\right), \tag{3.24}$$

G_T in the preceding section becomes the log-likelihood function. If Assumptions 3.1–3.4 are satisfied, then Theorem 3.3 implies the almost sure, uniform convergence of the sample log-likelihood function to $Q_0(\beta)$.[8]

3.3.2. Generalized Method of Moment Estimators

The GMM criterion function is based on the model-implied M-vector of moment conditions $E[h(z_t,\theta_0)] = 0$. With use of the sample counterpart to this expectation, the sample and population criterion functions are constructed as quadratic forms with distance matrices W_T and W_0, respectively:

[8] See DeGroot (1970) for a discussion of the use of first-moment continuity of $\log f(y_t|y_{t-1}^{J};\beta)$ in proving the strong consistency of ML estimators. DeGroot refers to first-moment continuity as "supercontinuity."

$$Q_T(\theta) = H_T(\vec{z}_T, \theta)' W_T H_T(\vec{z}_T, \theta), \tag{3.25}$$

$$Q_0(\theta) = H_0(\theta)' W_0 H_0(\theta), \tag{3.26}$$

where $H_T(\vec{z}_T, \theta) = T^{-1} \sum_{t=1}^{T} h(z_t, \theta)$ and $H_0(\theta) = E[h(z_t, \theta)]$. Since $H_0(\theta)$ is zero at θ_0, the function $Q_0(\cdot)$ achieves its minimum (zero) at θ_0.

To apply Theorem 3.3 to these criterion functions we impose an additional assumption.

Assumption 3.5. *$\{W_T : T \geq 1\}$ is a sequence of $M \times M$ positive semidefinite matrices of random variables with elements that converge almost surely to the corresponding elements of the $M \times M$ constant, positive semidefinite matrix W_0 with* $\text{rank}(W_0) \geq K$.

In addition, we let

$$\rho_T^* = \sup\{|Q_T(\theta) - Q_0(\theta)| : \theta \in \Theta\} \tag{3.27}$$

denote the maximum error in approximating Q_0 by its sample counterpart Q_T. The following lemma shows that Assumptions 3.1–3.5 are sufficient for this approximation error to converge almost surely to zero.

Lemma 3.1. *Suppose Assumptions 3.1–3.5 are satisfied. Then $\{\rho_T^* : T \geq 1\}$ converges almost surely to zero.*

Proof (Lemma 3.1). *Repeated application of the Triangle and Cauchy-Schwartz Inequalities gives*

$$\begin{aligned} |Q_T(\theta) - Q_0(\theta)| \leq\ & |H_T(\theta) - H_0(\theta)| \quad |W_T| \quad |H_T(\theta)| \\ &+ |H_0(\theta)| \quad |W_T - W_0| \quad |H_T(\theta)| \\ &+ |H_0(\theta)| \quad |W_0| \quad |H_T(\theta) - H_0(\theta)|, \end{aligned} \tag{3.28}$$

where $|W| = \left[\text{Tr } WW'\right]^{\frac{1}{2}}$. Therefore, letting $\phi_0 = \max\{|H_0(\theta)| : \theta \in \Theta\}$ and $\rho_T \equiv \sup\{|H_T(\theta) - H_0(\theta)| : \theta \in \Theta\}$,

$$0 \leq \rho_T^* \leq \rho_T |W_T| [\phi_0 + \rho_T] + \phi_0 |W_T - W_0| [\phi_0 + \rho_T] + \phi_0 |W_0| \rho_T. \tag{3.29}$$

Since $h(z_t, \theta)$ is first-moment continuous, $H_0(\theta)$ is a continuous function of θ. Therefore, ϕ_0 is finite because a continuous function on a compact set achieves its maximum. Theorem 3.3 implies that ρ_T converges almost surely to zero. Since each of the three terms on the right-hand side of (3.29) converges almost surely to zero, it follows that $\{\rho_T^* : T \geq 1\}$ converges almost surely to zero.

When this result is combined with Theorems 3.2 and 3.3, it follows that the GMM estimator $\{\theta_T : T \geq 1\}$ converges almost surely to θ_0.

3.3.3. QML Estimators

Key to consistency of QML estimators is verifying that the population moment equation (2.50) based on the normal likelihood function is satisfied at θ_0. As noted in Chapter 2, this is generally true if the functional forms of the conditional mean and variance of y_t are correctly specified (the moments implied by a DAPM are those in the probability model generating y_t). It is informative to verify that (2.50) is satisfied at θ_0 for the interest rate Example 2.1. This discussion is, in fact, generic to any one-dimensional state process y_t, since it does not depend on the functional forms of the conditional mean μ_{rt-1} or variance σ^2_{rt-1}. Extensions to the multivariate case, with some increase in notational complexity, are immediate (see, e.g., Bollerslev and Wooldridge, 1992).

Recalling the first-order conditions (2.57) shows the limit of the middle term on the right-hand-side to be

$$\frac{1}{T}\sum_{t=2}^{T}\left(\frac{(r_t-\hat{\mu}_{rt-1})^2}{\hat{\sigma}^4_{rt-1}}\frac{\partial\hat{\sigma}^2_{rt-1}}{\partial\theta_j}\right)\rightarrow E\left[\frac{(r_t-\mu_{rt-1})^2}{\sigma^4_{rt-1}}\frac{\partial\sigma^2_{rt-1}}{\partial\theta_j}\right]. \tag{3.30}$$

Using the law of iterated expectations, we find that this expectation simplifies as

$$\begin{aligned}E\left[\frac{(r_t-\mu_{rt-1})^2}{\sigma^4_{rt-1}}\frac{\partial\sigma^2_{rt-1}}{\partial\theta_j}\right] &= E\left[E\left(\frac{(r_t-\mu_{rt-1})^2}{\sigma^4_{rt-1}}\;\middle|\; r_{t-1}\right)\frac{\partial\sigma^2_{rt-1}}{\partial\theta_j}\right]\\ &= E\left[\frac{1}{\sigma^2_{rt-1}}\frac{\partial\sigma^2_{rt-1}}{\partial\theta_j}\right].\end{aligned} \tag{3.31}$$

The expectation (3.31) is seen to be minus the limit of the first term in (2.57), so the first and second terms cancel.

Thus, for the population first-order conditions associated with (2.57) to have a zero at θ_0, it remains to show that the limit of the last term in (2.57), evaluated at θ_0, is zero. This limit is

$$\frac{1}{T}\sum_{t=2}^{T}\left\{\frac{(r_t-\hat{\mu}_{rt-1})}{\hat{\sigma}^2_{rt-1}}\frac{\partial\hat{\mu}_{rt-1}}{\partial\theta_j}\right\}\rightarrow E\left[\frac{(r_t-\mu_{rt-1})}{\sigma^2_{rt-1}}\frac{\partial\mu_{rt-1}}{\partial\theta_j}\right], \tag{3.32}$$

which is indeed zero, because $E[r_t-\mu_{rt-1}|r_{t-1}]=0$ by construction and all of the other terms are constant conditional on r_{t-1}.

Consistency of the QML estimator then follows under the regularity conditions of Theorem 3.3.

3.4. Asymptotic Normality of Extremum Estimators

The consistency of θ_T for θ_0 implies that the limiting distribution of θ_T is degenerate at θ_0. For the purpose of conducting inference about the population value θ_0 of θ, we would like to know the distribution of θ_T for finite T. This distribution is generally not known, but often it can be reliably approximated using the limiting distribution of $\sqrt{T}(\theta_T - \theta_0)$ obtained by a central limit theorem. Applicable central limit theorems have been proven under a wide variety of regularity conditions. We continue our focus on stationary and ergodic economic environments.

Suppose that θ_T is strongly consistent for θ_0. To show the asymptotic normality of θ_T, we focus on the first-order conditions for the maximization or minimization of Q_T, the sample mean of the function $\mathcal{D}_0(z_t; \theta)$ first introduced in Chapter 1. More precisely, we let

$$h(z_t, \theta) = \begin{cases} \dfrac{\partial \log f}{\partial \theta}\left(y_t | \vec{y}_{t-1}^{\,J}; \theta\right) & \text{for the ML estimator,} \\ h(z_t, \theta) & \text{for the GMM estimator,} \\ \left(y_t - x_t'\theta\right) x_t & \text{for the LLP estimator.} \end{cases} \tag{3.33}$$

In each case, by appropriate choice of z_t and θ, $E[h(z_t, \theta_0)] = 0$. Thus, the function $\mathcal{D}_0(z_t; \theta)$, representing the first-order conditions for Q_0, is

$$\mathcal{D}_0(z_t; \theta) = A_0 h(z_t; \theta), \tag{3.34}$$

where the $K \times M$ matrix A_0 is

$$A_0 = \begin{cases} I_K & \text{for the ML estimator,} \\ E[\partial h(z_t, \theta_0)'/\partial \theta]\, W_0 & \text{for the GMM estimator,} \\ I_K & \text{for the LLP estimator,} \end{cases} \tag{3.35}$$

where I_K denotes the $K \times K$ identity matrix. The choice of A_0 for the GMM estimator is motivated subsequently as part of the proof of Theorem 3.5.

Using this notation and letting

$$H_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} h(z_t, \theta), \tag{3.36}$$

we can view all of these estimators as special cases of the following definition of a GMM estimator (Hansen, 1982b).

Definition 3.8. *The GMM estimator* $\{\theta_T : T \geq 1\}$ *is a sequence of random vectors that converges in probability to* θ_0 *for which* $\{\sqrt{T}A_T H_T(\theta_T) : T \geq 1\}$ *converges in probability to zero, where* $\{A_T\}$ *is a sequence of* $K \times M$ *matrices converging in probability to the full-rank matrix* A_0.

For a sequence of random variables $\{X_T\}$, convergence in distribution is defined as follows.

Definition 3.9. *Let* $F_1, F_2, \ldots,$ *be distribution functions of the random variables* $X_1, X_2, \ldots.$ *Then the sequence* $\{X_T\}$ *converges in distribution to* X *(denoted* $X_T \Rightarrow X$*) if and only if* $F_T(b) \to F_X(b)$ *for all* b *at which* F_X *is continuous.*

The classical central limit theorem examines the partial sums $\sqrt{T}S_T = (1/\sqrt{T})\sum_t X_t$ of an independently and identically distributed process $\{X_t\}$ with mean μ and finite variance. Under these assumptions, the distribution of $\sqrt{T}S_T$ converges to that of normal with mean μ and covariance matrix $\mathrm{Var}[X_t]$. However, for the study of asset pricing models, the assumption of independence is typically too strong. It rules out, in particular, persistence in the state variables and time-varying conditional volatilities.

The assumption that $\{X_t\}$ is a stationary and ergodic time series, which is much weaker than the i.i.d. assumption in the classical model, is not sufficient to establish a central limit theorem. Essentially, the problem is that an ergodic time series can be highly persistent, so that the X_t and X_s, for $s \neq t$, are too highly correlated for $\sqrt{T}S_T$ to converge to a normal random vector. The assumption of independence in the classical central limit theorem avoids this problem by assuming away any temporal dependence. Instead, we will work with the much weaker assumption that $\{X_t\}$ is a Martingale Difference Sequence (MDS), meaning that

$$E[X_t | X_{t-1}, X_{t-2}, \ldots] = 0 \tag{3.37}$$

with probability one. The assumption that X_t is mean-independent of its past imposes sufficient structure on the dependence of $\{X_t\}$ for the following central limit theorem to be true.

Theorem 3.4 (Billingsley, 1968). *Let* $\{X_t\}_{t=-\infty}^{\infty}$ *be a stationary and ergodic MDS such that* $E\left[X_1^2\right]$ *is finite. Then the distribution of* $(1/\sqrt{T})\sum_{t=1}^{T} X_t$ *approaches the normal distribution with mean zero and variance* $E\left[X_1^2\right]$.

Though many financial time series are not MDSs, it will turn out that they can be expressed as moving averages of MDS, and this will be shown to be sufficient for our purposes.

Equipped with Billingsley's theorem, under the following conditions, we can prove that the GMM estimator is asymptotically normal.

Theorem 3.5 (Hansen, 1982b). *Suppose that*

(i) $\{z_t\}$ *is stationary and ergodic.*

(ii) Θ *is an open subset of* $\mathbb{R}^K$.

(iii) h *is a measurable function of* z_t *for all* θ,

$$d_0 \equiv E\left[\frac{\partial h}{\partial \theta}(z_t, \theta_0)\right]$$

is finite and has full rank, and $\partial h/\partial\theta$ *is first moment continuous at all* $\theta \in \Theta$.

(iv) θ_T *is a GMM estimator of* θ_0.

(v) $\sqrt{T}H_T(\vec{z}_T, \theta_0) \Rightarrow N(0, \Sigma_0)$, *where* $\Sigma_0 = \lim_{T\to\infty} TE[H_T(\theta_0)H_T(\theta_0)']$.

(vi) A_T *converges in probability to* A_0, *a constant matrix of full rank, and* $A_0 d_0$ *has full rank.*

Then $\sqrt{T}(\theta_T - \theta_0) \Rightarrow N(0, \Omega_0)$, *where*

$$\Omega_0 = (A_0 d_0)^{-1} A_0 \Sigma_0 A_0' (d_0' A_0')^{-1}. \tag{3.38}$$

In proving Theorem 3.5, we will need the following very useful lemma.

Lemma 3.2. *Suppose that* $\{z_t\}$ *is stationary and ergodic and the function* $g(z_t, \theta)$ *satisfies: (a)* $E[g(z_t, \theta_0)]$ *exists and is finite, (b)* g *is first-moment continuous at* θ_0, *and suppose that* θ_T *converges to* θ_0 *in probability. Then* $(1/T)\sum_{t=1}^{T} g(z_t, \theta_T)$ *converges to* $E[g(z_t, \theta_0)]$ *in probability.*

Proof (Theorem 3.5). *When we apply Taylor's theorem on a coordinate by coordinate basis,*

$$H_T(\theta_T) = H_T(\theta_0) + G_T\left(\theta_T^*\right)(\theta_T - \theta_0), \tag{3.39}$$

where θ_T^* *is a* $K \times M$ *matrix with the* m*th column,* θ_{mT}^*, *satisfying* $|\theta_{mT}^* - \theta_0| \le |\theta_T - \theta_0|$, *for* $m = 1, \ldots, M$, *and the* ij*th element of the* $M \times K$ *matrix* $G_T(\theta_T^*)$ *is the* j*th*

element of the $1 \times K$ *vector* $\partial H_T^i(\theta_{iT}^*)/\partial\theta$. *The matrix* $G_T(\theta_T^*)$ *converges in probability to the matrix* d_0 *by Lemma 3.2. Furthermore, since* $\sqrt{T}A_T H_T(\theta_T)$ *converges in probability to zero,* $\sqrt{T}(\theta_T - \theta_0)$ *and* $[-(A_0 d_0)^{-1} A_0 \sqrt{T} H_T(\theta_0)]$ *have the same limiting distribution. Finally, from (v) it follows that* $\sqrt{T}(\theta_T - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix* $(A_0 d_0)^{-1} A_0 \Sigma_0 A_0' (d_0' A_0')^{-1}$.

A key assumption of Theorem 3.5 is Condition (v), as it takes us a long way toward the desired result. Prior to discussing applications of this theorem, it will be instructive to discuss more primitive conditions for Condition (v) to hold and to characterize Σ_0. Letting I_t denote the information set generated by $\{z_t, z_{t-1}, \ldots\}$, and $h_t \equiv h(z_t; \theta_0)$, we begin with the special case (where ACh is shorthand for autocorrelation in h):

Case ACh(0). $E[h_t|I_{t-1}] = 0$.

Since I_{t-1} includes h_s, for $s \leq t-1$, $\{h_t\}$ is an MDS. Thus, Theorem 3.4, the central limit theorem (CLT), applies directly and implies Condition (v) with

$$\Sigma_0 = E\left[h_t h_t'\right]. \tag{3.40}$$

Case ACh($n-1$). $E[h_{t+n}|I_t] = 0$, *for some* $n \geq 1$. *When* $n > 1$, *this case allows for serial correlation in the process* h_t *up to order* $n-1$.

We cannot apply Theorem 3.4 directly in this case because it presumes that h_t is an MDS. However, it turns out that we can decompose h_t into a finite sum of terms that do follow an MDS and then Billingsley's CLT can be applied. Toward this end, h_t is written as

$$h_t = \sum_{j=0}^{n-1} u_{t,j}, \tag{3.41}$$

where $u_{t,j} \in I_{t-j}$ and satisfies the property that $E[u_{t,j}|I_{t-j-1}] = 0$. This representation follows from the observation that

$$\begin{aligned} h_t &= E[h_t|I_{t-1}] + u_{t,0} \\ &= E[h_t|I_{t-2}] + u_{t,0} + u_{t,1} = \ldots = \sum_{j=0}^{n-1} u_{t,j}, \end{aligned} \tag{3.42}$$

where the law of iterated expectations has been used repeatedly. Thus,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{j=0}^{n-1} u_{t,\,j}. \tag{3.43}$$

Combining terms for which $t-j$ is the same (and, hence, that reside in the same information set) and defining

$$u_t^* = \sum_{j=0}^{n-1} u_{t+j,\,j}, \tag{3.44}$$

gives

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_t = \frac{1}{\sqrt{T}}\sum_{t=0}^{T-n+1} u_t^* + V_T^n, \tag{3.45}$$

where V_T^n involves a fixed number of $u_{t,\,j}$ depending only on n, for all T. Since V_T^n converges to zero in probability as $T \to \infty$, we can focus on the sample mean of u_t^* in deriving the limiting distribution of the sample mean of h_t.

The series $\{u_t^*\}$ is an MDS. Thus, Billingsley's theorem implies that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T-n+1} u_t^* \Rightarrow N(0,\,\Sigma_0), \qquad \Sigma_0 = E\big[u_t^* u_t^{*\prime}\big]. \tag{3.46}$$

Moreover, substituting the left-hand side of (3.45) for the scaled average of the $u_t^{*\prime}$s in (3.46), gives

$$\begin{aligned}\Sigma_0 &= \lim_{T\to\infty} E\left[\frac{1}{T}\left(\sum_{t=1}^{T} h_t\right)\left(\sum_{t=1}^{T} h_t'\right)\right] \\ &= \lim_{T\to\infty}\sum_{j=-n+1}^{n-1}\left(\frac{T-|j|}{T}\right)E\big[h_t h_{t-j}'\big] = \sum_{j=-n+1}^{n-1} E\big[h_t h_{t-j}'\big].\end{aligned} \tag{3.47}$$

In words, the asymptotic covariance matrix of the scaled sample mean of h_t is the sum of the autocovariances of h_t out to order $n-1$.

Case AC*h*(∞). *$E[h_t h_{t-s}'] \neq 0$, for all s.*

Since, in case AC$h(n-1)$, $n-1$ is the number of nonzero autocovariances of h_t, (3.47) can be rewritten equivalently as

$$\Sigma_0 = \sum_{j=-\infty}^{\infty} E\big[h(z_t, \theta_0)h(z_{t-j}, \theta_0)'\big]. \tag{3.48}$$

This suggests that, for the case where $E[h_t h_{t-s}'] \neq 0$, for all s (i.e., $n = \infty$), (3.48) holds as well. Hansen (1982b) shows that this is indeed the case under the additional assumption that the autocovariance matrices of h_t are absolutely summable.

3.5. Distributions of Specific Estimators

In applying Theorem 3.5, it must be verified that the problem of interest satisfies Conditions (iii), (iv), and (v). We next discuss some of the implications of these conditions for the cases of the ML, GMM, and LLP criterion functions. In addition, we examine the form of the asymptotic covariance matrix Σ_0 implied by these criterion functions, and discuss consistent estimators of Σ_0.

3.5.1. Maximum Likelihood Estimation

In the case of ML estimation, we proved in Chapter 2 that

$$E\left[\mathcal{D}_0\big(y_t, \vec{y}_{t-1}^{\,J}, \beta_0\big)\big|\vec{y}_{t-1}^{\,J}\right] = 0,$$

where[9]

$$\mathcal{D}_0\big(y_t, \vec{y}_{t-1}^{\,J}, \beta\big) = \frac{\partial \log f}{\partial \beta}\big(y_t\big|\vec{y}_{t-1}^{\,J}; \beta\big). \tag{3.49}$$

Since the density of y_t conditioned on $\vec{y}_{t-1}^{\,J}$ is the same as the density conditioned on $\vec{y}_{t-1}$ by assumption, (2.7) implies that the "score" (3.49) is an MDS. Therefore, Theorem 3.4 and Case ACh(0) apply and

$$\sqrt{T}H_T(z_t, \theta_0) = \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f}{\partial \beta}\big(y_t\big|\vec{y}_{t-1}^{\,J}; \beta_0\big)\right) \tag{3.50}$$

[9] In deriving this result, we implicitly assumed that we could reverse the order of integration and differentiation. Formally, this is justified by the assumption that the partial derivative of $\log f(y_t|\vec{y}_{t-1}^{\,J}; \beta)$ is first-moment continuous at β_0. More precisely, consider a function $h(z, \theta)$. Suppose that for some $\delta > 0$, the partial derivative $\partial h(z, \theta)/\partial\theta$ exists for all values of z and all θ such that $|\theta - \theta_0| < \delta$, and suppose that this derivative is first-moment continuous at θ_0. If $E[h(z, \theta)]$ exists for all $|\theta - \theta_0| < \delta$ and if $E[|\partial h(z, \theta)/\partial\theta|] < \infty$, then

$$E\left[\frac{\partial h(z, \theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right] = \frac{\partial E[h(z, \theta)]}{\partial \theta}\Big|_{\theta=\theta_0}.$$

converges in distribution to a normal random vector with asymptotic covariance matrix

$$\Sigma_0 = E\left[\frac{\partial \log f}{\partial \beta}\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)\frac{\partial \log f}{\partial \beta}\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)'\right]. \tag{3.51}$$

Furthermore, the first-order conditions to the log-likelihood function give K equations in the K unknowns (β), so A_T is I_K in this case and $\Omega_0^{\mathrm{ML}} = d_0^{-1}\Sigma_0(d_0')^{-1}$.

Thus, it remains to determine d_0. Since $E[\mathcal{D}_0(z_t, \beta_0)] = 0$, differentiating both sides of this expression with respect to β gives[10]

$$\begin{aligned} d_0^{\mathrm{ML}} &= E\left[\frac{\partial^2 \log f}{\partial \beta \partial \beta'}\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)\right] \\ &= -E\left[\frac{\partial \log f}{\partial \beta}\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)\frac{\partial \log f}{\partial \beta}\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)'\right]. \end{aligned} \tag{3.52}$$

When we combine (3.38), (3.51), and (3.52) and use the fact that if $X \sim N(0, \Sigma_X)$, then $AX \sim N(0, A\Sigma_X A')$, it follows that

$$\sqrt{T}(b_T^{\mathrm{ML}} - \beta_0) \Rightarrow N\left(0, -E\left[\frac{\partial^2 \log f}{\partial \beta \partial \beta'}\left(y_t|\vec{y}_{t-1}^{\,J};\beta_0\right)\right]^{-1}\right). \tag{3.53}$$

In actual implementations of ML estimation, the asymptotic covariance in (3.53) is replaced by its sample counterpart. From (3.52) it follows that this matrix can be estimated either as the inverse of the sample mean of the "outer product" of the likelihood scores or as minus the inverse of the sample mean of the second-derivative matrix evaluated at b_T^{ML},

$$\left(-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \log f}{\partial \beta \partial \beta'}\left(y_t|\vec{y}_{t-1}^{\,J};\, b_T^{\mathrm{ML}}\right)\right)^{-1}. \tag{3.54}$$

[10] The second equality in (3.52) is an important property of conditional density functions that follows from (3.49). By definition, (3.49) can be rewritten as

$$0 = \int \frac{\partial \log f}{\partial \beta}\left(y_t\middle|\vec{y}_{t-1}^{\,J};\beta_0\right) f\left(y_t\middle|\vec{y}_{t-1}^{\,J};\beta_0\right) dy_t.$$

Differentiating under the integral sign and using the chain rule gives

$$0 = E\left[\frac{\partial^2 \log f}{\partial \beta \partial \beta'}\left(y_t\middle|\vec{y}_{t-1}^{\,J};\beta_0\right)\right] + E\left[\frac{\partial \log f}{\partial \beta}\left(y_t\middle|\vec{y}_{t-1}^{\,J};\beta_0\right)\frac{\partial \log f}{\partial \beta}\left(y_t\middle|\vec{y}_{t-1}^{\,J};\beta_0\right)'\right].$$

Assuming that the regularity conditions for Lemma 3.2 are satisfied by the likelihood score, we see that (3.54) converges to the covariance matrix of b_T^{ML} as $T \to \infty$.

The asymptotic covariance matrix of b_T^{ML} is the Cramer-Rao lower bound, the inverse of the so-called Hessian matrix. This suggests that, even though the ML estimator may be biased in small samples, as T gets large, the ML estimator is the most efficient estimator in the sense of having the smallest asymptotic covariance matrix among all consistent estimators of β_0. This is indeed the case and we present a partial proof of this result in Section 3.6.

3.5.2. GMM Estimation

Theorem 3.5 applies directly to the case of GMM estimators. The GMM estimator minimizes (3.25) so the regularity conditions for Theorem 3.5 require that $h(z_t, \theta)$ be differentiable, $\partial h(z_t, \theta)/\partial\theta$ be first-moment continuous at θ_0, and that W_T converge in probability to a constant, positive-semidefinite matrix W_0.

The first-order conditions to the minimization problem (3.25) are

$$\frac{\partial H_T(\theta_T)'}{\partial\theta} W_T H_T(\theta_T) = 0. \tag{3.55}$$

Therefore, the A_T implied by the GMM criterion function (3.25) is

$$A_T = \frac{\partial H_T(\theta_T)'}{\partial\theta} W_T. \tag{3.56}$$

By Lemma 3.2 and the assumption that W_T converges to W_0, it follows that A_T converges in probability to $A_0 = d_0' W_0$. Substituting this expression into (3.38), we conclude that $\sqrt{T}(\theta_T - \theta_0)$ converges in distribution to a normal with mean zero and covariance matrix

$$\Omega_0^{\mathrm{GMM}} = \left(d_0' W_0 d_0\right)^{-1} d_0' W_0 \Sigma_0 W_0' d_0 \left(d_0' W_0' d_0\right)^{-1}. \tag{3.57}$$

If the probability limit of the distance matrix defining the GMM criterion function is chosen to be $W_0 = \Sigma_0^{-1}$, then (3.57) simplifies to

$$\Omega_0^{\mathrm{GMM}} = \left(d_0' \Sigma_0^{-1} d_0\right)^{-1}. \tag{3.58}$$

We show in Section 3.6 that this choice of distance matrix is the optimal choice among GMM estimators constructed from linear combinations of the moment equation $E[h(z_t, \theta_0)] = 0$.

A consistent estimator of Ω_0^{GMM} is constructed by replacing all of the matrices in (3.57) or (3.58) by their sample counterparts. The matrix W_0

is estimated by W_T, the matrix used to construct the GMM criterion function, and d_0 is replaced by $\partial H_T(\theta_T)/\partial\theta$. The construction of a consistent estimator of Σ_0 depends on the degree of autocorrelation in $h(z_t, \theta_0)$. In Case AC$h(n-1)$, with finite n, the autocovariances of h comprising Σ_0 are replaced by their sample counterparts using fitted $h(z_t, \theta_T)$ in place of $h(z_t, \theta_0)$:

$$\frac{1}{T}\sum_{t=j+1}^{T} h(z_t, \theta_T)h(z_{t-j}, \theta_T)'. \tag{3.59}$$

An asymptotically equivalent estimator is obtained by subtracting the sample mean from $h(z_t, \theta_T)$ before computing the sample autocovariances.

If, on the other hand, $n = \infty$ or n is very large relative to the sample size T, then an alternative approach to estimating Σ_0 is required. In Case AC$h(\infty)$, Σ_0 is given by (3.48). Letting $\Gamma_{h0}(j) = E[h_t h'_{t-j}]$, we proceed by constructing an estimator Σ_T as a weighted sum of the autocovariances that can feasibly be estimated with a finite sample of length T:

$$\Sigma_T = \frac{T}{T-K}\sum_{j=-T+1}^{T-1} k\left(\frac{j}{B_T}\right)\Gamma_{hT}(j), \tag{3.60}$$

where the sample autocovariances are given by

$$\Gamma_{hT}(j) = \begin{cases} \dfrac{1}{T}\displaystyle\sum_{t=j+1}^{T} h(z_t, \theta_T)h(z_{t-j}, \theta_T)' & \text{for } j \geq 0, \\ \dfrac{1}{T}\displaystyle\sum_{t=-j+1}^{T} h(z_{t+j}, \theta_T)h(z_t, \theta_T)' & \text{for } j < 0, \end{cases} \tag{3.61}$$

and B_T is a "bandwidth" parameter discussed later. The scaling factor $T/(T-K)$ is a small-sample adjustment for the estimation of θ.

The function $k(\cdot)$, called a *kernel*, determines the weight given to past sample autocovariances in constructing Σ_T. The basic idea of this estimation strategy is that, for fixed j, sample size must increase to infinity for $\Gamma_{hT}(j)$ to be a consistent estimator of $\Gamma_{h0}(j)$. At the same time, the number of nonzero autocovariances in (3.60) must increase without bound for Σ_T to be a consistent estimator of Σ_0. The potential problem is that if terms are added proportionately as T gets large, then the number of products $h_t h'_{t-j}$ in the sample estimate of $\Gamma_{hT}(j)$ stays small regardless of the size of T. To avoid this problem, the kernel must be chosen so that the number of autocovariances included grows, but at a slower rate than T,

so that the number of terms in each sample estimate $\Gamma_{hT}(j)$ increases to infinity.

Two popular kernels for estimating Σ_0 are

$$\text{Truncated} \quad k(x) = \begin{cases} 1 & \text{for } |x| \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.62}$$

$$\text{Bartlett} \quad k(x) = \begin{cases} 1 - |x| & \text{for } |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.63}$$

For both of these kernels, the bandwidth B_T determines the number of autocovariances included in the estimation of Σ_T. In the case of the truncated kernel, all lags out to order B_T are included with equal weight. This is the kernel studied by White (1984). In the case of the Bartlett kernel, the autocovariances are given declining weights out to order $j \leq B_T$. Newey and West (1987b) show that, by using declining weights, the Bartlett kernel guarantees that Σ_T is positive-semidefinite. This need not be the case in finite samples for the truncated kernel. The choice of the bandwidth parameter B_T is discussed in Andrews (1991).

3.5.3. Quasi-Maximum Likelihood Estimation

The QML estimator is a special case of the GMM estimator. Specifically, continuing our discussion of the scalar process r_t with conditional mean μ_{rt-1} and variance σ^2_{rt-1} that depend on the parameter vector θ, let the jth component of $h(z_t, \theta)$ be the score associated with θ_j:

$$h_j(z_t, \theta) \equiv -\frac{1}{2\sigma^2_{rt-1}(\theta)} \frac{\partial \sigma^2_{rt-1}(\theta)}{\partial \theta_j} + \frac{1}{2} \frac{(r_t - \mu_{rt-1}(\theta))^2}{\sigma^4_{rt-1}(\theta)} \frac{\partial \sigma^2_{rt-1}(\theta)}{\partial \theta_j}$$

$$+ \frac{(r_t - \mu_{rt-1}(\theta))}{\sigma^2_{rt-1}(\theta)} \frac{\partial \mu_{rt-1}(\theta)}{\partial \theta_j}, \quad j = 1, \ldots, K. \tag{3.64}$$

The asymptotic distribution of the QML estimator is thus determined by the properties of $h(z_t, \theta_0)$. From (3.64) it is seen that $E[h_j(z_t, \theta_0)|I_{t-1}] = 0$; that is, $\{h(z_t, \theta_0)\}$ is an MDS. This follows from the observations that, after taking conditional expectations, the first and second terms cancel and the third term has a conditional mean of zero.

Therefore, the QML estimator θ_T^{QML} falls under Case ACh(0) with $M = K$ (the number of moment equations equals the number of parameters) and

$$\sqrt{T}\left(\theta_T^{\text{QML}} - \theta_0\right) \Rightarrow N\left(0, \left(d_0^{\text{QML}}\right)^{-1} \Sigma_0 \left(d_0^{\text{QML}}\right)^{-1}\right), \tag{3.65}$$

where $\Sigma_0 = E[h(z_t, \theta_0)h(z_t, \theta_0)']$, with h given by (3.64), and

$$d_0^{\text{QML}} = E\left[\frac{\partial^2 \log f_N}{\partial\theta\,\partial\theta'}(r_t|I_{t-1}; \theta_0)\right]. \tag{3.66}$$

Though these components are exactly the same as in the case of full-information ML estimation, d_0^{QML} and Σ_0 are not related by (3.52), so (3.65) does not simplify further.

3.5.4. Linear Least-Squares Projection

The LLP estimator is the special case of the GMM estimator with $z_t' = (y_t, x_t')$, $h(z_t, \delta) = (y_t - x_t'\delta)x_t$, $A_0 = I_K$. Also,

$$d_0^{\text{LLP}} = -E\left[x_t x_t'\right] \tag{3.67}$$

and with $u_t \equiv (y_t - x_t'\delta_0)$, where δ_0 is the probability limit of the least-squares estimator δ_T,

$$\Sigma_0 = \sum_{j=-\infty}^{\infty} E\left[x_t u_t u_{t-j} x_{t-j}'\right]. \tag{3.68}$$

It follows that

$$\Omega_0^{\text{LLP}} = E\left[x_t x_t'\right]^{-1} \sum_{j=-\infty}^{\infty} E\left[x_t u_t u_{t-j} x_{t-j}'\right] E\left[x_t x_t'\right]^{-1}. \tag{3.69}$$

In order to examine several special cases of LLP for forecasting the future, we assume that the variable being forecasted is dated $t + n$, $n \geq 1$, and let x_t denote the vector of forecast variables observed at date t:

$$y_{t+n} = x_t'\delta_0 + u_{t+n}. \tag{3.70}$$

We consider several different assumptions about the projection error u_{t+n}. Unless otherwise noted, throughout the following discussion, the information set I_t denotes the information generated by current and past x_t and u_t.

Consider first Case ACh(0) with $n = 1$ and $E[u_{t+1}|I_t] = 0$. One circumstance where this case arises is when a researcher is interested in testing whether y_{t+1} is unforecastable given information in I_t (see Chapter 1). For instance, if we assume that x_t includes the constant 1 as the first component, and partitioning x_t as $x_t' = (1, \tilde{x}_t')$ and δ_0 conformably as $\delta_0' = (\delta_c, \delta_{\tilde{x}}')$, then this case implies that $E[y_{t+1}|I_t] = \delta_c$, $\delta_{\tilde{x}} = 0$, and y_{t+1} is unforecastable given past information about $\tilde{x}_t$ and y_t. The alternative hypothesis is that

$$E[y_{t+1}|I_t] = \delta_c + x_t'\delta_{\tilde{x}}, \tag{3.71}$$

with the (typical) understanding that the projection error under this alternative satisfies $E[u_{t+1}|I_t] = 0$. A more general alternative would allow $\delta_{\tilde{x}} \neq 0$ and the projection error u_{t+1} to be correlated with other variables in I_t. We examine this case later.

Since $d_0 = -E[x_t x_t']$ and this case fits into Case ACh(0),

$$\sqrt{T}(\delta_T - \delta_0) \Rightarrow N\big(0, \Omega_0^{\mathrm{LLP}}\big), \tag{3.72}$$

where

$$\Omega_0^{\mathrm{LLP}} = E\big[x_t x_t'\big]^{-1} E\big[u_{t+1}^2 x_t x_t'\big] E\big[x_t x_t'\big]^{-1}. \tag{3.73}$$

Without further assumptions, Ω_0^{LLP} does not simplify. One simplifying assumption that is sometimes made is that the variance of u_{t+1} conditioned on I_t is constant:

$$E\big[u_{t+1}^2|I_t\big] = \sigma_u^2, \text{ a constant.} \tag{3.74}$$

Under this assumption, Σ_0 in (3.73) simplifies to ${\sigma_u}^2 E[x_t x_t']$ and

$$\Omega_0^{\mathrm{LLP}} = \sigma_u^2 E\big[x_t x_t'\big]^{-1}. \tag{3.75}$$

These characterizations of Ω_0^{LLP} are not directly applicable because the asymptotic covariance matrices are unknown (are functions of unknown population moments). Therefore, we replace these unknown moments with their sample counterparts. Let $\hat{u}_{t+1} \equiv (y_{t+1} - x_t'\delta_T)$. With the homoskedasticity assumption (3.74), the distribution of δ_T used for inference is

$$\delta_T \approx N\big(\delta_0, \Omega_T^{\mathrm{LLP}}\big), \tag{3.76}$$

where

$$\Omega_T^{\mathrm{LLP}} = \hat{\sigma}_u^2 \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}, \tag{3.77}$$

with $\hat{\sigma}_u^2 = (1/T)\sum_{t=1}^{T} \hat{u}_t^2$. This is, of course, the usual distribution theory used in the classical linear least-squares estimation problem. Letting $\hat{\sigma}_{\delta^i}$ denote the ith diagonal element of (3.77), we can test the null hypothesis $H_0 : \delta_0^i = \delta_0^{i*}$ using the distribution

$$\frac{\delta_T^i - \delta_0^{i*}}{\hat{\sigma}_{\delta^i}} \approx N(0, 1). \tag{3.78}$$

Suppose that we relax assumption (3.74) and let the conditional variance of u_{t+1} be time varying. Then Ω_0^{LLP} is given by (3.73) and now Σ_0 is estimated by

$$\Sigma_T = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{t+1}^2 x_t x_t', \tag{3.79}$$

and

$$\Omega_T^{\text{LLP}} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \Sigma_T \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}. \tag{3.80}$$

Testing proceeds as before, but with a different calculation of $\hat{\sigma}_{\delta^i}$.

Next, consider Case AC$h(n-1)$ which has $n > 1$ and $E[u_{t+n}|I_t] = 0$. This case would arise, for example, in asking the question whether y_{t+n} is forecastable given information in I_t. For this case, d_0 is unchanged, but the calculation of Σ_0 is modified so that

$$\Omega_0^{\text{LLP}} = E\left[x_t x_t'\right]^{-1} \left(\sum_{j=-n+1}^{n-1} E\left[u_{t+n} u_{t+n-j} x_t x_{t-j}'\right]\right) E\left[x_t x_t'\right]^{-1}. \tag{3.81}$$

Analogously to the case AC$h(0)$, this expression simplifies further if the conditional variances and autocorrelations of u_t are constants.

To estimate the asymptotic covariance matrix for this case, we replace $E[x_t x_t']$ by $(1/T)\sum_{t=1}^{T} x_t x_t'$ and Σ_0 by

$$\Sigma_T = \sum_{j=-n+1}^{n-1} \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{t+n}\, \hat{u}_{t+n-j}\, x_t x_{t-j}'. \tag{3.82}$$

Testing proceeds in exactly the same way as before.

3.6. Relative Efficiency of Estimators

The efficiency of an estimator can only be judged relative to an a priori set of restrictions on the joint distribution of the z_t that are to be used in estimation. These restrictions enter the formulation of a GMM estimator in two ways: through the choices of the h function and the A_0. The form of the asymptotic covariance matrix Ω_0 in (3.38) shows the dependence of

the limiting distribution on both of these choices. In many circumstances, a researcher will have considerable latitude in choosing either A_0 or $h(z_t, \theta)$ or both. Therefore, a natural question is: Which is the most *efficient* GMM estimator among all admissible estimators? In this section, we characterize the optimal GMM estimator, in the sense of being most efficient or, equivalently, having the smallest asymptotic covariance matrix among all estimators that exploit the same information about the distribution of $\vec{z}_T$.

3.6.1. GMM Estimators

To highlight the dependence of the distributions of GMM estimators on the information used in specifying the moment equations, it is instructive to start with the conditional version of the moment equations underlying the GMM estimation,

$$\frac{1}{T}\sum_{t=1}^{T} A_t h(z_t, \theta_T) = 0, \tag{3.83}$$

where A_t is a (possibly random) $K \times M$ matrix in the information set I_t, and $h(z_t, \theta)$ is an $M \times 1$ vector, with $K \leq M$, satisfying

$$E\big[h(z_t; \theta_0)\big|I_t\big] = 0. \tag{3.84}$$

In this section, we will treat z_t as a generic random vector that is not presumed to be in I_t and, indeed, in all of the examples considered subsequently $z_t \notin I_t$.

Initially, we treat $h(z_t; \theta)$ as given by the asset pricing theory and, as such, not subject to the choice of the researcher. We also let

$$\mathcal{A} = \left\{A_t \in I_t, \text{ such that } E\left[A_t \frac{\partial h(z_t; \theta_0)}{\partial \theta}\right] \text{ has full rank}\right\} \tag{3.85}$$

denote the class of admissible GMM estimators, where each estimator is indexed by the (possibly random) weights A_t. The efficiency question at hand is: In estimating θ_0, what is the optimal choice of A_t? (Which choice of A_t gives the smallest asymptotic covariance matrix for θ_T among all estimators based on matrices in $\mathcal{A}$?) The following lemma, based on the analysis in Hansen (1985), provides a general characterization of the optimal $A^* \in \mathcal{A}$.

Lemma 3.3. *Suppose that the assumptions of Theorem 3.5 are satisfied and* $\{A_t\} \in \mathcal{A}$ *is a stationary and ergodic process (jointly with* z_t*). Then the optimal choice* $A^* \in \mathcal{A}$ *satisfies*

$$\lim_{T\to\infty} T\,E\left[\left(\frac{1}{T}\sum_{t=1}^{T} A_t h(z_t,\theta_0)\right)\left(\frac{1}{T}\sum_{t=1}^{T} A_t^* h(z_t,\theta_0)\right)'\right]$$
$$= E\left[A_t \frac{\partial h(z_t;\theta_0)}{\partial\theta}\right] \equiv d_A. \tag{3.86}$$

Proof (Lemma 3.3). *Using arguments analogous to those in Theorem 3.5, one can show the asymptotic covariance matrix of any estimator $A \in \mathcal{A}$ to be*

$$\Omega_0^A = E\left[A_t \frac{\partial h(z_t;\theta_0)}{\partial\theta}\right]^{-1} \Sigma_0^A E\left[\frac{\partial h(z_t;\theta_0)'}{\partial\theta} A_t'\right]^{-1}, \tag{3.87}$$

where

$$\Sigma_0^A = \sum_{j=-\infty}^{\infty} E\big[A_t h(z_t;\theta_0) h(z_{t-j};\theta_0)' A_{t-j}'\big]. \tag{3.88}$$

Define

$$D_T^A \equiv E\left[A_t \frac{\partial h(z_t;\theta_0)}{\partial\theta}\right]^{-1} \frac{1}{\sqrt{T}} \sum A_t h(z_t;\theta_0)$$
$$- E\left[A_t^* \frac{\partial h(z_t;\theta_0)}{\partial\theta}\right]^{-1} \frac{1}{\sqrt{T}} \sum A_t^* h(z_t;\theta_0), \tag{3.89}$$

and note that under assumption (3.86),

$$\lim_{T\to\infty} E\left[D_T^A \left(\frac{1}{\sqrt{T}} \sum h(z_t;\theta_0)' A_t^{*\prime}\right)\right] = 0. \tag{3.90}$$

It follows immediately that

$$\Omega_0^A = \lim_{T\to\infty} E\big[D_T^A D_T^{A\prime}\big] + \Omega_0^{A*}. \tag{3.91}$$

Since $E\big[D_T^A D_T^{A\prime}\big]$ is positive-semidefinite for all T, the lemma follows.

In applying Lemma 3.3 to our various estimation environments, consider first our basic GMM problem, where an asset pricing theory implies that $E[h(z_t,\theta_0)] = 0$. In this case, all we know about the distribution of z_t is that the unconditional mean of h is zero. Therefore, in examining the optimality question we must restrict attention to constant $A_t = A_0$, for all

t. In other words, the optimality question in this case is simply the optimal choice of the $K \times M$ matrix of constants A_0. Optimality condition (3.86) in this case is

$$A_0 \Sigma_0 A_0^{*\prime} = A_0 d_0, \tag{3.92}$$

and substitution verifies that $A^* \equiv d_0' \Sigma_0^{-1}$. This is the "optimal" GMM estimator proposed by Hansen (1982b), which gives the asymptotic covariance matrix

$$\Omega_0^{\text{FGMM}} = \left(d_0' \Sigma_0^{-1} d_0\right)^{-1}. \tag{3.93}$$

The superscript FGMM indicates that this is the asymptotic covariance matrix for the optimal GMM estimator based on the fixed set of moment conditions $E[h(z_t, \theta_0)] = 0$.

To relate this observation back to the standard GMM criterion function, expressed as a quadratic form in $H_T(\theta)$, recall that $A_0 = d_0' W_0$, where W_0 is the distance matrix in the GMM criterion function. It follows immediately that the optimal GMM estimator is obtained by setting $W_0 = \Sigma_0^{-1}$. Intuitively, since Σ_0 is the asymptotic covariance matrix of the sample moment $H_T(\theta_0)$, this choice of W_0 gives the most weight to those moment conditions that are most precisely estimated in the sense of having a small (asymptotic) variance.

Next, suppose that the asset pricing theory under investigation provides the stronger restriction that

$$E[h(z_t; \theta_0) | I_{t-1}] = 0, \tag{3.94}$$

where I_{t-1} includes current and past values of z_{t-1}. Here we have changed notation to make precise the model's implication that $z_t \in I_t$ and, given the conditioning in (3.94) on I_{t-1}, the admissible weight matrices A_{t-1} must be in I_{t-1}. Once again, we presume that the function h is fixed by the theory. However, in contrast to the previous case, since any $A_{t-1} \in I_{t-1}$ satisfies $E[A_{t-1} h(z_t; \theta_0)] = 0$, we have considerable latitude in choosing $\{A_t\}$. A_{t-1} can be an essentially arbitrary function of z_{t-1} and its history, and possibly the history of other variables in I_{t-1}. In spite of this latitude, there is a solution to the problem of choosing the optimal A_{t-1}. Direct substitution into (3.86) shows that

$$A_{t-1}^* = E\left[\frac{\partial h(z_t; \theta_0)'}{\partial \theta} \,\middle|\, I_{t-1}\right] \times E\left[h(z_t; \theta_0) h(z_t; \theta_0)' \middle| I_{t-1}\right]^{-1}. \tag{3.95}$$

Substituting A_{t-1}^* into (3.87), using the simplified notation $h_t \equiv h(z_t, \theta_0)$, gives the asymptotic covariance matrix

$$\Omega_0^{\text{OGMM}} = E\left(E\left[\frac{\partial h_t'}{\partial\theta}\;\middle|\; I_{t-1}\right]E\big[h_t h_t'|I_{t-1}\big]^{-1}E\left[\frac{\partial h_t}{\partial\theta}\;\middle|\; I_{t-1}\right]\right)^{-1}. \quad (3.96)$$

See Hansen (1985) and Hansen et al. (1988) for further discussions of this case.

3.6.2. LLP Estimators

When we set out to estimate the coefficients of the optimal linear forecast of y_{t+n} based on x_t, we proceeded using the sample counterpart to the moment equation

$$E\big[\big(y_{t+n} - x_t'\delta_0\big)x_t\big] = 0, \quad (3.97)$$

which defines the LLP. If all we know about the relation between y_{t+n} and x_t is that (3.97) is satisfied (by construction), then essentially the only estimation strategy available to the econometrician is LLP and the sample counterpart to (3.97) is solved for the least-squares estimator δ_T.

However linear DAPMs often imply that $E[u_{t+n}|I_t] = 0$ and, hence, that u_{t+n} is orthogonal to any (measurable) function of x_t and not just x_t. This leads immediately to the question of whether there is a more efficient estimator of δ_0 that exploits additional orthogonality conditions beyond (3.97). Hansen (1985) shows that the answer to this question is yes. Using his results and those in several subsequent papers, we examine this optimality question for two special cases: (1) $n = 1$ and u_{t+1} is (possibly) conditionally heteroskedastic, and (2) $n > 1$ and u_{t+n} is conditionally homoskedastic.

If $n = 1$, then we set $u_{t+1} = h(z_{t+1}, \delta_0) = (y_{t+1} - \delta_0' x_t)$ and construct the optimal GMM estimator based on the moment equation $E[u_{t+1}|I_t] = 0$. Instead of using the orthogonality of u_{t+1} and x_t to define the LLP estimator, we consider the larger class of estimators based on the moment equation

$$\frac{1}{T}\sum_{t=1}^{T} A_t u_{t+1}, \quad (3.98)$$

where A_t is a $K \times 1$ vector whose elements are in I_t. Least squares is the optimal estimator of δ_0 if the optimal choice of A_t is $A_t^* = x_t$. From (3.95) it is seen that the A_t^* for the conditional moment restriction $E[u_{t+1}|I_t] = 0$ is constructed from the two components,

$$E\left[\frac{\partial h(z_{t+1}, \delta_0)}{\partial\delta}\;\middle|\; I_t\right] = -x_t, \quad \sigma_{ut}^2 \equiv E\big[u_{t+1}^2\big|I_t\big], \quad (3.99)$$

which gives

$$A_t^* = x_t/\sigma_{ut}^2. \quad (3.100)$$

Thus, instead of ordinary least squares, the optimal estimator is obtained by scaling the regressors by the inverse of the conditional variance of u_{t+1}. To interpret this result, note that the population orthogonality condition used in estimation with A^* can be rewritten as

$$E\left[\left(\frac{y_{t+1}}{\sigma_{ut}} - \frac{x_t'}{\sigma_{ut}}\delta_0\right)\frac{x_t}{\sigma_{ut}}\right] = 0. \tag{3.101}$$

This is the moment equation obtained from the population least-squares objective function for the projection equation that is scaled by $1/\sigma_{ut}$:

$$\frac{y_{t+1}}{\sigma_{ut}} = \frac{x_t'}{\sigma_{ut}}\delta_0 + \frac{u_{t+1}}{\sigma_{ut}}, \tag{3.102}$$

or what is commonly referred to as *generalized least squares*. In practice, the optimal estimator is obtained by first scaling each observation by a consistent estimator of σ_{ut}, $\hat{\sigma}_t^2$ constructed using fitted residuals, and then proceeding with a standard linear projection. The asymptotic covariance matrix of this estimator is

$$\Omega_0^{*\text{LLP}} = E\left[\frac{x_t x_t'}{\sigma_{ut}^2}\right]^{-1}. \tag{3.103}$$

The reason for scaling becomes clear when we recognize that, if $n = 1$ and $\sigma_{ut}^2 = \sigma_u^2$, a constant (homoskedasticity), then least squares is the optimal estimation strategy; $A_t^* = x_t$. Thus, in the presence of heteroskedasticity, we first scale the regression equation to arrive at a homoskedastic model and then implement the optimal estimator for this case, least-squares.

From a practical point of view, implementation of the optimal GMM estimator based on A_t^* requires an estimate of σ_{ut}^2 (i.e., an estimate of how σ_{ut}^2 depends on I_t). Our point of departure in constructing the optimal GMM estimator was that all we know from our DAPM is that u_{t+1} is mean-independent of I_t. Thus this optimality result based on (3.94) alone is a "limited-information" result in that it holds using any consistent estimator of σ_{ut}^2, including a nonparametric estimator (that does not presume knowledge of the functional form of σ_{ut}^2).

Of course, if it is known that σ_{ut}^2 is given, say, by $g(x_t, \gamma_0)$, then this information can be used in constructing A^*. However, in this case, there is the additional moment equation

$$E\big[u_{t+1}^2 - g(x_t, \gamma_0)\big|I_t\big] = 0, \tag{3.104}$$

which can be used in estimation. Indeed, the moment equations (3.94) and (3.104) can be combined and, using similar arguments, the associated optimal GMM estimator can be derived. In general, this is a more efficient

estimator than the optimal estimator based on (3.94) alone. The weights for the optimal estimator based on both (3.94) and (3.104) involve the fourth conditional moments of u_{t+1}, which are also unknown (without further assumptions). So, with knowledge of (3.104), the limited information is pushed back to fourth moments.

These observations suggest an intermediate, suboptimal estimation strategy that might lead to some efficiency gains over a naive GMM estimator that completely ignores the nature of the A^*. For instance, if the functional form of g in (3.104) is unknown, then reasonable efficiency might be achieved by projecting u_{t+1}^2 onto variables that are expected to influence σ_{ut}^2 and then scaling the model by the square root of the fitted values from this projection. If this strategy is pursued, then the asymptotic covariance matrix should be constructed without assuming that the conditional variance of the scaled u_t is 1, since knowledge of the correct functional form for σ_{ut}^2 has not been assumed.

Up to this point, we have been discussing best forecasts that are linear with forecast errors that are mean-independent of an information set I_t. So it is of interest to inquire how the preceding discussion is altered for the case of LLP of y_t onto x_t,

$$y_t = x_t'\delta_0 + u_t, \tag{3.105}$$

where $x_t'\delta_0$ is the best linear, not best, predictor. Without further assumptions on the properties of u_t, the asymptotic covariance matrix of the LLP estimator θ_T is given by (3.69). Analogously to the case of heteroskedasticity, one would expect that a more efficient estimator of δ_0 than the least-squares estimator could be obtained given some knowledge of the form of any serial correlation in u_t. To highlight one widely studied example of such an efficiency gain, we will focus on an example of serial correlation with homoskedastic errors.

Consider the case where it is known a priori that u_t follows an autoregressive process of order p ($AR(p)$):

$$u_t = \rho_1 u_{t-1} + \ldots + \rho_p u_{t-p} + \epsilon_t \tag{3.106}$$

or, equivalently, $\rho(L)u_t = \epsilon_t$, where L is the lag operator ($L^s x_t = x_{t-s}$), and $\rho(L)$ is the polynomial

$$\rho(L) = 1 - \rho_1 L - \ldots - \rho_p L^p. \tag{3.107}$$

The roots of the polynomial (3.107) are assumed to lie outside the unit circle in the complex plane. Consistent with most treatments of serial correlation in the classical setting, it is assumed that

$$E\big[u_t \big| J\big] = 0, \quad J = \{x_t, x_{t\pm1}, x_{t\pm2}, \ldots\}, \tag{3.108}$$

and that $E[\epsilon_t | J] = \sigma_\epsilon^2$, a constant. Finally, the assumption that u_t is correctly specified as an AR(p) process is captured in the assumption that $E[\epsilon_t | J_{t-1}^*] = 0$, where $J_{t-1}^* = \{J, \epsilon_{t-1}, \epsilon_{t-2}, \ldots\}$.

Under these assumptions, $\rho(L)u_t = \epsilon_t$ is an MDS relative to the information set J_{t-1}^*. Therefore, finding the optimal set of instruments amounts to finding the optimal $K \times 1$ vector $a_t \in J_t^*$ satisfying

$$E[a_t \epsilon_{t+1}] = E[a_t \rho(L) u_{t+1}] = 0. \tag{3.109}$$

If we use the same logic as before and assume homoskedasticity, the optimal instrument vector is given by[11]

$$a_t^* = E\left[\frac{\partial \epsilon_{t+1}}{\partial \theta} \;\middle|\; J_t^*\right] = -\rho(L)x_t. \tag{3.110}$$

An identical set of moment equations is obtained by first transforming the linear projection equation by $\rho(L)$,

$$\rho(L)y_t = \rho(L)x_t'\delta_0 + \epsilon_t, \tag{3.111}$$

and then examining the first-order conditions for least-squares estimation of this transformed equation. Not surprisingly, the optimal GMM estimator is the generalized least-squares estimator in the presence of autoregressive autocorrelation.

This result can be derived directly using Lemma 3.3. According to this lemma, the optimal weight vector A^* for weighting u_t (not ϵ_t) satisfies

$$\begin{aligned} E\big[A_t x_t'\big] &= \sum_{j=-\infty}^{\infty} E\big[A_t u_t u_{t-j}\big(A_{t-j}^*\big)'\big] \\ &= E\big[\rho(L)^{-1} A_t \rho(L)^{-1}\big(A_t^*\big)'\big] \\ &= E\big[A_t \rho\big(L^{-1}\big)^{-1}\rho(L)^{-1}\big(A_t^*\big)'\big]. \end{aligned} \tag{3.112}$$

The second equality is obtained, under our assumption of homoskedasticity, by normalizing the variance of ϵ_t to be 1 (absorbing the variance into the definition of ρ) and using stationarity of the series. The last equality is also

[11] With constant conditional variance, the scaling constant $1/\sigma_\epsilon$ can be ignored in computing the optimal instruments.

an implication of stationarity. It follows immediately that the optimal GMM estimator in this case is characterized by the instrument matrix

$$A_t^* = \rho\big(L^{-1}\big)\rho(L)x_t. \tag{3.113}$$

This result gives rise to exactly the same moment equation as (3.110) because

$$E\big[\rho\big(L^{-1}\big)\rho(L)x_t u_t\big] = E\big[\rho(L)x_t\rho(L)u_t\big] = E\big[\rho(L)x_t\epsilon_t\big], \tag{3.114}$$

owing to the stationarity of the $\{x_t\}$ process.[12]

The weight vector A_t^* is a linear function of current, past, and future x_t. It is orthogonal to u_t because u_t is mean-independent of all elements of the information set J. Clearly, in this case, it is not enough to assume that u_t is mean-independent of current or current and past x_t in order for A^* to give a consistent estimator.

An important practical limitation of these results on models with serially correlated errors is that they maintain the "exogeneity" assumption (3.108). A more typical situation arising in the context of asset pricing is that of Case AC$h(n-1)$ where u_{t+n} is mean-dependent of an information set I_t that does not include future values of variables. Moreover, the implied autocorrelation structure of u is that of an MA$(n-1)$ and not the autoregressive structure (3.106).

Hayashi and Sims (1983) and Hansen and Singleton (1990, 1996) provide an analogous optimality result that applies to this more relevant situation. To illustrate their results, we focus on the case of a scalar u_t [see Hansen and Singleton (1996) for a treatment of the vector case] and let

$$u_t = \alpha(L)\epsilon_t, \quad \alpha(L) = 1 + \alpha_1 L + \ldots + \alpha_{n-1}L^{n-1}, \tag{3.115}$$

and proceed assuming the DAPM implies that $E[u_{t+n}|I_t] = 0$ and that $E[\epsilon_{t+1}^2|I_t] = \sigma_\epsilon^2$, a constant (conditional homoskedasticity). The reason that the A^* given by (3.113) is not optimal for this setting is that u_{t+n} is orthogonal to x_s for $s \le t$, but not for $s > t$. Therefore, using A^* in (3.113) would lead to an inconsistent estimator.

Analogously to the classical treatment of serial correlation in GLS estimation, we would like to "filter" the error u_{t+n} to remove its autocorrelation, and then apply the known optimal A^* for this filtered model. The

[12] Suppose, e.g., that $\rho(L) = 1 + \rho_1 L$ and define $y_t \equiv \rho(L)x_t$. Then (3.114) states that $E[(y_t+\rho_1 y_{t+1})u_t] = E[y_t(u_t+\rho_1 u_{t-1})]$. This is an immediate implication of the joint stationarity of y_t and u_t. A very similar equality plays a central role in generating simplified moment equations and inference in Chapter 9.

added complication in this setting is that the filtering must be done in a manner that preserves the mean-independence of the filtered error from I_t. We accomplish this using the "forward filter" $\lambda(L^{-1}) \equiv \alpha(L^{-1})^{-1}$, since $\lambda(L^{-1})u_{t+n}$ is serially uncorrelated and

$$E\left[\lambda\left(L^{-1}\right)u_{t+n}\middle|I_t\right] = 0. \tag{3.116}$$

Pursuing the example of linear projection of y_{t+n} onto x_t, under homoskedasticity, the optimal instruments for estimation based on the conditional moment restriction (3.116) are given by $E[\lambda(L^{-1})x_t]$. As shown in Hansen and Singleton (1990), this is equivalent to using the instruments

$$A_t^* = \lambda(L)E\left[\lambda\left(L^{-1}\right)x_t\middle|I_t\right] \tag{3.117}$$

for the original projection error u_{t+n}.

3.6.3. ML Estimators

Maximum likelihood estimation does not, in general, fit into any of the optimality results obtained so far because we have taken h to be a given $M \times 1$ vector. ML estimation can be viewed as the solution to the problem of finding the optimal h function to use in estimation. To see this, we start by observing that if we are free to choose the function h, then we might as well set $M = K$ as this is the number of moment equations needed in estimation. Second, since the moment restriction $E[h(z_t, \theta_0)|I_{t-1}] = 0$ implies the moment restriction $E[h(z_{t+n}, \theta_0)|I_{t-1}] = 0$, we presume that our asset pricing theory implies the stronger condition $E[h(z_t, \theta_0)|I_{t-1}] = 0$. That is, if we are free to choose an $h(z_{t+n}; \theta_0)$ that is mean-independent of I_{t-1}, for any $n \geq 0$, then our search can be restricted to the case of $n = 1$. As we proved in Section 3.5.1, one candidate h satisfying this conditional moment restriction is $h^*(z_t; \beta_0) \equiv \partial \log f(y_t|\vec{y}_{1t-1}^{\,J}; \beta_0)/\partial\beta$. This choice is indeed optimal, as the following argument shows.

Since we have K equations in the K unknowns $\theta(= \beta)$ and our focus is on the selection of h, we set $A_t = I_K$, for all t. Given an arbitrary h satisfying (3.94), the GMM estimator based on the fact that the unconditional mean of h is zero satisfies

$$\frac{1}{T}\sum_{t=1}^{T} h(z_t; \theta_T) = 0. \tag{3.118}$$

Applying Theorem 3.5 and Theorem 3.4 [to establish Condition (v)], we get

$$\sqrt{T}(\theta_T - \theta_0) \Rightarrow N\left(0,\, d_0^{-1}\Sigma_0\left(d_0'\right)^{-1}\right), \tag{3.119}$$

where $\Sigma_0 = E[h(z_t;\theta_0)h(z_t;\theta_0)']$. Next, define $D \equiv d_0^{-1}h - d_0^{*-1}h^*$, where we have suppressed the arguments to conserve on notation and used "$*$" to indicate the terms associated with ML estimation. Consider the expectation

$$E\big[Dh^{*\prime}\big] = d_0^{-1}E\big[hh^{*\prime}\big] + d_0^{*-1}d_0^*. \tag{3.120}$$

The counterpart for h to the optimality condition (3.86) is $E[hh^{*\prime}] = -d_0$, which can be verified by direct calculation.[13] Thus, D is orthogonal to $h^{*\prime}$. It follows that, taking the expected value of $D + d_0^{*-1}h^*$ times its transpose,

$$d_0^{-1}\Sigma_0 d_0^{-1} = E\big[D(z_t)D(z_t)'\big] - d_0^{*-1}, \tag{3.121}$$

where the last matrix is the Cramer-Rao lower bound achieved by the ML estimator. Since $E[DD']$ is positive-semidefinite, we conclude that the optimal choice of h is the score of the log-likelihood, h^*.

[13] Using the fact that the mean of $h(z_t,\theta_0)$ is zero, we have

$$0 = \int \frac{\partial h}{\partial \theta} f\, dy + \int h \frac{\partial f}{\partial \theta}\, dy,$$

which implies that $-d_0 = E[hh^{*\prime}]$.