

## AN INFORMATION-THEORETIC ALTERNATIVE TO GENERALIZED METHOD OF MOMENTS ESTIMATION

BY YUICHI KITAMURA AND MICHAEL STUTZER<sup>1</sup>

While optimally weighted GMM estimation has desirable large sample properties, its small sample performance is poor in some applications. We propose a computationally simple alternative, for weakly dependent data generating mechanisms, based on minimization of the Kullback-Leibler Information Criterion. Conditions are derived under which the large sample properties of this estimator are similar to GMM, i.e., the estimator will be consistent and asymptotically normal, with the same asymptotic covariance matrix as GMM. In addition, we propose overidentifying and parametric restrictions tests as alternatives to analogous GMM procedures.

KEYWORDS: GMM, estimation, Kullback, entropy, information theory.

### 1. INTRODUCTION

FOLLOWING HANSEN (1982), the following model is used throughout the paper. There is a stochastic vector process  $\mathbf{x}_t$ ,  $t = 1, 2, \dots$ , a parameter vector  $\beta$  from a set  $\Theta$  of possible parameter vectors, and an  $r$ -component vector of observable, real-valued functions  $f(\mathbf{x}, \beta) = (f_1, \dots, f_r)'$ , where  $'$  denotes the transpose operation.<sup>2</sup> We denote the observed time series (i.e. the sample) of these functions by  $f(x_1, \beta), \dots, f(x_T, \beta)$ . Also following Hansen (1982, p. 1032), theory is represented by the prediction:

$$(1) \quad E_\mu[f(\mathbf{x}, \beta^*)] \equiv \int f(\mathbf{x}, \beta^*) d\mu(\mathbf{x}) = \mathbf{0}$$

where  $\beta^*$  is a parameter vector from  $\Theta$ ,  $E_\mu$  is the expectation with respect to the probability measure  $\mu$ , and  $\mathbf{0}$  denotes an  $r$ -component vector of zeroes.

Empirical content is given to (1) by assuming that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T f(x_t, \beta^*) \equiv \mathbf{0}$$

for most realizations  $\mathbf{x}_t = x_t$ ,  $t = 1, 2, \dots$ , of the process. Hansen's GMM estimator of  $\beta^*$  satisfying (1) is then achieved by finding  $\hat{\beta}$  which makes the observed vector of sample means  $\bar{f}_T(\beta) \equiv (1/T) \sum_{t=1}^T f(x_t, \beta)$  close to  $\mathbf{0}$ . More precisely,

$$(2) \quad \hat{\beta} = \arg \min_{\beta \in \Theta} \bar{f}_T(\beta)' W_T \bar{f}_T(\beta)$$

<sup>1</sup> We gratefully acknowledge the early encouragement of Halbert White and conversations with Chuck Whiteman, Narayana Kocherlakota, Mike Keane, John Geweke, Marty Eichenbaum, and Larry Christiano. Lewis Segal and Re-jin Guo conducted simulations that corroborated our interest in developing this paper's results.

<sup>2</sup> Throughout this paper, vectors are columns unless transposed.

where  $W$  is a symmetric, positive definite weighting matrix used to measure the closeness of the sample moment vector to  $\mathbf{0}$ , via closeness of the quadratic form (2) to zero.

Hansen and Singleton (1982) showed how to estimate a weighting matrix  $\hat{W}_T$  which makes (2) a consistent estimator  $\beta^*$  (when  $\beta^*$  exists) with minimum asymptotic covariance matrix among the class of estimators defined by (2), as well as among the somewhat larger class of minimum discrepancy estimators analyzed by Bates and White (1993, pp. 644–645). Use of this optimal weighting matrix in (2) is thus sometimes termed an optimal minimum distance (OMD) estimator. Hansen also derived a  $\chi^2$ -test for the existence of  $\beta^*$ , i.e. a test of the moment conditions (1).

However, there are good reasons to explore alternatives to this OMD estimator. For example, a simulation study by Altonji and Segal (1994) demonstrated that OMD is biased in small samples when used to estimate covariance models. In their words (p. 3) “The bias arises because sampling errors in the second moments are correlated with sampling errors in the estimate of the covariance matrix of the sample moments. The latter is the weighting matrix for OMD.” Under these circumstances, it seems reasonable to pursue alternatives which do not require the data-dependent tuning which causes this problem. Confirming this intuition, Altonji and Segal found that the use of (2) with an identity matrix for  $W$  (i.e. equal weighting of the moments) outperformed OMD and some other alternative estimators, despite the fact that OMD would be asymptotically more efficient than equal weighting.

Section 2 of this paper presents an alternative estimator which has the same data requirements and computational feasibility as the OMD estimator. Section 3 shows that this alternative is asymptotically as efficient as OMD, unlike the equally weighted GMM estimator. Also in Section 3, we use our estimator to construct a  $\chi^2$ -specification test of the moment conditions (1), as well as Wald, Lagrange Multiplier, and Likelihood Ratio-like tests of parametric restrictions, analogous to those commonly used in applications of OMD.

## 2. AN INFORMATION-THEORETIC ALTERNATIVE

In order to stay as close as possible to Hansen’s formulation of the problem, the alternative described in this paper also utilizes just the sample values of the observable vector  $f$  and the restriction (1). For each parameter vector  $\beta \in \Theta$ , define the following set of probability measures:

$$(3) \quad \mathcal{P}(\beta) \equiv \{P : E_P[f(\mathbf{x}, \beta)] = \mathbf{0}\}$$

which additionally are absolutely continuous with respect to the measure  $\mu$  in (1). Selection of a particular probability measure in  $\mathcal{P}(\beta)$  is a form of *linear inverse problem* (Jones (1989)). Consider the following convex optimization problem, which is an important nonlinear projection problem frequently used to

solve linear inverse problems:<sup>3</sup>

$$(4) \quad \min_{P \in \mathcal{P}(\beta)} D(P \| \mu) \equiv \min_P \int \log(dP/d\mu) dP$$

subject to  $E_P[f(\mathbf{x}, \beta)] = \mathbf{0}$ .

In (4),  $D(P \| \mu) = \int \log(dP/d\mu) dP$  is the Kullback-Leibler Information Criterion (KLIC) distance from  $P$  to  $\mu$  (White (1982)). It is well-known that  $D(P \| \mu) \geq 0$ , with equality if and only if  $P = \mu$  (a.e.). Thus, the existence of a  $\beta^*$  satisfying (1) implies that  $\mu \in \mathcal{P}(\beta^*)$ , in which case  $\mu$  solves (4) and realizes a minimized KLIC value of  $D(\mu \| \mu) = 0$ . The model fails when  $\mu$  is not an element of  $\mathcal{P}(\beta)$  for each  $\beta \in \Theta$ . In this case, for each  $\beta$ , there is a positive KLIC distance  $D(P(\beta) \| \mu) > 0$  attained by the solution  $P(\beta)$  to (4). The solution has a density  $dP(\beta)/d\mu \neq 1$ . Thus, as an alternative to GMM, it seems reasonable to search for a  $\beta$  making  $D(P(\beta) \| \mu)$  as close to zero as possible, in sample. In addition, we will show that the deviation from zero can be used in a specification test of the moment conditions (1), based on the size of the minimized value of (4).

## 2.1 Discussion

GMM estimators, like others in the class of discrepancy estimators defined by Bates and White (1993), focus attention on the *inability* of parameter vectors  $\beta \neq \beta^*$  to satisfy the moment conditions (1) defined by the fixed measure  $\mu$ . They thus try to get sample moments close to the (zero) vector of population moments fixed by  $\mu$ . Our “duality” approach focuses attention on a change of measure,  $dP(\beta)/d\mu$ , which enables  $\beta \neq \beta^*$  to satisfy the *transformed* moment conditions (3). The correct parameter value  $\beta^*$  is the only one also able to satisfy (1), in which case  $P(\beta^*)$  must equal  $\mu$  (a.e.). It thus uses closeness between measures, rather than closeness of moments, to find  $\beta^*$ . As noted by Robinson (1991), the KLIC is extremely sensitive to any deviations of one measure from another. It thus is very sensitive to even small deviations of the transformed measure  $P(\beta)$  from  $\mu$ , induced by small deviations of  $\beta$  from  $\beta^*$ . So it is not surprising that this procedure leads to an estimator, defined in the following section, which has good asymptotic properties.<sup>4</sup>

This paper arose out of earlier work by one of the authors. Subsequently, independent work by Imbens (1993) came to our attention, which describes a similar use of the KLIC. However, his approach did not use duality theory to drastically simplify the implementation, as described in the following section. In addition, Imbens did not show how to modify his procedure to treat the important case of serially dependent data generating mechanisms, as we do, nor

<sup>3</sup> See Csiszar (1975) for key results about this projection problem, and Stutzer (1995) for a related application.

<sup>4</sup> In fact, it is possible to derive GMM as a quadratic approximation of our estimator (derivation available upon request).

did he derive the parametric hypothesis testing framework that we developed in Section 3.1. An early contribution of Haberman (1984) used the KLIC criterion to select a probability measure satisfying a vector of moment conditions, but did not consider the parameter estimation problem at all. Qin and Lawless (1994) estimate parameters by maximizing the “empirical likelihood” subject to the moment conditions. It is not hard to show that after substituting  $D(\mu \| P(\beta))$  for  $D(P(\beta) \| \mu)$  in our problem, the empirical likelihood method results. This estimator is also asymptotically normal, and can be used to construct tests of the moment conditions and other parametric hypotheses. The work of Imbens et al. (1995) is similar to this paper and to ours. None of these authors considered serially dependent data generating mechanisms, as we do.

## 2.2 The Estimator

This estimation concept is quite easy to implement. The solution to (4) is well-known (see, e.g., Csiszar (1975, Sec. 3(A))) to have the following *Gibbs canonical* density:

$$(5) \quad \frac{dP(\beta)}{d\mu} = \frac{e^{\gamma(\beta) f(x, \beta)}}{E_{\mu}[e^{\gamma(\beta) f(x, \beta)}]}.$$

To compute the coefficient vector  $\gamma(\beta)$  in (5), define the function  $\mathcal{M}(\beta, \gamma) \equiv E_{\mu}[e^{\gamma f(x, \beta)}]$ , and solve the following unconstrained, convex problem:<sup>5</sup>

$$(6) \quad \gamma(\beta) = \arg \min_{\gamma} \mathcal{M}(\beta, \gamma)$$

which can also be used to obtain the following formula for the minimized value of KLIC in (4):

$$(7) \quad D(P(\beta) \| \mu) = -\log \mathcal{M}(\beta, \gamma(\beta)).$$

Estimation of the true parameter vector  $\beta^*$  is numerically simple. To see this, first note that (6) and (7) show that

$$\beta^* = \arg \min_{\beta} -\log \mathcal{M}(\beta, \gamma(\beta)) = \arg \max_{\beta} \mathcal{M}(\beta, \gamma(\beta)).$$

We thus must estimate a saddle point of the function  $\mathcal{M}(\beta, \gamma) \equiv E_{\mu}[e^{\gamma f(x, \beta)}]$ .

To estimate this saddle point, one might just substitute the sample time average for the expectation under the unknown measure  $\mu$ . We will show that while this estimator is consistent, it is asymptotically inefficient relative to an alternative estimator when there is a dependent data generating mechanism.

<sup>5</sup> For a proof, see Ben-Tal (1985, Sec. 3), noting that his dual variable  $y$  is our  $\gamma(\beta)$ , that his right-hand side constants  $a$  are zero in our problem, that his problem (H) may be transformed to our problem (6) by taking the supremum over the exponential of his problem, and that our equality constraints leave our dual variable  $\gamma$  unconstrained. If the moment constraints were inequalities, Ben-Tal's proof shows that our approach would still work with  $\gamma$  constrained to be nonnegative.

The alternative estimator smooths the observations first. More precisely, we replace the observation  $f(x_t, \beta)$  by

$$(8) \quad \hat{f}(t, \beta) \equiv \sum_{k=-K}^K \frac{1}{2K+1} f(x_{t-k}, \beta)$$

where  $K^2/T \rightarrow 0$ , and  $K \rightarrow \infty$  as  $T \rightarrow \infty$ ,<sup>6</sup> and compute the estimator

$$(9) \quad (\hat{\beta}_T, \hat{\gamma}_T) = \arg \max_{\beta} \min_{\gamma} \left[ \hat{Q}_T(\beta, \gamma) = \frac{1}{T} \sum_{t=1}^T e^{\gamma' \hat{f}(t, \beta)} \right].$$

A variety of unconstrained numerical minimization algorithms could be used to solve (9). The performance of these algorithms is enhanced by using a good initial guess  $(\beta^0, \gamma^0)$  reasonably close to the solution of (9). For example, one could use the unsmoothed observations (i.e.  $K = 0$  in (8)), set  $\gamma^0 = 0$ , and use the identity matrix  $W$  in (2) to produce a consistent estimate  $\beta^0$ . Let us next consider the computational burden of solving (9) by finding a zero of its gradient through, say, Newton-Raphson iteration. First, the number of components in the solution of (9) equals the number of parameters plus  $r$ , the number of constraints. It may seem that this makes it somewhat harder to solve than the GMM problem (2), where the  $r$  variables  $\gamma$  do not explicitly occur. But the exponential form of (9) yields simple analytical formulae for its first and second derivatives with respect to  $\gamma$ , considerably easing the computational burden. Still, in high dimensional problems, one could compute just one Newton-Raphson step on the gradient of (9) to find an approximate saddle point. Under regularity conditions in Robinson (1988), this two-step estimator will also be consistent and asymptotically normal, with the same asymptotic covariance matrix as Hansen's OMD.<sup>7</sup>

### 3. ASYMPTOTIC RESULTS

To prove the consistency of the estimator  $\hat{\beta}_T$  in (9), we make use of the approach suggested by Wald (1949) and Wolfowitz (1949) for proving the consistency of maximum likelihood estimation. Though we only prove weak consistency, a strong consistency result may be available by utilizing Wald's approach.

In what follows,  $\Gamma(\beta, \delta)$  denotes an open sphere with center  $\beta$  and radius  $\delta$ , and  $O_p$  and  $o_p$  are the stochastic order symbols of Mann and Wald (1943). Let

<sup>6</sup> Though various kernels can be used to weight the observations, we adopted the flat function to simplify the derivation of properties.

<sup>7</sup> This suggestion was made by Peter Robinson. In addition to Assumptions 1-6 made later in this paper, sufficient conditions for application of this result are that the first three derivatives of  $f$  with respect to  $\beta$  be dominated by integrable functions in the neighborhood of  $\beta^*$ . In addition, Guo (1995) utilized the simulation of Ferson and Foerster (1994) to show that this two-step estimator had less bias than a comparable two-step GMM estimator. Their root mean squared errors were almost identical.

$\|U\|$  denote the Euclidean norm  $\sqrt{\text{trace}(U'U)}$  of a column vector or matrix  $U$ .

We make the following eleven assumptions:

ASSUMPTION 1: *The process  $\mathbf{x}_t$  is stationary and ergodic.*

ASSUMPTION 2:  *$\beta \in \Theta$ , a compact,  $m$ -dimensional set.*

ASSUMPTION 3:  *$\exists$  unique  $\beta^* \in \Theta$  satisfying (1).*

ASSUMPTION 4: *For sufficiently small  $\delta > 0$ ,  $E_\mu[\sup_{\beta' \in \Gamma(\beta, \delta)} e^{g'f(\mathbf{x}, \beta')}] < \infty$ , for all vectors  $g$  in a neighborhood of the origin.*

ASSUMPTION 5:  *$E_\mu[f(\mathbf{x}, \beta)f(\mathbf{x}, \beta)']$  is nonsingular for all  $\beta$  in  $\Theta$ .*

ASSUMPTION 6:  *$\beta_j \rightarrow \beta \in \Theta \Rightarrow f_i(x, \beta_j) \rightarrow f_i(x, \beta)$ , for almost every  $x$ .*

ASSUMPTION 7: *The process  $\mathbf{x}_t$  is strongly mixing with mixing coefficients  $\alpha_n$  satisfying  $\exists b > 1: \sum_{n=1}^{\infty} \alpha_n^{1-1/b} < \infty$ .*

ASSUMPTION 8:  *$T \rightarrow \infty \Rightarrow \text{Var}[(1/\sqrt{T})\sum_{t=1}^T f(\mathbf{x}_t, \beta^*)] \rightarrow S > 0$ .*

ASSUMPTION 9:  *$\exists \delta$ ,  $f(x, \beta)$  is continuously differentiable at  $\beta \in \Gamma(\beta^*, \delta)$  for almost every  $x$ .*

ASSUMPTION 10:  *$\exists \delta, \epsilon' > 0: \forall \epsilon \in (0, \epsilon'), D_\mu[\sup_{\beta \in \Gamma(\beta^*, \delta)} \|\partial f(\mathbf{x}, \beta)/\partial \beta'\|^{2+\epsilon}] < \infty$ .*

ASSUMPTION 11: *The  $r \times m$  matrix  $D = E_\mu[\partial f(\mathbf{x}, \beta^*)/\partial \beta']$  has rank  $m$ .*

Assumptions 1–3 are standard regularity conditions employed in GMM estimation, although the compactness in Assumption 2 can be relaxed along the lines of Wald's (1949) proof. In our proofs, Assumption 4 plays the same role as Wald's Assumption 2 does in his proof. While Assumption 4 is stronger than the moment existence assumption in Hansen (1982), it is commonly assumed in the KLIC literature (e.g. in Csiszar (1975, Theorem 3.3) or Haberman (1984, Theorem 5) and in exponential models (Berk (1972))). The nonsingularity Assumption 5 may be relaxed, as in Sheehy (1988), at the expense of greater complexity. The continuity Assumption 6 will be satisfied in most applications. Assumption 7–11 will later be shown to be sufficient for asymptotic normality.

Under Assumptions 1–6 above, we have the following consistency result:

THEOREM 1: *Under Assumptions 1–6,  $\hat{\beta}_T$  in (9) converges to  $\beta^*$  in probability.*

PROOF: See the Appendix.

The additional Assumptions 7–11 are sufficient to show that (9) is, to first order, asymptotically equivalent to OMD. These additional assumptions are frequently assumed in the nonlinear econometrics literature, e.g. Andrews (1993). These conditions are applicable in many situations, but are not the weakest ones ensuring the following theorem:

**THEOREM 2:** *Under the additional Assumptions 7–11 above, to first order, (9) is asymptotically equivalent to OMD. Thus,*

$$\sqrt{T}(\hat{\beta}_T - \beta^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, V)$$

where  $V = (D'S^{-1}D)^{-1}$ ,  $D = E_\mu[\partial f(\mathbf{x}, \beta^*)/\partial \beta']$ , and  $S = \sum_{j=-\infty}^{\infty} E_\mu[f(\mathbf{x}_t, \beta^*)f(\mathbf{x}_{t-j}, \beta^*)]$ .

**PROOF:** See the Appendix.

The following corollary considers the consequences of using unsmoothed observations (i.e.  $K = 0$  in (8)).

**COROLLARY 1:** *With  $K = 0$ ,*

$$\sqrt{T}(\hat{\beta}_T - \beta^*) \xrightarrow{d} \mathcal{N}[\mathbf{0}, V]$$

where  $V = (D'M^{-1}D)^{-1}D'M^{-1}SM^{-1}D(D'M^{-1}D)^{-1}$ ,  $M = E[f(\mathbf{x}, \beta^*)f(\mathbf{x}, \beta^*)']$ , and  $S$  is defined by Assumption 8. When  $\mathbf{x}_t$  is iid, there is asymptotic equivalence to OMD with  $V = (D'M^{-1}D)^{-1}$ .

**PROOF:** See the Appendix.

Corollary 1 shows that, to first order, use of unsmoothed observations results in an estimator asymptotically equivalent to GMM with the weighting matrix  $\hat{W}_T = (\sum_t f(x_t, \beta^*)f(x_t, \beta^*)'/T)^{-1}$ . By replacing  $f(x_t, \beta)$  with (8), we are essentially using the weighting matrix  $(\sum_{k=-2K}^{2K} \omega_K(k)\Gamma_k)^{-1}$ , where  $\omega_K(\cdot)$  is the Bartlett kernel<sup>8</sup> and  $\Gamma_k$  is the  $k$ th order sample autocovariance matrix of  $f(\mathbf{x}_t, \beta^*)$ . But this is an estimator of the optimal weighting matrix for OMD.

### 3.1 Hypothesis Testing

When there are more constraints in (1) than there are parameters (i.e.  $r > m$ ), the size of the minimized (positive) quadratic form in (2), when multiplied by the sample size, is asymptotically  $\chi^2$ . An excessively large value of this “ $J$ -statistic” may thus be used to test the hypothesis (1).  $\hat{Q}_T$  in (9) is unity when  $\gamma = 0$ , and is

<sup>8</sup> Because the optimal bandwidth growth rate for the Bartlett kernel is  $T^{1/3}$  (see Parzen (1957, p. 344) and Grenander and Rosenblatt (1957, p. 154)), there is no problem with our assumption about the growth rate of  $K$ .

always greater than zero. Under Assumptions 1–11, the estimate of the minimized KLIC may also be used in an analogous test of the overidentifying restrictions, stated below.

**THEOREM 3:** *Under the null hypothesis (1), the test statistic*

$$(10) \quad \hat{\kappa}_T \equiv -\frac{2T}{2K+1} \log \hat{Q}_T(\hat{\beta}_T, \hat{\gamma}_T) \xrightarrow{d} \chi_{r-m}^2.$$

*Thus, an excessively large value of (10) permits rejection of the hypothesis (1).*

**PROOF:** See the Appendix.

Suppose we wish to construct tests of the following, possibly nonlinear, restrictions:

$$(11) \quad H_0: R(\beta^*) = a,$$

where  $a$  is an  $s \leq m$  dimensional vector of constants. Now constrain the search for a saddle point (9) by imposing the restrictions (11), and denote the constrained solution by  $\hat{\beta}_T^c$ . Then, the Wald, Lagrange Multiplier, and Likelihood Ratio-like test statistics are stated below.

**THEOREM 4:** *Three alternative test statistics of the parametric restrictions (11) are:*

$$\begin{aligned} \text{Wald}_T &\equiv T(R(\hat{\beta}_T) - a)' [A_T V_T A_T']^{-1} (R(\hat{\beta}_T) - a), \\ \text{LM}_T &\equiv T \bar{f}_T'(\hat{\beta}_T^c) S_T^{-1} D_T V_T D_T' S_T^{-1} \bar{f}_T(\hat{\beta}_T^c), \\ \text{LR}_T &\equiv \frac{2T}{2K+1} \left[ \log \hat{Q}_T(\hat{\beta}_T, \hat{\gamma}(\hat{\beta}_T)) - \log \hat{Q}_T(\hat{\beta}_T^c, \hat{\gamma}(\hat{\beta}_T^c)) \right], \end{aligned}$$

where  $D_T$ ,  $S_T$ , and  $V_T$  are consistent estimates of  $D$ ,  $S$ , and  $V$  (defined in Theorem 2), and  $A_T$  is a consistent estimate of  $A$ , the Jacobian matrix of  $R$  evaluated at  $\beta^*$ , assumed to be of full row rank.

$$(12) \quad \text{Wald}_T, \text{LM}_T, \text{LR}_T, \xrightarrow{d} \chi_s^2.$$

**PROOF:** The results follow from Theorems 2 and 3 and the derivations in Amemiya (1985, Sec. 4.5.1).

#### 4. CONCLUSIONS AND FUTURE RESEARCH

A simple alternative optimally-weighted GMM (a.k.a. OMD) estimation is presented, that handles weakly dependent data generating mechanisms. It is based on minimization of the Kullback-Leibler Information Criterion.

Under reasonable regularity assumptions, the large sample properties of this estimator are similar to OMD. In particular, we identified assumptions under which the estimator is consistent and asymptotically normal. In addition, its asymptotic covariance matrix is the same as OMD.

As is the case with any new estimator, much work remains. A good large sample topic will be to study the higher order efficiency of our estimator relative to OMD. Good small sample topics are also available. Extensive small sample simulation studies should be conducted to better determine the situations under which our estimator will outperform OMD estimates and hypothesis tests. For example, it would be interesting to compare the small sample properties of Theorem 4's tests (12) against their GMM counterparts.

*Department of Economics, University of Minnesota, 271 19th Ave. S., Minneapolis, MN 55455, U.S.A.,*

*and*

*Department of Finance, Carlson School of Management, University of Minnesota, 271 19th Ave. S., Minneapolis, MN 55455, U.S.A.*

*Manuscript received May, 1995; final revision received August, 1996.*

APPENDIX

PROOF OF THEOREM 1

Assumption 3 implies that there is a unique saddle point  $(\beta^*, \gamma(\beta^*))$  of the function  $\mathcal{M}$  in (6). Because  $P(\beta^*) = \mu$ ,  $\gamma(\beta^*) = \mathbf{0}$  in (5), and the value of the saddle function  $\mathcal{M}(\beta^*, \gamma(\beta^*)) = 1$ . Assumption 3 also implies that  $\mathcal{P}(\beta)$  in (3) does not contain  $\mu$  when  $\beta \neq \beta^*$ , so we have

$$(13) \quad \mathcal{M}(\beta, \gamma(\beta)) < \mathcal{M}(\beta^*, \gamma(\beta^*)) = 1.$$

Note that  $\gamma(\beta)$  is continuous in its argument. This can be verified by noting that the Jacobian  $E_\mu[f(x, \beta)f(x, \beta)^\top e^{\gamma(\beta)^\top f(x, \beta)}]$  of the first order condition  $E_\mu[f(x, \beta)e^{\gamma(\beta)^\top f(x, \beta)}] = 0$  is nonsingular under Assumption 5. From Assumption 6,  $\lim_{\delta \downarrow 0} \sup_{\beta' \in \Gamma(\beta, \delta)} e^{\gamma(\beta')^\top f(x, \beta')} = e^{\gamma(\beta)^\top f(x, \beta)}$  ( $\mu$ -a.e.). By Assumption 4 and the continuity of  $\gamma(\beta)$ , there exists an  $r$ -dimensional vector  $\bar{g} = (\bar{g}_1, \dots, \bar{g}_r)$  such that  $\gamma_i(\beta'), \beta' \in \Gamma(\beta, \delta)$  is on the line segment joining 0 and  $\bar{g}_i$  and

$$E_\mu \left[ \sup_{\beta' \in \Gamma(\beta, \delta)} e^{\max[\gamma(\beta')^\top f(x, \beta'), 0]} \right] \leq E_\mu \left[ \sup_{\beta' \in \Gamma(\beta, \delta)} e^{\max[\bar{g}' f(x, \beta'), 0]} \right] < \infty$$

for a sufficiently small  $\delta$ . Therefore, for such  $\delta$ ,

$$E_\mu \left[ \sup_{\beta' \in \Gamma(\beta, \delta)} e^{\gamma(\beta')^\top f(x, \beta')} \right] < \infty.$$

Thus, the Dominated Convergence Theorem ensures that

$$(14) \quad \lim_{\delta \downarrow 0} E_\mu \left[ \sup_{\beta' \in \Gamma(\beta, \delta)} e^{\gamma(\beta')^\top f(x, \beta')} \right] = \mathcal{M}(\beta, \gamma(\beta)).$$

By the compactness Assumption 2, one can cover a compact set of parameters  $\Theta - \Gamma(\beta^*, \delta)$  with a suitably large number  $H$  of spheres  $\Gamma(\beta_j, \delta_j)$ , taking each  $\delta_j$  small enough so that (13) and (14)

can be utilized to ensure

$$E_{\mu} \left[ \sup_{\beta' \in \Gamma(\beta_j, \delta_j)} e^{\gamma(\beta')'f(x_t, \beta')} \right] < \mathcal{H}(\beta^*, \gamma(\beta^*)) = 1.$$

We can thus find positive numbers  $h_j$  so that

$$E_{\mu} \left[ \sup_{\beta' \in \Gamma(\beta_j, \delta_j)} e^{\gamma(\beta')'f(x_t, \beta')} \right] = 1 - 2h_j \quad (j = 1, \dots, H).$$

Now by Assumptions 1 and 4, there exists a sufficiently large integer  $T_j$  so that for small positive  $\epsilon$ :

$$\text{Prob} \left[ \frac{1}{T} \sum_{t=1}^T \sup_{\beta' \in \Gamma(\beta_j, \delta_j)} e^{\gamma(\beta')'f(x_t, \beta')} > 1 - h_j \right] < \epsilon/2H \quad (j = 1, \dots, H)$$

for all  $T > T_j$ . These inequalities imply

$$(15) \quad \text{Prob} \left[ \sup_{\beta' \in \Theta - \Gamma(\beta^*, \delta)} \frac{1}{T} \sum_{t=1}^T e^{\gamma(\beta')'f(x_t, \beta')} > 1 - h \right] < \epsilon/2$$

for all  $T > \max_j T_j$ , where  $h = \min_j h_j$ .

Note that for any  $\beta \in \Theta$ , the minimizer  $\hat{\gamma}_T(\beta)$  satisfies

$$(16) \quad \frac{1}{T} \sum_{t=1}^T e^{\hat{\gamma}_T(\beta)'f(t, \beta)} \leq \frac{1}{T} \sum_{t=1}^T e^{\gamma(\beta)' \sum_k f(x_{t-k}, \beta)/(2K+1)}$$

and that convexity implies that the right-hand side is no greater than

$$(17) \quad \frac{1}{T} \sum_{t=1}^T \sum_k e^{\gamma(\beta)'f(x_{t-k}, \beta)/(2K+1)} + o_p(1) = \frac{1}{T} \sum_{t=1}^T e^{\gamma(\beta)'f(x_t, \beta)} + o_p(1)$$

where the term  $o_p(1)$  adjusts for the "endpoint effect."

By (15)

$$(18) \quad \text{Prob} \left[ \sup_{\beta' \in \Theta - \Gamma(\beta^*, \delta)} \frac{1}{T} \sum_t e^{\hat{\gamma}_T(\beta')'f(t, \beta')} > 1 - h \right] < \epsilon/2 \quad \text{for large } T.$$

Note that

$$(19) \quad e^{\hat{\gamma}_T(\beta^*)' \sum_t \hat{f}(t, \beta^*)/T} + o_p(1) \leq \frac{1}{T} \sum_t e^{\hat{\gamma}_T(\beta^*)' \hat{f}(t, \beta^*)} \leq 1.$$

By Assumptions 1 and 3,  $\sum_t \hat{f}_i(t, \beta^*)/T \xrightarrow{P} 0$ , and one sees that  $\hat{\gamma}_{T_i}(\beta^*) \xrightarrow{P} 0, i = 1, \dots, r$  because

$$\begin{aligned} & (2K+1)(\hat{Q}_T(\beta^*, \gamma) - \hat{Q}_T(\beta^*, 0)) \\ &= \frac{2K+1}{T} \gamma' \sum_t \hat{f}(t, \beta^*) + \gamma' \left( \frac{2K+1}{2T} \sum_t \hat{f}(t, \beta^*) \hat{f}(t, \beta^*) \right) \gamma + o_p(1); \end{aligned}$$

thus

$$\hat{\gamma}_T(\beta^*) = - \frac{2K+1}{T} S^{-1} \sum_t \hat{f}(t, \beta^*) + o_p(1) = o_p(1).$$

This implies that  $e^{\hat{\gamma}_T(\beta^*)' \sum_t \hat{f}(t, \beta^*)/T} \xrightarrow{P} 1$ . Thus, by (19) there exists a sufficiently large  $T'$  such that

$$(20) \quad \text{Prob} \left[ \sum_t e^{\hat{\gamma}_T(\beta^*)' \hat{f}(t, \beta^*)/T} < 1 - h/2 \right] < \epsilon/2, \quad T > T'.$$

Probabilities (18) and (20) imply the consistency of  $\hat{\beta}_T$ .

For use in what follows, we also show that  $\hat{\gamma}_T \equiv \gamma_T(\hat{\beta}_T) \xrightarrow{p} 0$ . To show this, it suffices to prove that  $\sum_t \hat{f}(t, \hat{\beta}_T)$  is  $O_p(\sqrt{T})$ . The rest of the proof is identical to the proof of the consistency of  $\hat{\gamma}_T(\beta^*)$ .

Let  $g_T = ((2K + 1)/\sqrt{T})g$ , where  $g$  is an arbitrary  $r$ -dimensional vector. Note that

$$\begin{aligned} -\frac{2T}{2K+1} \log \hat{Q}_T(\hat{\beta}_T, g_T) &\leq -\frac{2T}{2K+1} \log \hat{Q}_T(\hat{\beta}_T, \gamma_T(\hat{\beta}_T)) \\ &\leq -\frac{2T}{2K+1} \log \hat{Q}_T(\beta^*, \gamma_T(\beta^*)). \end{aligned}$$

The last expression is  $T^{-1}[\sum_t f(x_t, \beta^*)]S^{-1}[\sum_t f(x_t, \beta^*)] + o_p(1)$ , and so it is  $O_p(1)$ , and asymptotically  $\chi_r^2$  (see the proof of Theorem 3). Also,

$$\begin{aligned} -\frac{2T}{2K+1} \log \hat{Q}_T(\hat{\beta}_T, g_T) &= -\frac{2T}{2K+1} g_T' \sum_t \hat{f}(t, \hat{\beta}_T) / T \\ &\quad - \frac{2T}{2(2K+1)} g_T' \sum_t \hat{f}(t, \hat{\beta}_T) \hat{f}(t, \hat{\beta}_T)' / T g_T + o_p(1) \end{aligned}$$

which in turn equals

$$-\frac{2}{2K+1} g_T' \sum_t \hat{f}(t, \hat{\beta}_T) + O_p(1) = -\frac{2}{\sqrt{T}} g' \sum_t \hat{f}(t, \hat{\beta}_T) + O_p(1).$$

Since the last expression is bounded above by the asymptotically  $\chi_r^2$  random variable for any vector  $g$ ,  $\sum_t \hat{f}(t, \hat{\beta}_T) = O_p(\sqrt{T})$ . In addition, considering the "endpoint effect,"  $\sum_t f(x_t, \hat{\beta}_T) = \sum_t \hat{f}(t, \hat{\beta}_T) + O_p(K^2/(2K + 1)) = O_p(\sqrt{T})$ . This finding will be used in deriving the last line of (22) below.

PROOF OF THEOREM 2

We first derive an asymptotic approximation for  $\hat{\gamma}_T$  in (9). To examine the asymptotics of the solution for its first order condition,

$$(21) \quad \sum_t \hat{f}(t, \hat{\beta}_T) e^{\hat{\gamma}_T \hat{f}(t, \hat{\beta}_T)} / T = 0,$$

it is useful to expand  $e^{\hat{\gamma}_T \hat{f}(t, \hat{\beta}_T)}$  in a Taylor series about 0, to obtain the following approximation:

$$\begin{aligned} (22) \quad &\frac{1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) e^{\hat{\gamma}_T \hat{f}(t, \hat{\beta}_T)} \\ &= \frac{1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) + \frac{1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) \hat{f}(t, \hat{\beta}_T)' \hat{\gamma}_T + \frac{1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) \sum_{j=2}^{\infty} \frac{1}{j!} (\hat{\gamma}_T' \hat{f}(t, \hat{\beta}_T))^j \\ &\equiv \frac{1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) + \frac{1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) \hat{f}(t, \hat{\beta}_T)' \hat{\gamma}_T + O_p((2K + 1)^{-3/2} \|\hat{\gamma}_T\|^2). \end{aligned}$$

In deriving the order term in (22), note that  $\hat{f}(t, \hat{\beta}_T) = O_p((2K + 1)^{-1/2})$  follows from the last comment in the proof of Theorem 1. Now let  $\hat{S}_T = ((2K + 1)/T) \sum_t \hat{f}(t, \hat{\beta}_T) \hat{f}(t, \hat{\beta}_T)'$ . Then

$$\frac{2K+1}{T} \sum_t \hat{f}(t, \hat{\beta}_T) + \hat{S}_T \hat{\gamma}_T = O_p((2K + 1)^{-1/2} \|\hat{\gamma}_T\|^2)$$

and we obtain

$$(23) \quad \frac{\sqrt{T}}{2K+1} \hat{\gamma}_T = -\hat{S}_T^{-1} \sum_i \hat{f}(t, \beta^*) / \sqrt{T} - \hat{S}_T^{-1} D \sqrt{T} (\hat{\beta}_T - \beta^*) + o_p(1).$$

Thus,

$$(24) \quad \frac{\sqrt{T}}{2K+1} \|\hat{\gamma}_T\| = O_p(\max(1, \sqrt{T} \|\hat{\beta}_T - \beta^*\|)).$$

We now use this approximation result to analyze the asymptotic behavior of  $\hat{\beta}_T$ . To do so, compute the first order condition for  $\hat{\beta}_T$  in (9), and use (21) to simplify it to obtain

$$(25) \quad \frac{\partial \hat{Q}_T}{\partial \beta} = \frac{1}{T} \sum_i \hat{\gamma}_{Ti} \sum_{t=1}^T \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta} e^{\hat{\gamma}_{Ti} \hat{f}(t, \hat{\beta}_T)} = 0.$$

For each  $t = 1, \dots, T$ , expand the exponential term in Taylor's series about 0 to first order, and note that Taylor's Theorem ensures that there exist vectors  $\gamma_t$  such that

$$(26) \quad \frac{1}{T} \sum_i \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta} e^{\hat{\gamma}_{Ti} \hat{f}(t, \hat{\beta}_T)} = \frac{1}{T} \sum_i \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta} + \frac{1}{T} \sum_i \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta} e^{\gamma_t \hat{f}(t, \hat{\beta}_T)} \hat{\gamma}'_T \hat{f}(t, \hat{\beta}_T).$$

By Hölder's Inequality, the component of the second term of the right-hand side in (26) that corresponds to  $\beta_j$  is no greater than

$$(27) \quad \begin{aligned} & \frac{1}{T} \sum_i \left| \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta_j} \right| \|\hat{f}(t, \hat{\beta}_T)\| e^{\gamma_t \hat{f}(t, \hat{\beta}_T)} \|\hat{\gamma}_T\| \\ & \leq \left[ \frac{1}{T} \sum_i \left| \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta_j} \right|^p \right]^{1/p} \left[ \frac{1}{T} \sum_i \|\hat{f}(t, \hat{\beta}_T)\|^q \right]^{1/q} \left[ \frac{1}{T} \sum_i e^{s \gamma_t \hat{f}(t, \hat{\beta}_T)} \right]^{1/s} \|\hat{\gamma}_T\| \\ & \leq \left[ \frac{1}{T} \sum_i \left| \frac{\partial \hat{f}_i(x_i, \hat{\beta}_T)}{\partial \beta_j} \right|^p \right]^{1/p} \left[ \frac{1}{T} \sum_i \|f(x_i, \hat{\beta}_T)\|^q \right]^{1/q} \left[ \frac{1}{T} \sum_i e^{s \gamma_t f(x_i, \hat{\beta}_T)} \right]^{1/s} \|\hat{\gamma}_T\| \\ & \quad + o_p(1) \end{aligned}$$

where  $1/p + 1/q + 1/s = 1$  and  $p, q, s > 1$ . Take  $2 < p < 2 + \epsilon$  and  $2 < q < 2 + \epsilon$ , so that  $s$  will possibly be a large number. The first two terms of (27) are of  $O_p(1)$  by Assumptions 4 and 10 and the consistency of  $\hat{\beta}_T$ . Assumption 4 also implies that the third term is of  $O_p(1)$ . In summary,

$$(28) \quad \frac{1}{T} \sum_i \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta} e^{\hat{\gamma}_{Ti} \hat{f}(t, \hat{\beta}_T)} = \frac{1}{T} \sum_i \frac{\partial \hat{f}_i(t, \hat{\beta}_T)}{\partial \beta} + O_p(\|\hat{\gamma}_T\|) = D_i + o_p(1)$$

where the second equality is implied by ergodicity and Assumption 10. Thus,

$$(29) \quad D' \hat{\gamma}_T = o_p(\|\hat{\gamma}_T\|).$$

By (23), (24), and (29),

$$(30) \quad D' \hat{S}_T^{-1} \sum_i \hat{f}(t, \beta^*) / \sqrt{T} + D' \hat{S}_T^{-1} D \sqrt{T} (\hat{\beta}_T - \beta^*) + o_p(\max(1, \sqrt{T} \|\hat{\beta}_T - \beta^*\|)) = 0;$$

therefore,

$$\sqrt{T} (\hat{\beta}_T - \beta^*) = -(D'S^{-1}D)^{-1} D'S^{-1} \sum_i \hat{f}(t, \beta^*) / \sqrt{T} + o_p(1).$$

Assumptions 7 and 8 and central limit theory in Ibragimov and Linnik (1971, Theorem 18.5.3) imply the result. Note that (24) shows that kernel smoothing lowers the rate of convergence of  $\hat{\gamma}_T$  to  $\sqrt{T}/(2K+1)$ .

PROOF OF COROLLARY 1

Let  $K=0$  in the proof of Theorems 1 and 2. Then, the same derivation results in the following analog of (30):

$$D'\hat{M}_T^{-1} \sum_t f(t, \beta^*)/\sqrt{T} + D'\hat{M}_T^{-1} D\sqrt{T}(\hat{\beta}_T - \beta^*) + o_p(\max(1, \sqrt{T})\|\hat{\beta}_T - \beta^*\|) = 0,$$

where  $\hat{M}_T = (1/T)\sum_t f(x_t, \beta^*)f(x_t, \beta^*)'$ . The results follow.

PROOF OF THEOREM 3

We again approximate  $\hat{Q}_T(\hat{\beta}_T, \hat{\gamma}_T)$ , this time to second order:

$$\hat{Q}_T(\hat{\beta}_T, \hat{\gamma}_T) = 1 + \frac{1}{T} \sum_t \hat{\gamma}_T' \hat{f}(t, \hat{\beta}_T) + \frac{1}{2T} \sum_t (\hat{\gamma}_T' \hat{f}(t, \hat{\beta}_T))^2 + o_p(\|\hat{\gamma}_T\|^3).$$

Since (23) implies that  $\hat{\gamma}_T = -(2K+1)S^{-1}\sum_t \hat{f}(t, \hat{\beta}_T)/T + o_p(K/\sqrt{T})$ ,

$$\begin{aligned} \hat{Q}_T(\hat{\beta}_T, \hat{\gamma}_T) &= 1 - \frac{2K+1}{T^2} \sum_t f(x_t, \hat{\beta}_T)' S^{-1} \sum_t f(x_t, \hat{\beta}_T) \\ &\quad + \frac{2K+1}{2T^2} \sum_t f(x_t, \hat{\beta}_T)' S^{-1} \left( \sum_t \hat{f}(t, \hat{\beta}_T) \hat{f}(t, \hat{\beta}_T)' \right) S^{-1} \sum_t f(x_t, \hat{\beta}_T) \\ &\quad + o_p(K/T). \end{aligned}$$

So

$$\log \hat{Q}_T(\hat{\beta}_T, \hat{\gamma}_T) = -\frac{2K+1}{2T} \left( \sum_t f(x_t, \hat{\beta}_T)' / \sqrt{T} \right) S^{-1} \sum_t f(x_t, \hat{\beta}_T) / \sqrt{T} + o_p(K/T)$$

and by the conventional theory of GMM-based overidentifying restrictions tests, the desired result follows:

$$\hat{\kappa}_T \xrightarrow{d} \chi_{r-m}^2.$$

REFERENCES

ALTONJI, J. G., AND L. M. SEGAL (1994): "Small Sample Bias in GMM Estimation of Covariance Structures," Northwestern University, Center for Urban Affairs and Policy Research.  
 AMEMIYA, T. (1985): *Advanced Econometrics*. Cambridge: Harvard University Press.  
 ANDREWS, D. W. K. (1993): "Test of Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-856.  
 BATES, C. E., AND H. WHITE (1993): "Determination of Estimators with Minimum Asymptotic Covariance Matrices," *Econometric Theory*, 9, 633-648.  
 BEN-TAL, A. (1985): "The Entropy Penalty Approach to Stochastic Programming," *Mathematics of Operational Research*, 10, 263-279.  
 BERK, R. H. (1972): "Consistency and Asymptotic Normality of MLE's for Exponential Models," *The Annals of Mathematical Statistics*, 43, 193-204.

- CSISZAR, I. (1975): "I-Divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, 3, 146–158.
- FERSON, W., AND S. R. FOERSTER (1994): "Finite Sample Properties of the Generalized Method of Moments in Tests of Conditional Asset Pricing Models," *Journal of Economics*, 34, 29–55.
- GRENANDER, U., AND M. ROSENBLATT (1957): *Statistical Analysis of Stationary Time Series*. New York: John Wiley.
- GUO, R. J. (1995): "Small Sample Properties of the KLIC Estimator Applied to Latent Variable Models," Mimeographed, Carlson School of Management Summer Paper, University of Minnesota.
- HABERMAN, S. J. (1984): "Adjustment by Minimum Discriminant Information," *Annals of Statistics*, 12, 971–988.
- HANSEN, L. P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 50, 1029–1054.
- HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286.
- IBRAGIMOV, I. A., AND Y. V. LINNIK (1971): *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters Noordhoff.
- IMBENS, G. W. (1993): "A New Approach to Generalized Method of Moments Estimation," Harvard Institute of Economic Research Discussion Paper No. 1633, Harvard University.
- IMBENS, G. W., P. JOHNSON, AND R. H. SPADY (1995): "Information Theoretic Approaches to Inference in Moment Condition Models," Mimeographed, Economics Department, Harvard University.
- JONES, L. K. (1989): "Approximation-Theoretic Derivation of Logarithmic Entropy Principles for Inverse Problems and Unique Extension of the Maximum Entropy Method to Incorporate Prior Knowledge," *SIAM Journal of Applied Mathematics*, 49, 650–661.
- MANN, H. B., AND A. WALD (1943): "On Stochastic Limit and Order Relationship," *Annals of Mathematical Statistics*, 14, 217–226.
- PARZEN, E. (1957): "On Consistent Estimates of the Spectrum of Stationary Time Series," *Annals of Mathematical Statistics*, 28, 329–347.
- QIN, J., AND J. LAWLESS (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–325.
- ROBINSON, P. M. (1988): "The Stochastic Difference Between Econometric Statistics," *Econometrica*, 56, 531–548.
- (1991): "Consistent Nonparametric Entropy-Based Testing," *Review of Economic Studies*, 58, 437–453.
- SHEEHY, A. (1988): "Kullback-Leibler Constrained Estimation of Probability Measures," Department of Statistics Technical Report No. 250, Stanford University.
- STUTZER, M. (1995): "A Bayesian Approach to Diagnosis of Asset Pricing Models," *Journal of Econometrics*, 68, 367–397.
- WALD, A. (1949): "Note on the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 20, 595–601.
- WHITE, H. (1982): "Maximum Likelihood of Misspecified Models," *Econometrica*, 50, 1–25.
- WOLFOWITZ, J. (1949): "On Wald's Proof of the Consistency of the Maximum Likelihood Estimate," *Annals of Mathematical Statistics*, 20, 601–603.