Econometrica, Vol. 72, No. 5 (September, 2004), 1601–1615

CONSISTENT ESTIMATION OF MODELS DEFINED BY CONDITIONAL MOMENT RESTRICTIONS

BY MANUEL A. DOMÍNGUEZ AND IGNACIO N. LOBATO[1]

In econometrics, models stated as conditional moment restrictions are typically estimated by means of the generalized method of moments (GMM). The GMM estimation procedure can render inconsistent estimates since the number of arbitrarily chosen instruments is finite. In fact, consistency of the GMM estimators relies on additional assumptions that imply unclear restrictions on the data generating process. This article introduces a new, simple and consistent estimation procedure for these models that is directly based on the definition of the conditional moments. The main feature of our procedure is its simplicity, since its implementation does not require the selection of any user-chosen number, and statistical inference is straightforward since the proposed estimator is asymptotically normal. In addition, we suggest an asymptotically efficient estimator constructed by carrying out one Newton–Raphson step in the direction of the efficient GMM estimator.

KEYWORDS: Generalized method of moments, identification, unconditional moments, marked empirical process, integrated regression function, efficiency bound.

1. INTRODUCTION

IN MANY AREAS OF ECONOMETRICS such as panel data, discrete choice, macroeconomics, and finance, there exist models that are defined in terms of conditional moment restrictions. That is, the models establish that certain parametric functions have zero conditional mean when evaluated at the true parameter value. Note that these conditional restrictions imply that the expectation of the parametric functions evaluated at the true parameter value times any function that depends on the conditioning variables is equal to zero. Therefore, when the conditioning variables have a support with infinite cardinality, the conditional moment restrictions imply an infinite number of unconditional moment restrictions. This fact underlies the generalized method of moments (GMM), which is the method commonly employed to estimate these models. Basically, this method consists of the following two stages. First, choose a finite number of unconditional moment restrictions out of the infinite number implied by the conditional moment restrictions. Second, define the estimator as the parameter value that makes the empirical analogs of the selected unconditional moments closest to 0. In linear models, any subset of linearly independent unconditional moment restrictions identifies globally the parameters of interest as long as the dimension of this subset equals the dimension of the parameter vector. Hence, the GMM procedure provides consistent estimators in linear models. However, in nonlinear models the selected unconditional

[1] We thank a co-editor and two referees for very useful suggestions and comments and M. Arellano, M. Carrasco, S. Donald, Y. Kitamura, H. Koul, W. Newey, and W. Stute for useful conversations. We also thank participants at the 2003 Joint Statistical Meeting and at the 2003 Latin American Meeting of the Econometric Society for interesting comments. Both authors acknowledge financial support from Asociación Mexicana de Cultura and from the Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT) under Project Grant J38276D (Domínguez) and Project Grant 41893-S (Lobato).

moment restrictions may hold for several parameter values even if the conditional restrictions just hold for a single value. This means that the GMM objective function may have several global minima. In these cases the arbitrarily chosen unconditional moment restrictions do not identify globally the parameters of interest, and hence, the GMM estimators are inconsistent. The next two examples illustrate this idea.

EXAMPLE 1: Assume that the random variable Y satisfies $E(Y|X) = X^{\theta_0}$ where $\theta_0 = 4$, and X is a random variable that is symmetric around zero and whose fourth and sixth moments are equal, such as an $\mathrm{N}(0, 1/5)$. Assume that the researcher specifies correctly the model $E(Y|X) = X^{\theta}$ where $\theta \in \Theta = [2, \infty)$, and sets out to estimate θ_0. The model implies that $E[(Y - X^{\theta_0})g(X)] = 0$ for any function g provided that $E|(Y - X^{\theta_0})g(X)| < \infty$. Since there is only one parameter, the researcher needs to select at least one function $g(X)$. Let us assume that she selects the functions (typically called instruments) 1 and X. The problem is that these two instruments do not identify the parameter value $\theta_0 = 4$ since the system of equations $E(Y - X^{\theta}) = E((Y - X^{\theta})X) = 0$ also holds for the value $\theta = 6$, at least. Of course, more arbitrary instruments could be added, but it would be simple to find a particular distribution for X, such that θ_0 and additional values for θ satisfy the new set of orthogonality conditions.

EXAMPLE 2: Assume that the random variable Y satisfies the simple nonlinear model $E(Y|X) = \theta_0^2 X + \theta_0 X^2$. Suppose that $\theta_0 = 5/4$ and that $V(Y|X)$ is constant. Assume that the researcher properly specifies the model and, instead of an arbitrary instrument, she chooses the optimal instrument, given by $W_0 = 2\theta_0 X + X^2$; see Amemiya (1974) and Chamberlain (1987). In this case, the parameter θ_0 is not identified, since the equation $E[(Y - \theta^2 X - \theta X^2)W_0] = 0$ is also satisfied for $\theta = -5/4$ when X follows an $\mathrm{N}(-1, 1)$ random variable. Moreover, W_0 is an unfeasible instrument because θ_0 is unknown. Hence, in practice the researcher just knows the form of the optimal instrument, given by $W = 2\theta X + X^2$. In this case the parameter θ_0 is not identified again, since the equation $E[(Y - \theta^2 X - \theta X^2)W] = 0$ is also satisfied for $\theta = -5/4$ and for $\theta = -3$ when X follows an $\mathrm{N}(1, 1)$ random variable.

These simple examples illustrate that the procedure based on selecting an arbitrary finite number of instruments (even the optimal ones) can lead to inconsistent estimation since it does not guarantee that the parameters of interest are globally identified. Hence, GMM typically introduces the additional assumption that the selected unconditional restrictions identify globally the parameters of interest. As we have seen on the examples, this additional assumption depends on the selected instruments and on the unknown true value of the parameters, and in fact, it restricts the marginal distribution of the conditioning variables. Thus, the introduction of this additional assumption leads to the following paradox: while the distribution of the conditioning variables should be irrelevant for the consistent estimation of conditional models, it turns out that this distribution is crucial for GMM estimators because it guarantees global identification of the parameters of interest.

In this article we propose an alternative estimation procedure where the identification problem does not arise, since the method is directly based on the conditional moment restrictions that define the parameters of interest. Implementing our procedure is very simple since no additional user-chosen objects (such as a smoothing number) are needed. As far as we know, ours is the first estimator proposed in the literature that is consistent and does not require the introduction of additional user-chosen objects. Carrying out statistical inference with our estimator is very simple since its asymptotic distribution is normal. In addition, by carrying out a single Newton–Raphson step in the direction of the efficient GMM estimator, an asymptotically efficient estimator can be constructed.

The paper is organized as follows. Section 2 introduces the framework and our estimator, Section 3 establishes the asymptotic theory, Section 4 considers efficient estimation, Section 5 examines a brief Monte Carlo exercise, and Section 6 concludes. The proofs are contained in the Appendix.

2. NOTATION AND FRAMEWORK

Let Z_t be a time series vector and for all t, let $\{Y_t, X_t\}$ be two subvectors of Z_t (that could have common coordinates). We consider Y_t as a k-dimensional time series vector that may contain endogenous and exogenous variables and a finite number of these variables lagged and X_t as a d-dimensional time series vector that contains the exogenous variables (again, a finite number of these variables lagged can be included). The coordinates of Z_t are related by an econometric model that establishes that the true distribution of the data satisfies the following conditional moment restrictions:

$$E\big(h(Y_t, \theta_0)|X_t\big) = 0, \quad \text{a.s.} \tag{1}$$

for a unique value $\theta_0 \in \Theta$, where $\Theta \subset \mathbb{R}^m$. Equation (1) defines the parameter value of interest θ_0, which is unknown to the econometrician. The function h that maps $\mathbb{R}^k \times \Theta$ into $\mathbb{R}^l$ is supposed to be known. In general, $h(Y_t, \theta_0)$ can be understood as the errors in a multivariate nonlinear dynamic regression model. In this paper for simplicity we will consider the case where $l = 1$.

This model has been repeatedly considered in the econometrics literature and several estimators have been proposed; see among others, Amemiya (1974, 1977), Hansen (1982), Newey (1990, 1993), and Robinson (1987, 1991). However, none of these references address the identification problem commented above. For instance, Newey (1990) considers a similar model (see his equation (2.1) on p. 810) in a more restrictive framework (he considers independent and identically distributed data with homoskedasticity) and focuses on the optimality properties of a selected estimator. Note that he assumes that the parameter vector is globally identified by the selected unconditional moment restrictions; see his assumption 3.3(a) on p. 817.

Recently, Donald, Imbens, and Newey (2003) have addressed the identification problem in a different setting. They consider efficient estimation of conditional moment restrictions models. Their analysis is different from ours. They need to introduce a sequence of approximating functions such as splines or power or Fourier series and the researcher needs to select the number of terms of these series to be considered in the analysis. This number is a smoothing or bandwidth number that compared to the sample size has to verify certain rate restrictions in order to achieve efficient estimation. Although this bandwidth number allows their estimators to be root-n asymptotically normal and efficient, statistical inference with this estimator can be sensitive to the selection of the bandwidth number. Furthermore, their procedure is restricted to the independent and identically distributed setting, and for most of their results, the conditioning variables should have a compact support and their joint density has to be bounded away from zero. In the same spirit as Donald, Imbens, and Newey (2003), Newey and Powell (2003) have provided consistent estimators of semiparametric models defined by conditional moment restrictions. Contrary to this approach, our procedure is very simple, does not require the introduction of an arbitrary user-chosen number to achieve an asymptotically normal distribution, allows for instruments with unbounded support, and can be used for time series data.

Kitamura, Tripathi, and Ahn (2000) have also analyzed the problem of efficient estimation in conditional moment restrictions models. By employing a localized empirical likelihood, they propose an estimator that also achieves the semiparametric efficiency bound without estimating the optimal instrument. Similarly to Donald, Imbens, and Newey (2003), Kitamura, Tripathi, and Ahn (2000) also need to introduce a bandwidth number and restrict to the independent and identically distributed setting.

Another related reference is Carrasco and Florens (2000). They consider optimal GMM estimation for the case where there is a continuum of moment conditions in an independent and identically distributed framework. Our estimator is similar to theirs in spirit. However, our estimator cannot be written in their framework, as we will see below, because our norm in the objective function is random and changes with the sample size, whereas their norm is deterministic and does not change with the sample size. Carrasco and Florens' estimator is efficient, but efficiency is achieved at the cost of introducing a user-chosen smoothing number that permits inversion of the covariance operator. As in the case of Donald, Imbens, and Newey (2003) the sensitivity of the estimator to that number is unknown.

Next, we introduce our estimator. As discussed in the previous section, the typical estimation procedure based on selecting some orthogonality conditions does not guarantee global identification of the parameters of interest. In this paper we propose an alternative estimation procedure that uses the whole information about θ_0 contained in expression (1). From Billingsley (1995, Theorem 16.10iii), note that

$$E\big(h(Y_t,\theta_0)|X_t\big)=0 \quad \text{a.s.} \quad \Longleftrightarrow \quad H(\theta_0,x)=0 \quad \text{for almost all } x\in\mathbb{R}^d, \tag{2}$$

where $H(\theta,x)=E(h(Y_t,\theta)I(X_t\le x))$ is the integrated regression function (Brunk (1970)) and the indicator function $I(X_t\le x)$ equals 1 when each component in X_t is less than or equal to the corresponding component in x, and equals 0 otherwise. In addition, from (1), it follows that $P(E(h(Y_t,\theta)|X_t)=0)<1$ when $\theta\neq\theta_0$, so that $H(\theta,x)\neq 0$ in a nonnull set of the sample space of X_t. Therefore, denoting by P_{X_t} the probability distribution function of the random vector X_t, $\int H(\theta_0,x)^2\,dP_{X_t}(x)=0$ but $\int H(\theta,x)^2\,dP_{X_t}(x)>0\ \forall\,\theta\neq\theta_0$. Hence, we can write

$$\theta_0=\arg\min_{\theta\in\Theta}\int H(\theta,x)^2\,dP_{X_t}(x), \tag{3}$$

and θ_0 is the unique value that satisfies (3). Denote the sample integrated regression function by $H_n(\theta,x)=n^{-1}\sum_{t=1}^n h(Y_t,\theta)I(X_t\le x)$, where n is the sample size. For any g, the sample analog of $\int g^2(x)\,dP_{X_t}(x)$ is $n^{-1}\sum_{\ell=1}^n g^2(X_\ell)$. Then, we propose estimating θ_0 by the sample analog of (3), that is,

$$\widehat{\theta}=\arg\min_{\theta\in\Theta}\frac{1}{n^3}\sum_{\ell=1}^n\left(\sum_{t=1}^n h(Y_t,\theta)I(X_t\le X_\ell)\right)^2.$$

This estimator is a minimum distance estimator; see Ch. 5 in Koul (2002). From a computational point of view, the previous objective function has an additional summation of n terms compared to the standard GMM objective function. However, it does not involve either matrix inversion or nonparametric estimation, which are computationally more demanding procedures.

3. ASYMPTOTIC THEORY

We start by enumerating the assumptions for the consistency of our estimator. Let $|\cdot|$ denote the Euclidean norm in the corresponding Euclidean space, and assume that all the considered functions are Borel measurable.

ASSUMPTION 1: *$h(y,\cdot)$ is continuous in Θ for each y in $\mathbb{R}^k$, $|h(Y_t,\theta)| < k(Y_t)$ with $Ek(Y_t) < \infty$ and $E(h(Y_t,\theta)|X_t) = 0$ a.s. if and only if $\theta = \theta_0$.*

ASSUMPTION 2: *Z_t is ergodic and strictly stationary.*

ASSUMPTION 3: *$\Theta \subset \mathbb{R}^m$ is compact.*

Assumptions 1–3 are standard in the GMM literature. Assumption 1 defines the model and identifies globally θ_0. It also establishes that the function h is smooth in Θ, but this smoothness condition is weaker than the Lipschitz condition in Assumption 3 in Donald, Imbens, and Newey (2003). Notice that the assumptions concerning the existence of a bounding function k and the compactness of Θ can be replaced by other assumptions imposing that for all $\theta \in \Theta$ there exists $\rho_\theta > 0$ such that $E[\sup_{\{\|\theta-\theta'\|<\rho_\theta\}\cap\Theta} |h(Y_t,\theta) - h(Y_t,\theta')|] < \infty$ and that $\underline{\lim}_{|\theta|\to\infty} E|h(Y_t,\theta) - h(Y_t,\theta_0)| > 0$. This first condition is a smoothness assumption that is still weaker than the condition in Donald, Imbens, and Newey (2003), whereas the second condition rules out redescending functions. Opposite to standard GMM, all our assumptions refer to the unconditional or to the conditional distribution of h, and nothing is imposed on the marginal distribution of X_t, except for Assumption 2, which just restricts dependence and heterogeneity of the data. Next, we state the consistency theorem whose proof is in the Appendix.

THEOREM 1: *Under Assumptions* 1–3 $\widehat{\theta} \to_{a.s.} \theta_0$.

In order to obtain asymptotic normality, some additional assumptions are required.

ASSUMPTION 4: *$h(y,\cdot)$ is once continuously differentiable in a neighborhood of θ_0 and satisfies $E[\sup_{\theta\in\aleph_0} |\dot{h}(Y_t,\theta)|] < \infty$ where $\aleph_0$ denotes a neighborhood of θ_0 and $\dot{h}(Y_t,\theta) = \partial h(Y_t,\theta)/\partial\theta$.*

ASSUMPTION 5: *$h(Y_t,\theta_0)$ is a martingale difference sequence with respect to $\{Z_s, s \le t\}$.*

ASSUMPTION 6: $\theta_0 \in \operatorname{int}(\Theta)$.

ASSUMPTION 7: $E[h^4(Y_t,\theta_0)\|X_t\|^{1+\delta}] < \infty$.

ASSUMPTION 8: *The density of the conditioning variables given the past is bounded and continuous.*

Assumption 4 is a standard smoothness assumption that is weaker than Assumption 4 in Donald, Imbens, and Newey (2003), which requires twice continuous differentiability. In addition, contrary to Donald, Imbens, and Newey (2003) we do not require any smoothness condition of h with respect to y. Assumption 5 bounds the amount of dependence in the sample. These assumptions are very weak and allow for many types of weak and strong dependence for the process Z_t. Assumption 6 is standard. Assumptions 7 and 8 also restrict the dependence of the conditioning variables with respect to the past. Conditions similar to Assumption 7 and 8 have been employed by Koul and Stute (1999); see their assumptions (A)(b) and (B) on pp. 218–219. Notice that under independence, Assumption 7 can be relaxed to $Eh^2(Y_t, \theta_0) < \infty$, and Assumption 8 can be deleted, similarly to Stute (1997). Hence, for the independence case, no assumption concerning X_t would be required.

Next, we state the asymptotic normality theorem, proof of which is in the Appendix.

THEOREM 2: *Under Assumptions* 1–8

$$\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d}\left(\int \dot{H}\dot{H}'\,dP_{X_1}\right)^{-1}\int \dot{H}B_\Gamma\,dP_{X_1},$$

where $\dot{H}(x)=E(\dot{h}(Y_t,\theta_0)I(X_t\le x))$ *and* B_Γ *denotes a centered Gaussian process in* $D[R]^d$ (*where* $D[R]^d$ *is the space of real functions that are continuous from above and with limits from below*; *see Bickel and Wichura* (1971)), *with covariance structure given by* $\Gamma(r,s)=E(h^2(Y_t,\theta_0)I(X_t\le r\wedge s))$.

Note that, when h is homoskedastic and $d=1$, B_Γ particularizes to a scaled Brownian motion. Using the previous theorem and the fact that the integrated weighted Gaussian process follows a normal distribution (see, for instance, Tanaka (1996, Ch. 2)) the following corollary holds.

COROLLARY 1: *Under Assumptions* 1–8

$$\sqrt{n}(\widehat{\theta}-\theta_0)\xrightarrow{d}\mathrm{N}(0,\Omega),$$

where

$$\Omega=\left(\int \dot{H}\dot{H}'\,dP_X\right)^{-1}$$
$$\times\int\!\!\int \dot{H}(x_1)\dot{H}'(x_2)\Gamma(x_1,x_2)\,dP_X(x_1)\,dP_X(x_2)\left(\int \dot{H}\dot{H}'\,dP_X\right)^{-1}.$$

Our proposed estimator is consistent and asymptotically normal but inefficient. It is difficult to compare Ω with the minimum asymptotic variance for a general case. For the simplest linear location model with independents errors, Ω is 20% higher than the asymptotic variance of the sample mean. In Section 5 we evaluate its finite sample performance in a brief Monte Carlo exercise. In order to perform statistical inference, the matrix Ω needs to be estimated consistently. A simple consistent estimator of Ω is its sample analog

$$\left(\sum_{i=1}^{n}\dot{H}_n(X_i)\dot{H}_n'(X_i)\right)^{-1} \times \sum_{i=1}^{n}\sum_{j=1}^{n}\dot{H}_n(X_i)\dot{H}_n'(X_j)\Gamma_n(X_i,X_j)\left(\sum_{i=1}^{n}\dot{H}_n(X_i)\dot{H}_n'(X_i)\right)^{-1},$$

where

$$\dot{H}_n(t)=n^{-1}\sum_{i=1}^{n}\dot{h}(Y_i,\widehat{\theta})I(X_i\le t) \quad \text{and}$$

$$\Gamma_n(t,s)=n^{-1}\sum_{i=1}^{n}h^2(Y_i,\widehat{\theta})I(X_i\le t\wedge s).$$

4. A TWO-STEP EFFICIENT ESTIMATOR

The previously introduced estimator $\widehat{\theta}$ is consistent but inefficient. As commented in Section 2, the literature has focused on efficient estimation; see Donald, Imbens, and Newey (2003) and Kitamura, Tripathi, and Ahn (2000). In this section we briefly discuss a two-step efficient estimator.

Let $\widetilde{\theta}$ denote the efficient GMM estimator and $Q_n(\theta)$ denote the efficient GMM objective function that converges to $Q(\theta)$. Assume that θ_0 is locally identified by $Q(\theta)$, and let $\aleph_0$ denote a neighborhood of θ_0 such that $Q(\theta)>Q(\theta_0)$ for all $\theta\in\aleph_0$. Note that consistency of $\widehat{\theta}$ guarantees that $\widehat{\theta}\in\aleph_0$ with probability one for n large enough, and that under regularity assumptions, $\widetilde{\theta}$ is consistent and asymptotically efficient when the parameter space Θ is restricted to $\aleph_0$. Hence, using these two facts, we can modify $\widehat{\theta}$ to construct an estimator that is asymptotically efficient, by carrying out a single Newton–Raphson iterative step in the direction of the efficient GMM estimator.

Denote the gradient and Hessian of $Q_n(\theta)$ as

$$\dot{Q}_n(\theta)=\frac{\partial Q_n(\theta)}{\partial\theta} \quad \text{and} \quad \ddot{Q}_n(\theta)=\frac{\partial^2 Q_n(\theta)}{\partial\theta\,\partial\theta'}.$$

Note that in the general case, both $\dot{Q}_n(\theta)$ and $\ddot{Q}_n$ involve estimating some conditional expectations; see Newey (1993). In this section we assume that consistent nonparametric estimators for these quantities exist, such as those based on kernels, series expansions, or nearest neighbors methods; see Newey (1990) and Robinson (1991). From Young's theorem (see Serfling (1980, p. 45)),

$$0=\dot{Q}_n(\widetilde{\theta})=\dot{Q}_n(\widehat{\theta})+\ddot{Q}_n(\widehat{\theta})(\widetilde{\theta}-\widehat{\theta})+o_p(|\widetilde{\theta}-\widehat{\theta}|).$$

Since both $\widetilde{\theta}$ and $\widehat{\theta}$ are $\sqrt{n}$-consistent estimators, $o_p(|\widetilde{\theta}-\widehat{\theta}|)=o_p(n^{-1/2})$. Hence, a consistent and efficient estimator is given by

$$\widehat{\theta}_E=\widehat{\theta}-\ddot{Q}_n(\widehat{\theta})^{-1}\dot{Q}_n(\widehat{\theta}),$$

with $\widehat{\theta}_E$ satisfying

$$\widehat{\theta}_E - \widetilde{\theta} = o_p(n^{-1/2}).$$

The result is straightforward; see Robinson (1988). Although theoretically the asymptotic distribution is achieved after the first iteration, in practice carrying out additional iterations may improve the finite sample performance. Contrary to the standard two-step efficient GMM procedure, which employs some initial arbitrary estimator that may be inconsistent, the previous result suggests that a sensible empirical strategy is to compute the two-step efficient GMM estimator using $\widehat{\theta}$ as a starting point. Note that for the homoskedastic nonlinear regression model, this efficient estimator does not require any smoothing, as opposed to Donald, Imbens, and Newey (2003) and Kitamura, Tripathi, and Ahn (2000).

5. SIMULATIONS

In this section we consider Example 2 from the Introduction and report some brief Monte Carlo evidence on the finite sample performance of our method. We compare the performance of our consistent estimator $(\widehat{\theta})$, the two-step efficient estimator $(\widehat{\theta}_E)$, and the feasible efficient GMM estimator $(\widetilde{\theta})$, that uses the optimal instrument, $W = 2\theta X + X^2$. Note that in this case, $\widetilde{\theta}$ coincides with the nonlinear least squares estimator. We assume that X follows a normal distribution with unit variance and consider two values for the mean, zero and one. Recall that when X follows an $\mathrm{N}(1, 1)$ distribution, the optimal instrument does not identify the true value $\theta_0 = 5/4$, while for the other case, the optimal instrument does identify θ_0.

In Table I we report the bias, standard error (SE), and root mean squared error (RMSE) for the three estimators for three sample sizes, $n = 50$, 100, and 200. The number of replications is 5,000 in all experiments. Table I indicates that when the optimal instrument does not identify the true value θ_0, $\widetilde{\theta}$ is unreliable as the theory predicts. When $n = 50$ the RMSE of $\widetilde{\theta}$ is 26 times higher than that of our consistent estimator and when $n = 200$ this ratio increases to 47. For the two-step estimator these ratios are even higher, 54 when $n = 50$ and 106 when $n = 200$. Note that for $\widetilde{\theta}$, neither its bias nor its standard error decreases with the sample size. For the $\mathrm{N}(0, 1)$ case, for $n = 50$, both

TABLE I

BIAS, STANDARD ERROR, AND ROOT MEAN SQUARED ERROR

		Bias			SE			RMSE		
X	n	$\widehat{\theta}$	$\widehat{\theta}_E$	$\widetilde{\theta}$	$\widehat{\theta}$	$\widehat{\theta}_E$	$\widetilde{\theta}$	$\widehat{\theta}$	$\widehat{\theta}_E$	$\widetilde{\theta}$
N(0, 1)	50	.004	.005	−.013	.112	.079	.181	.112	.079	.182
	100	.002	−.001	−.001	.080	.035	.050	.081	.035	.050
	200	.001	−.001	.000	.058	.024	.024	.058	.024	.024
N(1, 1)	50	.000	−.004	−.406	.048	.022	1.167	.048	.023	1.235
	100	.000	−.005	−.362	.035	.015	1.102	.035	.016	1.160
	200	.000	.000	−.364	.025	.011	1.113	.025	.011	1.171

TABLE II

COVERAGE PERCENTAGES

		90%			95%			99%		
X	n	$\widehat{\theta}$	$\widehat{\theta}_E$	$\widetilde{\theta}$	$\widehat{\theta}$	$\widehat{\theta}_E$	$\widetilde{\theta}$	$\widehat{\theta}$	$\widehat{\theta}_E$	$\widetilde{\theta}$
N(0, 1)	50	89.2	89.0	90.3	94.2	93.8	95.1	98.2	98.3	98.6
	100	89.3	89.5	90.0	94.7	94.5	95.2	98.6	99.0	99.1
	200	90.1	89.7	89.7	94.9	94.7	94.5	98.7	98.7	98.9
N(1, 1)	50	90.4	91.2	81.2	95.3	95.9	85.4	99.2	99.3	88.4
	100	90.5	91.5	82.2	95.4	96.0	86.1	99.1	99.3	89.1
	200	90.1	91.0	82.0	94.7	95.8	86.4	99.1	98.9	89.3

$\widehat{\theta}$ and $\widehat{\theta}_E$ perform better than $\widetilde{\theta}$, although for $n = 200$, the RMSE of $\widehat{\theta}$ is larger than that of $\widetilde{\theta}$ and $\widehat{\theta}_E$.

In Table II we report the coverage percentages for 90%, 95%, and 99% confidence intervals for the three estimators. For the N(0, 1) case these coverage percentages are quite accurate for the three estimators for any sample size. However, for the N(1, 1) case, the coverage probabilities of the efficient GMM estimator present substantial distortions that do not vanish by increasing the sample size.

In order to gain more insight for this example, in Figure 1 we have plotted the asymptotic objective functions that our consistent estimator (solid lines) and the

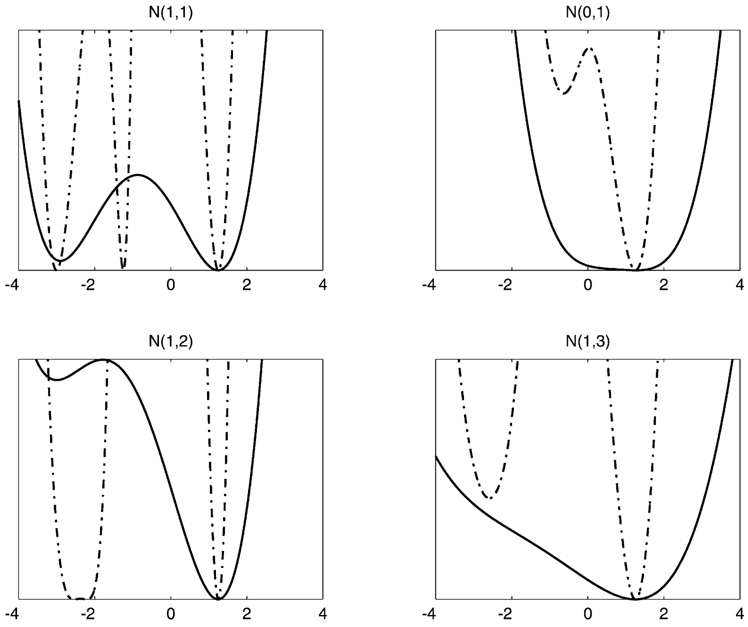

FIGURE 1.—Asymptotic objective functions.

efficient GMM estimator (dashed lines) minimize for different distributions for X. Note that these objective functions are polynomials on θ of degrees 4 and 6, respectively. As commented in Section 1, when X follows an $\mathrm{N}(1, 1)$, the efficient GMM objective function presents three global minima (at 1.25, -1.25, and -3) whereas our objective function presents a unique global minimum at 1.25 and a local minimum at approximately -2.93. When X follows an $\mathrm{N}(0, 1)$, the efficient GMM objective function identifies globally θ_0 but it also has a local minimum whereas our objective function is convex. Figure 1 also plots the asymptotic objective functions when X follows an $\mathrm{N}(1, 2)$ and an $\mathrm{N}(1, 3)$.

6. DISCUSSION

There are two approaches to consistently estimate models defined by conditional moment restrictions. The first approach, which we follow in this article, substitutes the conditional restrictions by an infinite number of unconditional moment restrictions that fully characterize the conditional restrictions. In our case, the infinite unconditional restrictions arise by considering the expectation of the function of interest times a class of indicators functions. Alternative classes of functions, such as the exponentials, could have been employed; see Bierens (1990) and Carrasco and Florens (2000). The second approach fits the conditional expectation that defines the model by means of nonparametric methods. This approach has been followed by Donald, Imbens, and Newey (2003) and by Kitamura, Tripathi, and Ahn (2000). The main difference between both approaches resides in the number of unconditional restrictions effectively employed in finite samples. Whereas infinite moment restrictions are employed in the first approach, the second approach employs a finite number of them, where this number is determined by a smoothing parameter that increases to infinity. The main advantage of introducing this smoothing number is that it allows the derivation of estimators that are asymptotically efficient. However, in the absence of automatic data-dependent methods for selecting this smoothing number, such as cross-validation procedures, a researcher faces the difficulty of selecting it for her particular case. In many cases, statistical inference is very sensitive to this selection.

Asymptotically efficient estimators can also be derived in the first approach. However, deriving them would also require the introduction of a bandwidth parameter necessary to avoid a singularity problem; see Carrasco and Florens (2000). The estimator proposed in this article is consistent and very simple to implement since it does not require the introduction of any user chosen object such as the order of a lag or a bandwidth number. It possesses the additional advantages of being applicable to cross-sectional and to a wide variety of time series data, of allowing for instruments with unbounded support, and of imposing mild smoothness conditions on the function that defines the model. In addition, the techniques employed in this article are different from those used in the previous references. Finally, in Section 4 we have shown that our estimator can be modified to achieve the semiparametric efficiency bound, although this modification may require the introduction of smoothing estimators.

We finish with a suggestion on further research. Similarly to the GMM overidentifying restriction test, by evaluating our objective function at our estimate, we can perform specification testing for conditional parametric models such as (1). Whereas the GMM overidentifying restriction test is not consistent since the number of unconditional restrictions tested is finite, our test would be consistent since it would use an infinite number of them.

Centro de Investigación Económica, Instituto Tecnológico Autónomo de México (ITAM), Av. Camino a Santa Teresa #930, Col. Héroes de Padierna, 10700 México, D.F., Mexico; madt@itam.mx

and

Centro de Investigación Económica, Instituto Tecnológico Autónomo de México (ITAM), Av. Camino a Santa Teresa #930, Col. Héroes de Padierna, 10700 México, D.F., Mexico; ilobato@itam.mx.

Manuscript received February, 2003; final revision received February, 2004.

APPENDIX

Unless explicitly stated, the summations run from 1 to n.

PROOF OF THEOREM 1: In Section 2 we have shown that $\int H(\theta,x)^2\,dP_{X_t}(x)$ has a unique minimum at θ_0. Then, using theory of M-estimators we just have to show that

$$\int H_n(\theta,x)^2\,dP_n(x)\overset{\text{a.s.}}{\to}\int H(\theta,x)^2\,dP_{X_t}(x)\quad\text{uniformly in }\theta,$$

where $P_n(x)=n^{-1}\sum_{i=1}^n I(X_i=x)$ is the empirical analog of $P_{X_t}(x)$. This result holds applying the Continuous Mapping theorem since

$$H_n(\theta,x)\overset{\text{a.s.}}{\to}H(\theta,x)\quad\text{uniformly in }(x,\theta),$$

which follows from Ranga Rao (1962). *Q.E.D.*

PROOF OF THEOREM 2: The first-order conditions of the minimization problem are

$$\sum_\ell\Big[\sum_t \dot{h}(Y_t,\widehat{\theta})I(X_t\le X_\ell)\Big]\Big[\sum_t h(Y_t,\widehat{\theta})I(X_t\le X_\ell)\Big]=0.$$

Let denote $h_t(\theta)=h(Y_t,\theta)$ and $\dot{h}_t(\theta)=\dot{h}(Y_t,\theta)$. Assumption 6 and the mean value theorem imply that for some random $\lambda\in[0,1]$ and $\theta^*=\lambda\theta_0+(1-\lambda)\widehat{\theta}$, we can write

$$\sum_\ell\Big[\sum_t \dot{h}_t(\widehat{\theta})I(X_t\le X_\ell)\Big]\Big[\sum_t h_t(\theta_0)I(X_t\le X_\ell)\Big]+G_n(\widehat{\theta}-\theta_0)=0,$$

where

$$G_n=\sum_\ell\Big[\sum_t \dot{h}_t(\widehat{\theta})I(X_t\le X_\ell)\Big]\Big[\sum_t \dot{h}_t(\theta^*)I(X_t\le X_\ell)\Big].$$

Therefore,

$$\sqrt{n}(\widehat{\theta}-\theta_0)=n^3G_n^{-1}\bigg(\frac{1}{n}\sum_\ell\Big[\frac{1}{n}\sum_t \dot{h}_t(\widehat{\theta})I(X_t\le X_\ell)\Big]\Big[\frac{1}{\sqrt{n}}\sum_t h_t(\theta_0)I(X_t\le X_\ell)\Big]\bigg).$$

Then, the result follows from the continuous mapping theorem, Lemmas 1 and 2 below, and using Assumption 4, which guarantees that $n^{-3}G_n\to_{\text{a.s.}}\int \dot{H}\dot{H}'\,dP_{X_t}$. *Q.E.D.*

LEMMA 1: *Let θ^* be a consistent estimator of θ_0. Under Assumptions* 1–8,

$$\frac{1}{n}\sum_t \dot{h}_t(\theta^*)I(X_t\le x)\overset{a.s.}{\to}E\big(\dot{h}(\theta_0)I(X_t\le x)\big)=\dot{H}(x)\quad\textit{uniformly in } x.$$

The proof of this lemma is omitted since it follows from Ranga Rao (1962).

LEMMA 2: *Under Assumptions* 1–8,

$$\frac{1}{\sqrt{n}}\sum_t h_t(\theta_0)I(X_t \leq \cdot) \Rightarrow B_\Gamma$$

where $\Rightarrow$ *denotes weak convergence in* $D[\mathbb{R}]^d$, *and* $D[\mathbb{R}]^d$ *is the natural extension of* $D[0,1]^d$ *in the sense of Stute* (1997) *and* $D[0,1]^d$ *is defined in Bickel and Wichura* (1971), *Neuhaus* (1971), *or Straf* (1970).

PROOF OF LEMMA 2: For simplicity, we introduce the notation $H_n(x) = H_n(\theta_0, x)$. According to Bickel and Wichura (1971), we need to show that the finite-dimensional distributions of the process $\sqrt{n}H_n(x)$ are asymptotically normal with the appropriate covariance matrix and that the process $\sqrt{n}H_n(x)$ is tight.

Convergence of finite-dimensional distributions refers to the weak convergence of vectors of the form $(\sqrt{n}H_n(x_1), \sqrt{n}H_n(x_2), \ldots, \sqrt{n}H_n(x_q))$, for arbitrary $q \in \mathbb{N}$ and $x_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, q$. This result can be obtained using the Corollary 3.1 in Hall and Heyde (1980).

In order to prove tightness, some definitions are required. Let $\{W_n(t) : t \in \mathbb{R}^d,\ n = 1, 2, \ldots\}$ be a sequence of stochastic processes in some metric space of functions $\mathbb{G}$. Then, $\{W_n\}$ is *tight* if and only if for any $\delta > 0$ there exists a compact set $\mathbb{K} \subset \mathbb{G}$ depending on δ, such that

$$\sup_n P(W_n \in \mathbb{K}) > 1 - \delta. \tag{4}$$

Let $D_1 = (s^1, t^1] = \times_{j=1}^d (s_j^1, t_j^1]$, and $D_2 = (s^2, t^2] = \times_{j=1}^d (s_j^2, t_j^2]$ be two *intervals* in $\mathbb{R}^d$. Then, D_1 and D_2 are *neighbor intervals* if and only if for some $j^* \in \{1, 2, \ldots, d\}$, $(s_{j^*}^1, t_{j^*}^1] \neq (s_{j^*}^2, t_{j^*}^2]$, $\times_{j \neq j^*}(s_j^1, t_j^1] = \times_{j \neq j^*}(s_j^2, t_j^2]$ and $t_{j^*}^1 = s_{j^*}^2$, that is, if and only if they are next to each other and share the j^*th face. Each stochastic process indexed by a parameter in $\mathbb{R}^d$ has an associated *process indexed by the intervals* that is defined as

$$W_n(D_h) = \sum_{e_1=0}^{1} \cdots \sum_{e_d=0}^{1} (-1)^{d-\sum_j e_j} W_n\big(s_1^j + e_1(t_1^j - s_1^j), \ldots, s_d^j + e_d(t_d^j - s_d^j)\big) \quad (h = 1, 2).$$

In this proof we verify Kolmogorov–Chentsov's criterion, which is a sufficient condition for (4) according to Bickel and Wichura (1971, p. 1658).

In what follows we will simplify further the notation by writing h_t instead of $h_t(\theta_0)$. In our case, the process

$$\sqrt{n}H_n(x) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} h_t I(X_t \leq x),$$

has associated with it the following process indexed by the intervals

$$\sqrt{n}H_n(D_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} [h_t I_t(D_j)],$$

where $I_t(D_j) = I(X_t \in D_j)$. Then

$$\begin{aligned} &E\big((\sqrt{n}H_n(D_1))^2(\sqrt{n}H_n(D_2))^2\big) \\ &\quad = \frac{1}{n^2}E\left\{\sum_{t=1}^{n}\sum_{s=1}^{n}\sum_{u=1}^{n}\sum_{v=1}^{n}[h_t I_t(D_1)][h_s I_s(D_1)][h_u I_u(D_2)][h_v I_v(D_2)]\right\}. \end{aligned}$$

Case 1: Assume that Z_t and Z_{t-s} have no common coordinates for any $s \neq 0$. Using that h_t is a centered MDS, the nonzero terms are those such that the greater subindex appears at least twice. Moreover, notice that when a subindex appears three times, the corresponding term is zero because D_1 and D_2 are disjoint sets. For the same reason, terms like $[h_t^2 I_t(D_1) I_t(D_2)][h_u I_u(D_1)][h_v I_v(D_2)]$ are also zero. Therefore,

$$
\begin{aligned}
E\big((\sqrt{n}H_n(D_1))^2(\sqrt{n}H_n(D_2))^2\big) &= \frac{1}{n^2}E\left\{\sum_{t=1}^{n}[h_t^2 I_t(D_1)]\left(\sum_{s=1}^{t-1}[h_s I_s(D_2)]\right)^2\right\} \\
&\quad + \frac{1}{n^2}E\left\{\sum_{t=1}^{n}[h_t^2 I_t(D_2)]\left(\sum_{s=1}^{t-1}[h_s I_s(D_1)]\right)^2\right\}.
\end{aligned}
$$

Under our assumptions, these expectations exist. Note that both terms are analyzed similarly since the only difference is the indexing set D_j.

Now, denote the σ-algebra generated by $\{Z_{t-1}, Z_{t-2}, \ldots\}$ by $\Im_{t-1}$. Then

$$
\begin{aligned}
&\frac{1}{n^2}E\left\{\sum_{t=1}^{n}[h_t^2 I_t(D_1)]\left(\sum_{s=1}^{t-1}[h_s I_s(D_2)]\right)^2\right\} \\
&\quad = \frac{1}{n^2}\sum_{t=1}^{n}E\left\{\sigma^2(\Im_{t-1}, X_t) I_t(D_1)\left(\sum_{s=1}^{t-1}[h_s I_s(D_2)]\right)^2\right\}
\end{aligned}
$$

where $\sigma^2(\Im_{t-1}, X_t)$ denotes the conditional variance of h_t given $\Im_{t-1}$ and X_t. This last expression equals

$$
\frac{1}{n^2}\sum_{t=1}^{n}E\left\{\int_{D_1}\sigma^2(\Im_{t-1}, e) f_{X_t|\Im_{t-1}}(e)\,de\left(\sum_{s=1}^{t-1}[h_s I_s(D_2)]\right)^2\right\},
$$

and applying Fubini's theorem, it is

$$
\frac{1}{n^2}\sum_{t=1}^{n}\int_{D_1}E\left\{\sigma^2(\Im_{t-1}, e) f_{X_t|\Im_{t-1}}(e)\left(\sum_{s=1}^{t-1}[h_s I_s(D_2)]\right)^2\right\}de.
$$

Now, using Cauchy–Schwarz's inequality, the term is bounded above by

$$
\frac{1}{n^2}\sum_{t=1}^{n}\int_{D_1}E^{1/2}\big[\sigma^2(\Im_{t-1}, e) f_{X_t|\Im_{t-1}}(e)\big]^2 E^{1/2}\left[\sum_{s=1}^{t-1}h_s I_s(D_2)\right]^4 de.
$$

Using the Burkhölder's inequality, and denoting by K some generic constant, this term is bounded by

$$
\begin{aligned}
&\frac{K}{n^2}\sum_{t=1}^{n}\int_{D_1}E^{1/2}\big[\sigma^2(\Im_{t-1}, e) f_{X_t|\Im_{t-1}}(e)\big]^2 E^{1/2}\left[\sum_{s=1}^{t-1}h_s^2 I_s(D_2)\right]^2 de \\
&\quad \le \frac{K}{n^2}\sum_{t=1}^{n}\int_{D_1}E^{1/2}\big[\sigma^2(\Im_{t-1}, e) f_{X_t|\Im_{t-1}}(e)\big]^2 (t-1)E^{1/2}h_1^4 I_1(D_2)\,de \\
&\quad \le \frac{K}{n^2}\sum_{t=1}^{n}\mu_1(D_1)(t-1)\mu_2^{1/2}(D_2) \le K\mu_1(D_1\cup D_2)\mu_2^{1/2}(D_1\cup D_2),
\end{aligned}
$$

where

$$\mu_1(D) = \int_D E^{1/2}\big[\sigma^2(\Im_{t-1}, e) f_{X_t|\Im_{t-1}}(e)\big]^2 \, de \quad \text{and}$$

$$\mu_2(D) = E h_1^4 I_1(D),$$

which is a Chentsov's criterion provided that μ_1 and μ_2 are finite. This final condition is straightforward to check using Hölder's inequality and Assumptions 7 and 8, as in Koul and Stute (1999).

Case 2: In case Z_t and Z_{t-s} have common coordinates, define b as the minimum s such that Z_t and Z_{t-j} have no common coordinates for all $j \geq s$. Note that $b \leq k + d$. In this case the process

$$\frac{1}{\sqrt{n}} \sum_t h_t I(X_t \leq x)$$

can be rewritten as

$$\sum_{s=0}^{b-1} \left\{ \sum_{j=1}^{[n/b+1]} \frac{1}{\sqrt{n}} h_{sb+j} I(X_{sb+j} \leq x) \right\},$$

which are b different processes. Tightness of each of these b processes follows similarly as above, and, since b is fixed and finite, tightness of the original process follows from the tightness of each of its components. *Q.E.D.*

REFERENCES

AMEMIYA, T. (1974): "The Nonlinear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 2, 105–110.

——— (1977): "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model," *Econometrica*, 45, 955–968.

BICKEL, P. J., AND M. J. WICHURA (1971): "Convergence Criteria for Multiparameter Stochastic Processes and Some Applications," *Annals of Mathematical Statistics*, 42, 1656–1670.

BIERENS, H. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458.

BILLINGSLEY, P. (1995): *Probability and Measure*. New York: Wiley and Sons.

BRUNK, H. D. (1970): "Estimation for Isotonic Regression," in *Nonparametric Techniques in Statistical Inference*, ed. by M. L. Puri. Cambridge: Cambridge University Press, 177–197.

CARRASCO, M., AND J. P. FLORENS (2000): "Generalization of GMM to a Continuum of Moment Conditions," *Econometric Theory*, 16, 797–834.

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

DONALD, S. G., G. W. IMBENS, AND W. NEWEY (2003): "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions," *Journal of Econometrics*, 117, 55–93.

HALL, P., AND C. C. HEYDE (1980): *Martingale Limit Theory and Its Application*. New York: Academic Press.

HANSEN, L. P. (1982): "Large-Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

KITAMURA, Y., G. TRIPATHI, AND H. AHN (2000): "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models," forthcoming in *Econometrica*.

KOUL, H. L. (2002): *Weighted Empirical Processes in Dynamic Nonlinear Models*. New York: Springer-Verlag.

KOUL, H. L., AND W. STUTE (1999): "Nonparametric Model Checks for Time Series," *The Annals of Statistics*, 27, 204–236.

NEUHAUS, G. (1971): "On Weak Convergence of Stochastic Processes with Multidimensional Time Parameter," *Annals of Mathematical Statistics*, 42, 1285–1295.

NEWEY, W. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

——— (1993): "Efficient Estimation of Models with Conditional Moment Restrictions," in *Handbook of Statistics*, *Volume 11*, *Econometrics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod. Amsterdam: North-Holland.

NEWEY, W., AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

RANGA RAO, R. (1962): "Relations Between Weak and Uniform Convergence of Measures with Applications," *Annals of Mathematical Statistics*, 33, 659–680.

ROBINSON, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–891.

——— (1988): "The Stochastic Difference Between Econometric Statistics," *Econometrica*, 56, 531–548.

——— (1991): "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," *Econometrica*, 59, 755–786.

SERFLING, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. New York: Wiley and Sons.

STRAF, M. L. (1970): "Weak Convergence of Stochastic Processes with Several Parameters," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 187–221.

STUTE, W. (1997): "Nonparametric Model Checks for Regression," *The Annals of Statistics*, 25, 613–641.

TANAKA, K. (1996): *Time Series Analysis*. New York: Wiley and Sons.