CONSISTENT ESTIMATION OF MODELS DEFINED BY CONDITIONAL MOMENT RESTRICTIONS

Manuel A. Domínguez and Ignacio N. Lobato

Instituto Tecnológico Autónomo de México
Av. Camino a Santa Teresa #930
Col. Héroes de Padierna
10700 México, D.F., MEXICO
E-mail: ilobato@itam.mx

February, 2003

Abstract

Models stated as conditional moment restrictions are typically estimated in econometrics by means of the generalized method of moments (GMM). The GMM estimation procedure can render inconsistent estimates since the instruments are arbitrarily chosen and the method 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 which 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.

Keywords and Phrases: Generalized Method of Moments, Identification, Unconditional Moments, Marked Empirical Process, Integrated Regression Function.

JEL classification numbers: C12 and C52

1 Introduction

In many areas of econometrics such as panel data, dicrete choice, macroeconomics and finance, models 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. Commonly, these models are estimated using the Generalized Method of Moments (GMM) that basically 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 conditions (of dimension equal to the dimension of the parameter vector) identifies globally the parameters of interest, and hence, the GMM procedure provides consistent estimators for them. However, in many nonlinear models the selected unconditional moment restrictions may hold for several parameter values. In these cases the arbitrarily chosen unconditional conditions 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 $E(Y \mid X) = X^{\theta_0}$ where $\theta_0 = 4$, and X is a symmetric around zero random variable whose fourth and sixth moments are equal, such as a N(0,1/5). Now, the researcher specifies correctly the model $E(Y \mid X) = X^{\theta}$ where $\theta \in \Theta = [2, \infty)$, and sets out to estimate θ_0. The model implies that $(Y - X^{\theta_0})$ is orthogonal to any function g such that $E\left|(Y - X^{\theta_0})g(X)\right| < \infty$. Since there is only one parameter, the researcher needs to select at least one function $g(X)$. Let assume that she selects the functions 1 and X. The problem is that these two instruments do not identify the parameter value $\theta_0 = 4$ since the value $\theta = 6$ also solves the system of equations $E(Y - X^{\theta}) = E((Y - X^{\theta})X) = 0$. Of course, more arbitrary instruments could be added, but it would always be simple to find a particular distribution for X, such that θ_0 and additional values for θ would satisfy the new set of orthogonality conditions.

Example 2. Assume that the random variable Y satisfies $E(Y \mid X) = \theta_0^2 X + \theta_0 X^2$ where X is a N(1,1) random variable and $\theta_0 = 5/4$. In addition, suppose that $V(Y \mid X)$ is constant. Now, assume that the researcher properly specifies the model and, instead of an arbitrary instrument, she chooses 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 when $\theta = -5/4$.

These simple examples illustrate that the estimation procedure based on selecting an arbitrary finite number of instruments (even the optimal ones) does not guarantee that the parameters of interest are globally identified. Hence, in order to achieve global identification, GMM needs to introduce the additional assumption that the selected unconditional restrictions identify globally the parameter of interest. This additional assumption implies additional restrictions on the marginal distribution of the conditioning variables which are introduced for statistical convenience and without any relation to the underlying economic (conditional) model. Thus, the introduction of these restrictions leads to the following paradox: while the distribution of the conditioning variables should be irrelevant for the 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 addition, applied researchers are typically unaware of these restrictions, and faced to estimating (possibly highly complicated) nonlinear models, they just choose arbitrary instruments and estimate by GMM *assuming* that the parameter vector is globally identified. Obviously, this estimation procedure can lead to completely misleading inferences.

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 which 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. The paper is organized as follows. Section 2 introduces the framework and our estimator. Section 3 establishes the asymptotic theory and Section 4 concludes. The proofs are contained in the Appendix.

2 Notation and framework

Let Z_t be a vector time series 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 instrumental variables (again, a finite number of these variables lagged can be included). The coordinates of Z_t are related by an econometric model which establishes that the true distribution of the data satisfies

the following conditional moment restrictions

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

for a unique $\theta_0 \in \Theta$ where $\Theta \subset \mathbb{R}^m$. This conditional moment model is given to the econometrician by economic theory. 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 instrumental variables estimators have been proposed, see among others, Amemiya (1974, 1977), Jorgenson and Laffont (1974), Berndt, Hall, Hall and Hausman (1974), Burguete, Gallant and Souza (1982), Hansen (1982), Newey (1990a, b). However, none of these references addressed the identification problem commented above. For instance, Newey (1990a) considered a similar model (see his equation (2.1) in p.810) in a more restrictive framework (he considered i.i.d data with homoskedasticity) and focused on the optimality properties of a selected estimator. However, he overlooked the identification problem by assuming that the parameter vector is globally identified by the selected unconditional conditions, see his assumption 3.3 (a) in p.817.

Recently, Donald, Imbens and Newey (2001) 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. This bandwidth number allows their estimators to be root$-n$ asymptotically normal and efficient, but it is unclear the sensitivity of the estimator to the selection of this bandwidth number. On the contrary, our approach does not require the introduction of an arbitrary user-chosen number to achieve an asymptotically normal distribution. Although the asymptotic variance for the Donald et al. (2001) estimator is lower than ours, statistical inference with this estimator can be sensitive to the selection of the bandwidth number. Furthermore, their procedure restricts to the i.i.d. setting, and for most of their results, the density of the conditioning variables has to be bounded from zero on a compact, rectangular support. On the contrary, our procedure is

very simple, allows for instruments with unbounded support and can be used for time series data.

Another related reference is Carrasco and Florens (2000). They consider optimal GMM estimation for the case where there is a continuum of moment conditions. 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 constant. Carrasco and Florens' estimator is efficient, but efficiency is achieved at the cost of introducing a user-chosen smoothing number necessary to avoid a singularity problem associated with the inversion of a linear bounded operator. As in the case of Donald, Imbens and Newey 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. Hence, 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

$$E(h(Y,\theta_0) \mid X) = 0 \ a.s. \Leftrightarrow H(\theta_0, x) = 0 \text{ for all } x \in \mathbb{R}^d, \tag{2}$$

where $H(\theta, x) = E(h(Y,\theta)I(X \leq x))$ is the integrated regression function (Brunk, 1970) and the indicator function $I(X \leq x)$ equals 1 when each component in X is less or equal than the corresponding component in x; and equals 0 otherwise. In addition, from (1), $P\left(E(h(Y,\theta) \mid X) = 0\right) < 1$ when $\theta \neq \theta_0$, and then $H(\theta, x) \neq 0$ in a non null set of the sample space of X. Therefore, denoting by P_X the probability distribution function of the random vector X, $\int H(\theta_0, x)^2 dP_X(x) = 0$ but $\int H(\theta, x)^2 dP_X(x) > 0 \ \ \forall \theta \neq \theta_0$. Equivalently, let S denote a random vector with probability distribution function P_X and independent of (Y, X), then $E(H(\theta, S)^2) \geq 0$ with equality if and only if $\theta = \theta_0$. Hence, by the law of iterated expectations we can write

$$\theta_0 = \arg\min_{\theta \in \Theta} E\left(E\left[(h(Y,\theta)I(X \leq S) \mid S\right]^2\right) = \arg\min_{\theta \in \Theta} \int E^2\left[(h(Y,\theta)I(X \leq s)\right] dP_X(s). \tag{3}$$

Now, calling n the sample size, $n^{-1}\sum_{t=1}^{n} h(Y_t,\theta)I(X_t \leq s)$ and $n^{-1}\sum_{\ell=1}^{n} g^2(X_\ell)$ are the sample analogs of $E(h(Y,\theta)I(X \leq s))$ and $\int g^2(s) dP_X(s)$ respectively. Then, we propose estimating θ_0 by the sample analogue of (3), that is,

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

In the next section we establish the asymptotic theory of the proposed estimator.

3 Asymptotic Theory

We start by enumerating the assumptions necessary for the consistency and asymptotic normality of our estimator. In what follows $|\cdot|$ denotes the Euclidean norm in the corresponding Euclidean space, and we assume that all the considered functions are Borel measurable.

Assumption A1. $E\left|h(Y_t,\theta)\right| < \infty$ for all $\theta \in \Theta$, and $E(h(Y_t,\theta) \mid X_t) = 0$ $a.s.$ if and only if $\theta = \theta_0$.

Assumption A2. Z_t is ergodic and strictly stationary.

Assumption A3. $h(y,\cdot)$ is continuous in Θ for each y in $\mathbb{R}^k$ and for all $\theta \in \Theta$ there exists $\rho_\theta > 0$ such that $E\left[\sup_{\{\|\theta-\theta'\|<\rho_\theta\}\cap\Theta}\left|h(Y_t,\theta) - h(Y_t,\theta')\right|\right] < \infty$.

Assumption A4. $\Theta \subset \mathbb{R}^m$ is compact.

Assumption A5. $h(Y_t,\theta_0)$ given X_t has a bounded conditional density function which is continuous on any conditioning argument.

Assumption A6. $h(\cdot,\theta)$ is once continuously differentiable in a neighbourhood of θ_0 and satisfies that $E\left[\sup_{\theta\in\aleph_0}\left|\dot{h}\left(Y_t,\theta\right)\right|\right] < \infty$ where $\aleph_0$ denotes a neighbourhood of θ_0 and $\dot{h}\left(Y_t,\theta\right) = \partial h(Y_t,\theta)/\partial\theta$.

Assumption A7. $h(Y_t,\theta_0)$ is a martingale difference sequence with respect to $\{X_s,\, s \le t\}$.

Assumption A8. $\theta_0 \in int(\Theta)$.

Assumption A9. Z_t satisfies that $E\left|h(Y_t,\theta_0)\right|^{4+\delta} < \infty$, for some $\delta > 0$.

Assumption A1 defines the model and identifies globally θ_0. This identification condition is given by the economic theory. Assumption A3 is a smoothness condition which is weaker than the Lipschitz condition in Assumption 3 in Donald, Imbens and Newey (2001). Assumption A5 imposes boundedness of the conditional density while assumptions 3 and 4 in Donald, Imbens and Newey (2001) impose boundedness of conditional moments restricting some forms of conditional heteroskedasticity. Assumption A6 is a standard smoothness assumption that is weaker than Assumption 4 in Donald, Imbens and Newey (2001) which require twice continuous differentiability. Assumptions A2 and A7 bound 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. Regarding assumption A9, notice that for the independent sampling case the condition $E\left|h(Y_t,\theta_0)\right|^{4+\delta} < \infty$ could be relaxed to $E\left|h(Y_t,\theta_0)\right|^{2} < \infty$. Opposite to standard GMM, all our assumptions refer to the uncon-

ditional or to the conditional distribution of h, and nothing is imposed on the marginal distribution of X_t.

Next, we state the consistency and asymptotic normality theorems. Their proofs are in the Appendix.

Theorem 1. Under assumptions A1-A4 $\widehat{\theta} \rightarrow_{a.s.} \theta_0$.

Theorem 2. Under assumptions A1-A9

$$\sqrt{n}(\widehat{\theta} - \theta_0) \rightarrow_d \left(\int \dot{H} \dot{H}' dP_X \right)^{-1} \int \dot{H} \, B_\Gamma dP_X$$

where $\dot{H}(t) = E(\dot{h}(Y, \theta_0) I(X \leq t))$ and B_Γ denotes a centered Brownian motion with covariance structure given by $\Gamma(t, s) = E(h^2(Y, \theta_0) I(X \leq t \wedge s))$.

Using the previous theorem and the fact that the integrated weighted Brownian motion follows a normal distribution (see, for instance, Tanaka (1996, Chapter 2)) the following corollary holds.

Corollary. Under assumptions A1-A9

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

where

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

Our proposed estimator is consistent and asymptotically normal but inefficient. It is difficult to compare Ω with the minimum asympotic 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 order to perform statistical inference the matrix Ω needs to be estimated consistently. A simple consistent estimator of Ω is its sample analogue

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

where $\dot{H}_n(t) = n^{-1} \sum_{i=1}^{n} \dot{h}(Y_i, \widehat{\theta}) I(X_i \leq t))$ and $\Gamma_n(t, s) = n^{-1} \sum_{i=1}^{n} h^2(Y_i, \widehat{\theta}) I(X_i \leq t \wedge s))$.

4 Discussion

There are two approaches to estimate consistently models stated with conditional moment restrictions. The first approach, which we follow in this article, substitute the conditional

restriction by an infinite number of unconditional moment restrictions that fully characterizes 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 indexed by a set of nuisance parameters. 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 et al. (2001), where they consider a variety of nonparametric estimators such as orthogonal series or splines. The main difference between both approaches resides in the number of unconditional restrictions effectively employed in finite samples. Whereas an infinity (continuum) of moment restrictions is employed in the first approach, the second approach employs a finite number of them where this number is determined by a smoothing parameter. The main adavantage of introducing this smoothing number is that it allows to derive 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 a wide variety of time series data, allowing for instruments with unbounded support and imposing mild smoothness conditions on the function that defines the model. Finally, the techniques employed in this article are different from those used in the second approach. We end with three suggestions on further research.

First, similar to the GMM literature, specification test for conditional moment models could be developed by using procedures similar to the ones employed in this paper, see Domínguez and Lobato (2002). Second, one of the main points of this paper is to show that the GMM procedure to estimate conditional moment models is fundamentally flawed because it imposes the additional assumption that the selected instruments identify globally the parameters of interest. However, if this additional assumption holds, the GMM estimator is consistent for the true θ_0. Then, it is of interest to test the null hypothesis that the GMM estimator is consistent for θ_0. This could be tested with a Hausman-type test which would

measure the distance between our estimator and the GMM estimator. Finally, since the proposed estimator is asymptotically pivotal, employing the bootstrap would lead to obtain an asymptotic refinement.

Acknowledgment

We thank M. Carrasco and W. Newey for useful conversations. Domínguez acknowledges financial support from Consejo Nacional de Ciencia y Tecnología (CONACYT) under project grant J38276D and Lobato acknowledges financial support from Asociación Mexicana de Cultura.

5 Appendix

Unless stated the summatories run from 1 to n.

Proof of the Theorem 1. Call $H_n(\theta, x) = n^{-1} \sum_t h(Y_t, \theta) I(X_t \leq x)$. In Section 2 we have shown that $\int H(\theta, x)^2 dP_X(x)$ has a unique minimum at θ_0. Then, using theory of M-estimators we just have to show that

$$\int H_n(\theta, x)^2 dP_n(x) \to_{a.s.} \int H(\theta, x)^2 dP_X(x) \text{ uniformly in } \theta,$$

where $P_n(x) = n^{-1} \sum_{i=1}^n I(X_i = x)$. This result holds applying the Continuous Mapping Theorem if it can be shown that

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

Start by defining $h^+(Y_t, \theta) = \max\{0, h(Y_t, \theta)\}$ and $h^-(Y_t, \theta) = h(Y_t, \theta) - h^+(Y_t, \theta)$. Assumption A3 implies that $E\left[\sup_{\{\|\theta - \theta'\| < \rho_\theta\} \cap \Theta} |h^+(Y_t, \theta) - h^+(Y_t, \theta')|\right] < \infty$ and similarly for h^-. If we establish the Uniform Law of Large Numbers for h^+ and for h^-, the result trivially holds for h, so without loss of generality we can assume that $h \geq 0$. Define

$$h_1(Z_t, \theta, x, \rho) = \sup_{\{\|\theta - \theta'\| < \rho_\theta\} \cap \Theta} |h(Y_t, \theta) I(X_t \leq x) - h(Y_t, \theta') I(X_t \leq x)|,$$

and

$$\overline{h}(Z_t, \theta, \rho) = \sup_x h_1(Z_t, \theta, x, \rho).$$

Note that

$$\overline{h}(Z_t, \theta, \rho) = \sup_{\{\|\theta - \theta'\| < \rho_\theta\} \cap \Theta} |h(Y_t, \theta) - h(Y_t, \theta')|$$

By the Monotone Convergence Theorem, since for all θ, $\overline{h}(Z_t, \theta, \rho)$ is decreasing in ρ, then $\lim_{\rho \to 0^+} E\left[\overline{h}(Z_t, \theta, \rho)\right] = 0$. Now, define $H(\theta, \rho) = E\overline{h}(Z_t, \theta, \rho)$. The proof goes as following. Fix $\varepsilon > 0$, then $\forall \theta$, $\exists \rho_\theta > 0$ such that $H(\theta, \rho_\theta) < \varepsilon$. Next, call $B(\theta, \rho_\theta)$ the ball with center θ and radius ρ_θ, by compactness $\exists\{B(\theta_1, \rho_{\theta_1}), ..., B(\theta_k, \rho_{\theta_k})\}$ with finite k that covers Θ. Notice that k does not depend on x. Hence,

$$
\begin{aligned}
& \sup_x \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_t h(Y_t, \theta) I(X_t \le x) - E(h(Y, \theta) I(X \le x)) \right| \\
\le \; & \sup_x \min_{j=1,...k} \left\{ \sup_{\left\{\|\theta - \theta'\| < \rho_{\theta_j}\right\} \cap \Theta} \left| \frac{1}{n} \sum_t h(Y_t, \theta) I(X_t \le x) - \frac{1}{n} \sum_t h(Y_t, \theta_j) I(X_t \le x) \right. \right. \\
& + E(h(Y, \theta_j) I(X \le x)) - E(h(Y, \theta) I(X \le x)) \\
& \left. \left. + \frac{1}{n} \sum_t h(Y_t, \theta_j) I(X_t \le x) - E(h(Y, \theta_j) I(X \le x)) \right| \right\} \\
\le \; & \sup_x \min_{j=1,...k} \left\{ \left| \frac{1}{n} \sum_t h_1(Z_t, \theta_j, x, \rho_{\theta_j}) \right| + H(\theta_j, \rho_{\theta_j}) \right. \\
& \left. + \left| \frac{1}{n} \sum_t h(Y_t, \theta_j) I(X_t \le x) - E(h(Y, \theta_j) I(X \le x)) \right| \right\} \\
\le \; & \max_{j=1,...k} \frac{1}{n} \sum_t \overline{h}(Z_t, \theta_j, \rho_{\theta_j}) + \varepsilon \\
& + \max_{j=1,...k} \sup_x \left| \frac{1}{n} \sum_t h(Y_t, \theta_j) I(X_t \le x) - E(h(Y, \theta_j) I(X \le x) \right|,
\end{aligned}
$$

where we have applied the definition of ρ_{θ_j} in the second term of the right hand side. Then,

$$
\begin{aligned}
& \overline{\lim_{n \to \infty}} \sup_x \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_t h(Y_t, \theta) I(X_t \le x) - E(h(Y, \theta) I(X \le x)) \right| \\
\le \; & \max_{j=1,...k} \lim_{n \to \infty} \frac{1}{n} \sum_t \overline{h}(Z_t, \theta_j, \rho_{\theta_j}) + \varepsilon \\
& + \max_{j=1,...k} \lim_{n \to \infty} \sup_x \left| \frac{1}{n} \sum_t h(Y_t, \theta_j) I(X_t \le x) - E(h(Y, \theta_j) I(X \le x)) \right| \\
= \; & \max_{j=1,...k} H(\theta_j, \rho_{\theta_j}) + \varepsilon + 0 \quad a.s,
\end{aligned}
$$

applying the Glivenko-Cantelli Theorem, and finally note that $H(\theta_j, \rho_{\theta_j}) < \varepsilon$ by the definition of ρ_{θ_j}. Then, it follows that

$$
\overline{\lim_{n \to \infty}} \sup_x \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_t h(Y_t, \theta) I(X_t \le x) - E(h(Y, \theta) I(X \le x)) \right| \le 2\varepsilon \quad a.s.
$$

Therefore, the result follows because ε was selected arbitrarily.

Proof of the Theorem 2. The first order conditions of the minimization problem are

$$\sum_{\ell}\left[\sum_{t}\dot{h}\left(Y_t,\widehat{\theta}\right)I(X_t\leq X_\ell)\right]\left[\sum_{t}h(Y_t,\widehat{\theta})I(X_t\leq X_\ell)\right]=0.$$

Now introducing the notation $h_t(\theta)=h(Y_t,\theta)$ and $\dot{h}_t\left(\theta\right)=\dot{h}\left(Y_t,\theta\right)$, using A8 the mean value theorem implies that for some random $\lambda\in[0,1]$ and $\widetilde{\theta}=\lambda\theta_0+(1-\lambda)\widehat{\theta}$,

$$\sum_{\ell}\left[\sum_{t}\dot{h}_t\left(\widehat{\theta}\right)I(X_t\leq X_\ell)\right]\left[\sum_{t}h_t(\theta_0)I(X_t\leq X_\ell)\right]+G_n(\widehat{\theta}-\theta_0)=0.$$

where

$$G_n=\sum_{\ell}\left[\sum_{t}\dot{h}_t\left(\widehat{\theta}\right)I(X_t\leq X_\ell)\right]\left[\sum_{t}\dot{h}_t\left(\widetilde{\theta}\right)I(X_t\leq X_\ell)\right]$$

Then,

$$\sqrt{n}(\widehat{\theta}-\theta_0)=\frac{\frac{1}{n}\sum_{\ell}\left[\frac{1}{n}\sum_{t}\dot{h}_t\left(\theta_0\right)I(X_t\leq X_\ell)\right]\left[\frac{1}{\sqrt{n}}\sum_{t}h_t(\theta_0)I(X_t\leq X_\ell)\right]}{n^{-3}G_n}$$

Then, the result follows from the continuous mapping theorem, Lemmas 1 and 2 below and using assumption A6 that guarantees that $n^{-3}G_n\rightarrow_{a.s.}\int\dot{H}^2\,dP_X$.

Lemma 1: Let θ^* be a consistent estimator of θ_0. Under assumptions A1-A9

$$\frac{1}{n}\sum_{t}\dot{h}_t\left(\theta^*\right)I(X_t\leq x)\rightarrow_{a.s.}E\left(\dot{h}\left(\theta_0\right)I(X\leq x)\right)=\dot{H}\left(x\right)\text{ uniformly in }x.$$

The proof of this Lemma is omitted since it is a trivial extension of the Glivenko-Cantelli Theorem.

Lemma 2: Under assumptions A1-A9

$$\frac{1}{\sqrt{n}}\sum_{t}h_t(\theta_0)I(X_t\leq x)\Rightarrow B_\Gamma$$

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, let 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 $\sqrt{n}H_n\left(x\right)$ are asymptotically normal with the appropriate covariance matrix and that the process $\sqrt{n}H_n\left(x\right)$ is tight.

Convergence of finite-dimensional distributions refers to the weak convergence of vectors of the form $(\sqrt{n}H_n(x_1),\ \sqrt{n}H_n(x_2),...,\sqrt{n}H_n(x_k))$, for arbitrary $k \in \mathbb{N}$ and $x_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, k$. 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 $\{W_n(t) : t \in \mathbb{R}^d,\ n = 1, 2, ..\}$ 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 $D_1 = (s^1, t^1] = \times_{k=1}^d (s_k^1, t_k^1]$, and $D_2 = (s^2, t^2] = \times_{k=1}^d (s_k^2, t_k^2]$ be two *intervals* in $\mathbb{R}^d$. Then, D_1 and D_2 are *neighbor intervals* if and only if for some $j \in \{1, 2, ..., d\}$, $(s_j^1, t_j^1] \neq (s_j^2, t_j^2]$, $\times_{k \neq j}(s_k^1, t_k^1] = \times_{k \neq j}(s_k^2, t_k^2]$ 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

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

In this proof we verify Kolmogorov-Chentsov´s criterion that is a sufficient condition for (4) according to Bickel and Wichura (1971).

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 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

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

Using that h_t is a centered MDS, the non-zero 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 $\left[h_t^2 I_t(D_1) I_t(D_2)\right]\left[h_u I_u(D_1)\right]\left[h_v I_v(D_2)\right]$ are also zero. Therefore,

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

Under the assumptions of the Theorem, these expectations exist. Note that both terms are analyzed similarly since the only difference is the index set D_j. Before continuing with the proof we need to introduce some new notation. Let Q_1 and Q_2 be two random vectors of different dimensions on the same probability space. Borrowing the notation from set theory we call $Q_1 \setminus Q_2$ a random vector on the same probability space that keeps the coordinates in Q_1 that do not appear in Q_2 and drops the other coordinates. In addition, denote $\ell = \min_{s\in\mathbb{N}}\{Y_t \setminus Y_{t-r} = Y_t \quad \forall r \geq s\}$, that is, ℓ denotes the minimum lag such that the most past coordinate in Y_t predates the most recent coordinate in $Y_{t-\ell}$.

Using that $\left(\sum_{i=1}^{l} a_i\right)^2 \leq l\sum_{i=1}^{l} a_i^2$, the first term is bounded by

$$\frac{2\ell}{n^2}\sum_{s=1}^{\ell-1}E\left\{\sum_{t=1}^{n}\left[h_t^2 I_t(D_1)\right]\left[h_{t-s}^2 I_{t-s}(D_2)\right]\right\} \tag{5}$$
$$+\frac{2\ell}{n^2}E\left\{\sum_{t=1}^{n}\left[h_t^2 I_t(D_1)\right]\left(\sum_{s=1}^{t-l}\left[h_s I_s(D_2)\right]\right)^2\right\}. \tag{6}$$

First, consider any term in (5). Define the random variable vector $V_{t,t-s} = \left(V_{t,t-s}^1, Z_{t-s}\right)$ where $V_{t,t-s}^1 = X_t \setminus Z_{t-s}$ and $\sigma^2(V_{t,t-s}) = E\left(h_t^2 \mid V_{t,t-s}\right)$ then, for any $s = 1, ..., d$,

$$E\left\{\left[h_t^2 I_t(D_1)\right]\left[h_{t-s}^2 I_{t-s}(D_2)\right]\right\} = E\left\{\left[\sigma^2\left(V_{t,t-s}\right) I_t(D_1)\right]\left[h_{t-s}^2 I_{t-s}(D_2)\right]\right\}$$
$$= E\left\{E\left[\sigma^2\left(V_{t,t-s}\right) I_t(D_1) \mid Z_{t-s}\right]\left[h_{t-s}^2 I_{t-s}(D_2)\right]\right\}$$
$$= E\left[\int \sigma^2(e, Z_{t-s})\, I(e \in D_1^{(1)},\ V_{t,t-s} \in D_1^{(2)}) f_{V_{t,t-s}^1 \mid Z_{t-s}}(e)\, de \left[h_{t-s}^2 I_{t-s}(D_2)\right]\right]$$

where $D_1 = D_1^{(1)} \times D_1^{(2)}$ is arranged according to the decomposition of $X_t = \left[V_{t,t-s}^1, \quad X_t \setminus V_{t,t-s}^1\right]$ and $f_{V_{t,t-s}^1 \mid Z_{t-s}}(e)$ denotes the density of $V_{t,t-s}^1$ conditional on Z_{t-s}. Using Fubini´s theorem and Hölder´s inequality, the last expression is bounded by

$$\int_{D_1^{(1)}} E\left(\sigma^2(e, Z_{t-s})\, I\left(X_t \setminus V_{t,t-s}^1 \in D_1^{(2)}\right)\left[h_{t-s}^2 I_{t-s}(D_2)\right] f_{V_{t,t-s}^1 \mid Z_{t-s}}(e)\right) de$$

$$\leq \int_{D_1^{(1)}} \left(E\left[\sigma^2 (e, Z_{t-s}) I\left(X_t \setminus V_{t,t-s}^1 \in D_1^{(2)} \right) h_{t-s}^2 \right]^{1+\delta} \right)^{1/(1+\delta)} de \cdot \left(EI_{t-s}(D_2) \right)^{\delta/(1+\delta)}$$
$$\leq \mu_{1,s}(D_1 \cup D_2) \left[\mu_2(D_1 \cup D_2) \right]^{\delta/(1+\delta)}$$

with $0 < \delta < 1$, where

$$\mu_{1,s}(D_1 \cup D_2) = \int_{D_1^{(1)}} \left(E\left[\sigma^2 (e, Z_{t-s}) I\left(X_t \setminus V_{t,t-s}^1 \in D_1^{(2)} \right) h_{t-s}^2 \right]^{1+\delta} \right)^{1/(1+\delta)} de$$

and

$$\mu_2(D_1 \cup D_2) = EI_{t-s}(D_1 \cup D_2) .$$

Hence, the Kolmogorov-Chentsov criterion is satisfied.

References

Amemiya, T. (1974), "The nonlinear two-stage least squares estimator", *Journal of Econometrics*, 2, 105-110.

Amemiya, T. (1977), "The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonllinear simultaneous equations model, *Econometrica* 45, 955-968.

Berndt, E.R., B.H. Hall, R.E. Hall and J.A. Hausman (1974), "Estimation and inference in nonlinear structural models", *Annals of Economic Social Measurement* 3, 653-666.

Bickel, P.J. and Wichura, M.J. (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*, Wiley and Sons, New York.

Brunk, H. D. (1970), "Estimation for isotonic regression," in *Nonparametric Techniques in Statistical Inference*, Ed. M.L. Puri, pp. 177-197, Cambridge: Cambridge University Press.

Burguete, J.F., A.R. Gallant and G. Souza (1982), "On the unification of the asymptotic theory of nonlinear econometric models", *Econometric Reviews* 1, 151-190.

Carrasco, M. and Florens J.-P. (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.

Chentsov, N.N. (1956), "Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov-Smirnov tests", *Theory of Probability and its Applications*, 1, 140-144.

Donald, S.G., G.W. Imbens and W.K. Newey (2001), "Empirical Likelihood estimation and consistent tests with conditional moment restrictions, preprint, MIT.

Domínguez, M.A. and Lobato I.N. (2002), "Testing the martingale difference hypothesis", Econometric Reviews, forthcoming.

Hall, P. and Heyde, C.C. (1980), *Martingale Limit Theory and its Application*, Academic Press, New York.

Hansen (1982), "Large-sample properties of generalized method of moments estimators", *Econometrica* 50, 1029-1054.

Hansen, L.P. (1985), "A method for calculating bounds in the asymptotic covariance matrices of generalized method of moments estimators", *Journal of Econometrics*, 30, 203-238.

Jorgenson, D.W. and J. Laffont (1974), "Efficient estimation of nonlinear simultaneous equations with additive disturbances", *Annals of Economic Social Measurement* 3, 615-640.

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

Newey, W. (1990a), "Efficient instrumental variables estimation of nonlinear models", *Econometrica* 58, 809-837.

Newey, W. (1990b), "Efficient estimation of models with conditional moment restrictions", in G:S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics*, Volume 11: Econometrics. Amsterdam: North-Holland.

Straf, M.L. (1970), "Weak convergence of stochastic processes with several parameters", Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, 187-221.