ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 74 – 2004

Closest Moment Estimation under General Conditions

Chirok HAN *, Robert DE JONG **

ABSTRACT. – This paper considers Closest Moment (CM) estimation with a general distance function, and avoids the assumption of nonsingular quadratic local behavior. The results of MANSKI [1983], NEWEY [1988], PÖTSCHER and PRUCHA [1997], and DE JONG and HAN [2002] are obtained as special cases. Consistency and a root-n rate of convergence are obtained under mild conditions on the distance function and on the moment conditions. We derive the limit distribution of CM estimators in a general setting, and show that the limit distribution is not necessarily normal. Asymptotic normality is obtained as a special case when the distance function displays nonsingular quadratic behavior.

La méthode d'estimation plus proche sous les conditions générales

RÉSUMÉ. – This paper considers Closest Moment (CM) estimation with a general distance function, and avoids the assumption of nonsingular quadratic local behavior. The results of Manski (1983), Newey (1988), Pötscher and Prucha (1997), and de Jong and Han (2002) are obtained as special cases. Consistency and a root-n rate of convergence are obtained under mild conditions on the distance function and on the moment conditions. We derive the limit distribution of CM estimators in a general setting, and show that the limit distribution is not necessarily normal. Asymptotic normality is obtained as a special case when the distance function displays nonsingular quadratic behavior.

---

* School of Economics & Finance, Victoria University of Wellington, P.O Box 600, Wellington, New Zealand, email chirok.han@vuw.ac.nz

** Department of Economics, Ohio State University, 429 Arps Hall, 1945 N. High Street, Columbus OH 43210, U.S.A, email dejong@econ.ohio-state.edu

1 Introduction

When economic information is given in the form of moment restrictions, the Generalized Method of Moments (GMM) formulated by HANSEN [1982] is a convenient way to directly exploit the moment conditions and estimate parameters. When the number of moment restrictions is equal to the number of parameters to be estimated, the method estimates the parameters by equating the empirical moments to zero. But when there are more moment conditions than there are parameters, it is generally impossible to set the empirical moments to zero, and in that case we want to minimize a distance[1] between the moment vector and zero. The usual GMM estimator minimizes a quadratic distance measure.

There are only a few papers dealing with the asymptotic distribution of CM estimators using a distance measure other than a quadratic one. This may be partly because the quadratic distance function is "natural" as a distance measure, partly because GMM has well developed asymptotics, and partly because the optimal GMM estimator attains the semiparametric efficiency bound, as is shown in CHAMBERLAIN [1987]. Nevertheless, the question of what happens if other distance measures are used still has its own source of interest. MANSKI [1983] considered the use of general distance functions and called the estimation technique Closest Moment (CM) estimation. He assumed the existence and nonsingularity of a second derivative matrix for distance functions and derived asymptotic normality of the CM estimators. Based upon these results, NEWEY [1988] showed that under the same assumptions on the distance function, the CM estimator is asymptotically equivalent to the GMM estimator using the second derivative matrix (evaluated at 0) as weight. ANDREWS [1994] established root-n consistency and asymptotic normality with greater generality for what he called "MINPIN" estimators, and ANDREWS' results can be applied to CM estimation. But ANDREWS also assumes local twice differentiability and nonsingularity of the Hessian for the distance function, like MANSKI and NEWEY; see condition (h) of "Assumption N" of ANDREWS [1994]. PÖTSCHER and PRUCHA [1997]'s derivation of asymptotic normality for GMM estimators also relies on the nonsingularity of the Hessian matrix of the the distance function; see condition (c) of Assumption 11.7 of PÖTSCHER and PRUCHA [1997].

Though the regularity condition of a nonsingular local quadratic behavior for distance functions leads to asymptotically normal estimators, it is in fact quite restrictive, and nontrivially limits the class of applicable distance functions. For example, as is mentioned in DE JONG and HAN [2002], among the class of L_p distances, only the usual quadratic distance ($p = 2$) satisfies the regularity conditions, and interesting cases such as $p = 1$ and $p = \infty$ cannot be analyzed by the above method. The same thing is true for more complex,

1. In this paper, a "distance" does not necessarily mean a "metric" or a "norm." Though "discrepancy" might be a better choice, we use the word "distance" because the meaning is clearer and it would not give confusion.

interesting distance functions such as $||x||_1 + ||x||_2^2$, $||x||_1 + ||x||_\infty$, $\sum_{j=1}^{q} \log(1 + |x_j|)$, and $|x_1| + x_2^2 + \ldots + x_q^2$, for example, where $x = (x_1, \ldots, x_q)$ and $||\cdot||_p$ is the usual L_p distance.

Recently, DE JONG and HAN [2002] took a different approach towards the asymptotics of a special case of CM estimation. They analyzed the asymptotics of CM estimators using general L_p distances (which they named "L_p-GMM" estimators) and obtained a root-n rate of convergence, but asymptotic non-normality for $p \neq 2$. Their analysis does not give an explicit form for the asymptotic distribution, but presents it in an abstract form using the "argmin" functional on a Gaussian process.

In this paper, we will stretch the arguments in DE JONG and HAN [2002] to their outer limit. For a far more general class of distance functions, root-n consistency for CM estimators will be established, and their asymptotic distributions will be expressed as the argmin functional on a Gaussian stochastic process. The results in this paper will encompass both the traditional asymptotics found in MANSKI [1983], NEWEY [1988] and PÖTSCHER and PRUCHA [1997] and DE JONG and HAN [2002]'s new L_p-GMM asymptotics as special cases. The goal of this paper therefore is not to suggest a technique that is readily used empirically, but instead the intention is to provide a new piece of estimator theory for nonstandard forms of CM estimators.

In what follows, section 2 presents the main result of this paper with some examples. The concluding section is followed by a Mathematical Appendix in which all the proofs are gathered.

2 Main Theorem

Let $y_1, y_2, \ldots$ be a sequence of observable random vectors. Let $g(y_i, \theta)$ be the set of q moment restrictions with parameters $\theta \in \Theta \subset \mathbb{R}^p$ satisfying

$$Eg(y_i, \theta_0) = 0 \text{ for all } i. \tag{1}$$

Let $\overline{g}(\theta) = n^{-1} \sum_{i=1}^{n} g(y_i, \theta)$. The CM estimator $\widehat{\theta}$ is assumed to minimize the criterion function $\delta(\overline{g}(\theta))$, *i.e.* to satisfy

$$\delta(\overline{g}(\hat{\theta})) = \inf_{\theta \in \Theta} \delta(\overline{g}(\theta)) \tag{2}$$

where δ is a nonnegative real function on $\mathbb{R}^q$.

Now, as the first step, we make the following assumptions on the $(q \times 1)$ continuous moment function $g(\cdot)$. Let $\overline{D}(\theta) = \partial \overline{g}(\theta)/\partial \theta'$.

ASSUMPTION 2.1 (MOMENT CONDITIONS):

(C1) Θ *is a compact subset of* $\mathbb{R}^p$*;*

(C2) $\overline{g}(\theta)$ *converges in probability to a nonrandom function* $\gamma(\theta)$ *uniformly on* Θ;

(C3) $\gamma(\theta) = 0$ *if and only if* $\theta = \theta_0$ *where* θ_0 *is an interior point of* Θ;

(C4) $\overline{D}(\theta)$ *exists and converges in probability to a continuous function* $D(\theta)$ *uniformly in a neighborhood of* θ_0, *and* $D(\theta_0)$ *has full column rank;*

(C5) $n^{1/2}\overline{g}(\theta_0) \xrightarrow{d} N(0,\Omega)$.

The $\gamma(\theta)$ function defined in (C2) is usually $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Eg(y_i,\theta)$, and when y_i is stationary, it is equal to $Eg(y_i,\theta)$. The uniform convergence in probability is equivalent to pointwise convergence in probability and stochastic equicontinuity of $\overline{g}(\theta)$. Sufficient conditions can be found in DAVIDSON [2000]. Condition (C4) is also equivalent to pointwise convergence in probability and stochastic equicontinuity of $\overline{D}(\theta)$ in a neighborhood of θ_0. Condition (C5) can be regarded as the result of a central limit theorem.

The above conditions are standard, but the differentiability of $\overline{g}(\cdot)$ restricts unnecessarily the applicability of our main theorem. For example, the moment conditions involved with the "median" function do not usually satisfy them. We could circumvent this restrictedness by rewriting (C4) in terms of stochastic differentiability (*e.g.*, POLLARD [1985]) and a Taylor series expansion. However, we will not pursue that issue here, in order to keep us focused solely on distance functions. Note also that the $\overline{g}(\cdot)$ function does not necessarily have to be a sample moment function but can be a more general function, as long as the conditions in Assumption 2.1 are satisfied.

The following conditions are assumed to hold for the distance function $\delta(\cdot)$.

ASSUMPTION 2.2 (DISTANCE FUNCTION):

(D1) $\delta(\cdot)$ *is continuous on* $\mathbb{R}^q$;

(D2) $\delta(x) = 0$ *if and only if* $x = 0$;

(D3) $\delta(x) = \delta(-x)$;

(D4) $\delta(\cdot)$ *satisfies the triangle inequality up to a finite constant in a neighborhood of 0, i.e., there exist an* $\varepsilon > 0$ *and an* $M < \infty$ *such that if* $|x| < \varepsilon$ *and* $|y| < \varepsilon$, *then* $\delta(x+y) \leqslant M[\delta(x) + \delta(y)]$.

Conditions (D1) and (D2) are essential for the consistency of $\hat{\theta}$. The symmetry condition (D3) is reasonable because we do not want to get different estimates by changing $g(y_i,\theta)$ to $-g(y_i,\theta)$. Condition (D4) restricts $\delta(\cdot)$ so that it does not vary vigorously around 0, and is satisfied for a wide class of functions. If $\delta(\cdot)$ happens to be a norm, conditions (D2), (D3) and (D4) are automatically satisfied.

As will be explained later in our main theorem, we will first establish consistency for CM estimators. When an estimator is consistent, it will asymptotically be concentrated on an arbitrarily small neighborhood of the true parameter, and thus $\overline{g}(\hat{\theta})$ is asymptotically close to 0, due to the uniform convergence of $\overline{g}(\theta)$. So the local behavior of $\delta(\cdot)$ around 0 naturally plays a key role in determining the asymptotic distribution of the estimator. The key

quantity in our analysis is the sequence of *localized distance* functions $d_n(\cdot)$ defined as

$$d_n(x) = \frac{\delta(n^{-1/2}x)}{\delta(n^{-1/2}1)/\delta(1)} \quad \text{for } n = 1,2,\ldots \tag{3}$$

The $d_n(\cdot)$ function can be interpreted as a means to "zoom in" on $\delta(\cdot)$ at the origin. The factor $n^{-1/2}$ (the speed at which we zoom in) is related to the root-n rate of convergence for $\hat{\theta}$. The 1 appearing in the denominator is the vector of ones, and is chosen only for normalization.

The next set of assumptions puts restrictions on the behavior of the localized distance functions $d_n(\cdot)$. For a $(r \times 1)$ vector $x = (x_1,\ldots,x_r)$, let $|x| = \max_j |x_j|$. Following convention, let $D = D(\theta_0)$.

ASSUMPTION 2.3 (LOCALIZED DISTANCE):

(E1) *There exist a finite integer N and a nonnegative real function $\phi(\cdot)$ on $\mathbb{R}^q$ such that $\phi(x) \leqslant \inf_{n \geqslant N} d_n(x)$, and*

$$\phi(x) \longrightarrow \infty \text{ if } |Ax| \longrightarrow \infty \tag{4}$$

for some $(q \times q)$ matrix A such that AD has full column rank;

(E2) *$d_n(\cdot)$ converges uniformly on every compact subset of $\mathbb{R}^q$ to a continuous function $d(\cdot)$;*

(E3) *For $d(\cdot)$ defined in (E2), $d(z + Bt)$ achieves its minimum at a unique point of $t \in \mathbb{R}^p$ for each $z \in \mathbb{R}^q$ and for any $(q \times p)$ matrix B with full column rank.*

Condition (E1) is easier to understand if we choose $A = I$ and replace (4) and its subsequent sentence with a stronger condition that

$$\phi(x) \longrightarrow \infty \text{ if } |x| \longrightarrow \infty. \tag{5}$$

With the stronger condition (5) in place of (4), (E1) imposes that the local behavior of $\delta(\cdot)$ is such that the minimum of $\delta(x)$ is attained clearly at 0. The weaker condition (4) has been introduced in order to make our main theorem applicable to the cases in which some coordinates of $\mathbb{R}^q$ disappear in $\phi(\cdot)$ and $d(\cdot)$. Example 2.4 provides a nice illustration. Condition (E2) combined with (3) is key to bridge the gap between the central limit theorem for $n^{1/2}\overline{g}(\theta_0)$ and the limit distribution of the estimator. Condition (E2) looks very plausible for most functions, but we can construct a simple counterexample such as $\delta(x) = x^2[1 + \sin^2(1/x^2)]$ with $\delta(0) = 0$. Note that conditions (E1) and (E2) are automatically satisfied for $\phi(x) = d(x) = \delta(x)$ and $A = I$ if $\delta(\cdot)$ is a norm. Condition (E3) is, as is mentioned in DE JONG and HAN [2002], far from being innocent. For more on this point, the reader is referred to DE JONG and HAN [2002].

Below is an example that should enhance understanding.

EXAMPLE 2.4: Suppose the distance function is

$$\delta(x) = \frac{1}{2}\left(|x_1| + x_2^2\right),\ x = (x_1, x_2) \tag{6}$$

implying that the econometrician wants to minimize the absolute value of the first sample moment plus the square of the second sample moment (when there is a single parameter to estimate). Clearly (D4) holds for any $\varepsilon > 0$ because $(a+b)^2 \leqslant 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$. The corresponding localized distance is

$$d_n(x) = \frac{\frac{1}{2}(n^{-1/2}|x_1| + n^{-1}x_2^2)}{\frac{1}{2}(n^{-1/2} + n^{-1})} = \frac{1}{1+n^{-1/2}}\,|x_1| + \frac{1}{n^{1/2}+1}\,x_2^2 \tag{7}$$

This $d_n(\cdot)$ is bounded uniformly over n from below by

$$\phi(x) = \frac{1}{2}|x_1|, \tag{8}$$

which satisfies (E1) with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{9}$$

as long as $D_1 \neq 0$ where D_1 is the first element of D. And $d_n(x)$ converges uniformly on every compact subset of $\mathbb{R}^2$ to

$$d(x) = |x_1| \tag{10}$$

which satisfies (E3).

Finally, note that any strictly monotonic, continuous transformation of $\delta(\cdot)$ does not affect the minimization, and therefore the above assumptions in fact mean that there exists a strictly monotonic, continuous transformation of $\delta(\cdot)$ such that the transformed function satisfies the specified conditions.

Our main theorem is the following.

THEOREM 2.5: *Under Assumptions 2.1, 2.2 and 2.3, the estimator $\hat{\theta}$ converges in probability to θ_0. Furthermore, we have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{t \in \mathbb{R}^p} d(\zeta + Dt) \text{ where } \zeta \sim N(0, \Omega). \tag{11}$$

The proof is provided in Appendix.

EXAMPLE 2.6: Take the case of MANSKI [1983], NEWEY [1988], and PÖTSCHER and Prucha [1997] as an example. Suppose that $\delta(0) = 0$,

$\partial\delta(0)/\partial x' = 0$, and $\delta(x)$ has a continuous second derivative which is nonsingular when evaluated at 0. Consistency directly follows Theorem 2.5. A Taylor expansion for δ around 0 implies that $\delta(x) = \frac{1}{2}\, x'H(\tilde{x})x$ where $H(x) = \partial^2\delta(x)/\partial x\partial x'$ and $\tilde{x}$ is in between 0 and x. Thus,

$$d_n(x) = c_n x'H(\tilde{x}_n)x \quad \text{with } c_n = \delta(1)/1'H(a_n)1 \tag{12}$$

where a_n is in between 0 and $n^{-1/2}1$ and $\tilde{x}_n$ lies in between 0 and $n^{-1/2}x$. Denote $H = H(0)$. Then we have $c_n \longrightarrow c$ as $n \longrightarrow \infty$ where $c = \delta(1)/1'H1 > 0$ and $d_n(x) \longrightarrow d(x) = cx'Hx$ as $n \longrightarrow \infty$ where the convergence is uniform on every compact set in $\mathbb{R}^q$. Eventually for n large enough, $c_n \geqslant \frac{1}{2}c$ and $x'H(\tilde{x}_n) \geqslant \frac{1}{2}\lambda_{\min}x'x$, where $\lambda_{\min}$ is the smallest eigenvalue (which is positive) of H, and therefore condition (E1) is satisfied for $\phi(x) = \frac{1}{4}c\lambda_{\min}x'x$ and $A = I$. Conditions (E2) and (E3) are satisfied for $d(x) = cx'Hx$. Finally $d(\zeta + Dt)$ is minimized by $t^* = -(D'HD)^{-1}D'H\zeta$ (using notations used in Theorem 2.5), which has the limit distribution of the GMM estimator (after centering and rescaling) using H as weight.

EXAMPLES 2.7: Below are two examples taken from DE JONG and HAN [2002].

(i) The asymptotics of DE JONG and HAN [2002] for L_p-GMM ($p \in [1; \infty)$) can be obtained with no further analysis, since for the L_p distance, we have $d(x) = d_n(x) = \delta(x) = ||x||_p$, and condition (E1) is satisfied for $\phi(x) = \delta(x)$ and $A = I$. For these choises of $d_n(\cdot)$, $d(\cdot)$, $\phi(\cdot)$ and A, the results in DE JONG and HAN [2002] are obtained as a special case.

(ii) Consider the L_∞-GMM estimator corresponding to $\delta(x) = ||x||_\infty = \max_{1\leqslant j\leqslant q} |x_j|$. Then clearly $\phi(x) = d(x) = d_n(x) = \delta(x) = ||x||_\infty$ with $A = I$, and we have the asymptotic distribution

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{t\in\mathbb{R}^p} ||\zeta + Dt||_\infty. \tag{13}$$

EXAMPLES 2.8: Some more examples for other complicated distance functions are following.

(i) Let $\delta(x) = ||x||_1 + ||x||_\infty$. We have $\phi(x) = d(x) = d_n(x) = \delta(x)$ with $A = I$, and therefore,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{t\in\mathbb{R}^q}(||\zeta + Dt||_1 + ||\zeta + Dt||_\infty). \tag{14}$$

The asymptotic distribution is different from both the L_1-GMM limit distribution and the L_∞-GMM limit distribution.

(ii) Let $\delta(x) = \frac{1}{2}(||x||_1 + ||x||_2^2)$. We have $\phi(x) = \frac{1}{2}||x||_1$ with $A = I$ and $d(x) = ||x||_1$. Therefore, the asymptotic distribution is the same as that of the L_1-GMM estimator. Note that the limit distribution is different from $\underset{t\in\mathbb{R}^q}{\operatorname{argmin}}\, \delta(\zeta + Dt)$.

(iii) Consider $\delta(x) = \frac{1}{2}(|x_1| + x_2^2)$, as in Example 2.4, with a single parameter to estimate. We obtained that $d(x) = |x_1|$ in (10), which implies that as long as $D_1 \neq 0$, the estimator has an asymptotic distribution which is identical to that of the method of moments estimator using the first moment condition only, and therefore, it is asymptotically normally distributed. Note that when there are two parameters to estimate, conditions (E1) and (E3) are violated.

Now, let us consider the matter of weighting. In the context of usual GMM, a weighted GMM estimator can be regarded as an unweighted GMM estimator that uses the transformed moment conditions $EW\, g(y_i,\theta_0) = 0$, where W is a $(q \times q)$ matrix such that $W'W$ is equal to the weight. Extending this way of thinking to general CM estimation, we can define a "weighted" CM estimator $\hat{\theta}$ as the minimizer of $\delta\big(W\overline{g}(\theta)\big)$, where W is a $(q \times q)$ nonsingular matrix. Then, since $EW\, g(y_i,\theta) = 0$ is also a set of correct moment conditions, and since all the conditions in Assumption 2.1 are satisfied for $W\, g(y_i,\theta)$, the results of Theorem 2.5 hold, but Ω and D should be replaced with $W\Omega W'$ and WD, respectively. That is, we have the asymptotic distribution

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \underset{t\in\mathbb{R}^p}{\operatorname{argmin}}\, d(\zeta^* + D^*t) \tag{15}$$

where $\zeta^* \sim N(0, W\Omega W')$ and $D^* = WD$.

When a consistent estimate $\hat{W}_n$ of W is used in place of W, and therefore when $\hat{\theta}$ minimizes $\delta\big(\hat{W}_n\overline{g}(\theta)\big)$, the conditions of Assumption 2.1 are still satisfied for $\hat{W}_n\overline{g}(\theta)$ and $\hat{W}_n\overline{D}(\theta)$: For (C2), we have $\hat{W}_n\overline{g}(\cdot) \xrightarrow{p} W\gamma(\cdot)$ uniformly; for (C3), $W\gamma(\theta) = 0$ if and only if $\theta = \theta_0$ because W is nonsingular; for (C4), $\hat{W}_n\overline{D}(\theta_n) \xrightarrow{p} WD$ if $\theta_n \xrightarrow{p} \theta_0$, where WD has full column rank due to the nonsingularity of W; finally, for (C5), we have $n^{1/2}\hat{W}_n\overline{g}(\theta_0) \xrightarrow{d} N(0, W\Omega W')$. Therefore, in this case too we have the asymptotic distribution (15).

3 Conclusion

In this paper, we derived an abstract expression for the limit distribution of estimators which minimize an arbitrary distance function between population moments and sample moments, without the restriction of a nonsingular second derivative of the distance function evaluated at 0. MANSKI [1983], NEWEY [1988] and PÖTSCHER and PRUCHA [1997]'s traditional asymptotics of root-n consistency and asymptotic normality, as well as DE JONG and HAN [2002]'s asymptotics are produced as special cases.

A Mathematical appendix

In this section we prove Theorem 2.5. Consistency is easily established using standard technique from the uniform convergence of the criterion function, compactness of the parameter space, and the uniqueness (inside the interior of the parameter space) of the minimizer of the limit criterion function.

THEOREM A.1: *Under conditions (C1), (C2), (C3), (D1), and (D2),* $\hat{\theta} \xrightarrow{p} \theta_0$.

PROOF: See Theorem 9.3.1 of Davidson (2000).

To prove the convergence in distribution asserted in Theorem 2.5, we will apply a continuous mapping theorem to the argmin functional. But as is well known, the argmin functional is not continuous in general, and more restrictions should be imposed to the limit criterion function to make it continuous. (For more information, see VAN DER VAART and WELLNER [1996], Section 3.2.) A set of conditions that is useful in our case is found in Theorem 2.7 of Kim and Pollard (1990). More specifically, the theorem states that if

(i) a sequence of random processes $C_n(t)$ on $\mathbb{R}^p$ converges weakly on every compact set to a stochastic process $C(t)$ in a separable subset of locally bounded functions such that, for almost all sample path, (a) $C(\cdot)$ is continuous, (b) $C(\cdot)$ achieves its minimum at a unique point in $\mathbb{R}^p$, and (c) $C(t) \longrightarrow \infty$ as $|t| \longrightarrow \infty$; and

(ii) the minimizers $\hat{t}_n$ of $C_n(t)$ are $O_p(1)$,

then $\hat{t}_n$ converges in distribution to the minimizer of $C(t)$. (Note that the theorem as such is more complicated to handle possible non-measurability of the argmin estimators.)

To apply this result, define the stochastic processes $h_n(t)$ on $\mathbb{R}^p$ as

$$h_n(t) = \begin{cases} n^{1/2}\overline{g}(\theta_0 + tn^{-1/2}), & \text{if } \theta_0 + tn^{-1/2} \in \Theta \\ n^{1/2}\overline{g}(\theta_n^0), & \text{otherwise} \end{cases} \tag{16}$$

where θ_n^0 maximizes $\delta\big(\overline{g}(\theta)\big)$ over Θ for $n = 1,2,\ldots$ Noting that the minimizer $\hat{t}_n$ of $\delta\big(h_n(t)\big)$ is equal to $n^{1/2}(\hat{\theta} - \theta_0)$, Lemma A.3 will first establish root-n rate of convergence for $\hat{\theta}$. Lemmas A.4 and A.5 will show the weak convergence of $h_n(t)$ to the stochastic process $h(t)$ defined on $\mathbb{R}^p$ as

$$h(t) = \zeta + DT \text{ where } \zeta \sim N(0,\Omega) \tag{17}$$

and Lemma A.6 will establish weak convergence of $d_n\big(h_n(\cdot)\big)$. And finally all the facts and results are assembled to prove our main theorem.

To begin with, the next theorem establishes that $\hat{t}_n = n^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$. But first let us introduce a simple lemma tailored for our case.

LEMMA A.2: *Suppose that a sequence of fuctions $d_n(\cdot)$ on $\mathbb{R}^q$ converges to $d(\cdot)$ uniformly on every compact subset of $\mathbb{R}$. Let $\{x_n\}$ be a sequence of stochastically bounded random elements, i.e., $x_n = O_p(1)$. Then $d_n(x) = d(x_n) + o_p(1)$.*

PROOF: We need to prove that for every $\varepsilon > 0$ and $\eta > 0$, there exists a finite number n_0 such that if $n > n_0$ then $P\{|d_n(x_n) - d(x_n)| > \varepsilon\} < \eta$. To prove it, first choose M and n_1 such that $P\{|x_n| > M\} < \eta$ if $n > n_1$. Then M and n_1 are finite because $x_n = O_p(1)$. Next, for that M, choose n_2 such that $\sup_{|x| \leqslant M} |d_n(x) - d(x)| < \frac{1}{2}\varepsilon$ if $n > n_2$. Then n_2 is also finite because of the first supposition of the lemma. Let $n_0 = \max\{n_1,n_2\}$. Then for $n > n_0$,

$$\begin{aligned} P\{|d_n(x_n) - d(x_n)| > \varepsilon\} &= P\{|d_n(x_n) - d(x_n)| > \varepsilon,\ |x_n| \leqslant M\} \\ &\quad + P\{|d_n(x_n) - d(x_n)| > \varepsilon,\ |x_n| > M\} \\ &\leqslant 0 + P\{|x_n| > M\} < \eta, \end{aligned} \tag{18}$$

which gives the conclusion.

Another simple fact is that the differentiability of $\overline{g}(\theta)$ implies a Taylor expansion of $\overline{g}(\theta)$ at θ_0

$$\overline{g}(\theta) = \overline{g}(\theta_0) + \overline{D}(\theta^*)(\theta - \theta_0) \tag{19}$$

for a θ^* (which is stochastic) lying on the line segment between θ_0 and θ. Note that θ^* can lie outside Θ, but because θ_0 is an interior point of Θ, there exists a neighborhood $V \subset \Theta$ of θ_0 such that (19) is satisfied with $\theta^* \in \Theta$ if $\theta \in V$.

LEMMA A.3: *Under (C1)-(C5), (D1)-(D4), and (E1)-(E2),* $n^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$.

PROOF: By (19), we have

$$\delta\big(\overline{g}(\hat{\theta}) - \overline{g}(\theta)\big) = \delta\big(\overline{D}(\tilde{\theta})(\hat{\theta} - \theta)\big) \tag{20}$$

for some $\tilde{\theta}$ in between $\hat{\theta}$ and θ_0. Because $\overline{g}(\cdot)$ converges uniformly to $\gamma(\cdot)$ and $\hat{\theta}$ is consistent for θ_0, both $\overline{g}(\hat{\theta})$ and $\overline{g}(\theta_0)$ converge to 0, and thus both become small enough to satisfy (D4) as n increases. When this happens, we have

$$\delta\big(\overline{g}(\hat{\theta}) - \overline{g}(\theta_0)\big) \leqslant M\left[\delta\big(\overline{g}(\hat{\theta})\big) + \delta\big(\overline{g}(\theta_0)\big)\right] \leqslant 2M\delta\big(\overline{g}(\theta_0)\big) \tag{21}$$

for some $M < \infty$, where the first inequality comes from (D4) and the second inequality is obtained from the definition of $\hat{\theta}$ as minimizer. Now, dividing (20) and (21) by $\delta(n^{-1/2}1)/\delta(1)$, we get

$$d_n\big(\overline{D}(\tilde{\theta})n^{1/2}(\hat{\theta} - \theta_0)\big) \leqslant 2Md_n\big(n^{1/2}\overline{g}(\theta_0)\big) \tag{22}$$

The left hand side of (22) is bounded from below by $\varphi\big(\overline{D}(\tilde{\theta})n^{1/2}(\hat{\theta} - \theta_0)\big)$ for $\phi(\cdot)$ possessing the properties described in (E1), and the right hand side is $2Md\big(n^{1/2}\overline{g}(\theta_0)\big) + o_p(1)$ under (E2) by Lemma A.2. This last term is $O_p(1)$ because of (C5) and the continuity of $d(\cdot)$. Therefore, $\phi\big(\overline{D}(\tilde{\theta})n^{1/2}(\hat{\theta} - \theta_0)\big)$ is also bounded by an $O_p(1)$ sequence.
To conclude, this last result and the second part of (E1) imply that $A\overline{D}(\tilde{\theta})\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ for some A such that AD has full column rank. The desired result now follows because $\overline{D}(\tilde{\theta}) \xrightarrow{p} D$ by (C4) and the consistency of $\tilde{\theta}$, and because AD has full column rank.

The next two lemmas will prove that $h_n(\cdot)$ converges weakly to $h(\cdot)$ on every compact set $K \subset \mathbb{R}^p$ by showing that the finite dimensional distributions of $h_n(\cdot)$ converge to those of $h(\cdot)$ (Lemma A.4) and that $h_n(\cdot)$ is stochastically equicontinous (Lemma A.5).

LEMMA A.4: *Let $h_n(\cdot)$ and $h(\cdot)$ be defined by (16) and (17). Under assumptions (C3), (C4) and (C5), the finite dimensional distributions of $h_n(\cdot)$ converge to those of $h(\cdot)$, i.e., for any finite collection of $(t_1,\ldots,t_r)$, for any $r = 1,2,\ldots,$*

$$\big(h_n(t_1),\ldots,h_n(t_r)\big) \xrightarrow{d} \big(h(t_1),\ldots,h(t_r)\big) \tag{23}$$

PROOF: Fix $t \in \mathbb{R}^p$. Since θ_0 is an interior point of Θ by (C3), $\theta_0 + tn^{-1/2}$ eventually belongs to Θ. When this happens, by (19),

(24) $$h_n(t) = n^{1/2}\overline{g}(\theta_0) + \overline{D}(\theta_0 + \tilde{t}n^{-1/2})t,$$

with $\tilde{t}$ lying in between t and 0. So assumptions (C4) and (C5) imply that $h_n(t) \xrightarrow{d} \zeta + Dt = h(t)$. Finally, (23) follows from the CRAMÉR-WOLD device.

LEMMA A.5: *Under assumption (C4), the processes $h_n(.)$ defined by (16) are stochastically equicon-tinuous on every compact set $K \subset \mathbb{R}^p$.*

PROOF: Similarly to the above proof, $\theta_0 + tn^{-1/2}$ eventually falls into a neighborhood of θ_0 satisfying (C4) and (19) for all t uniformly on K. When it happens, by (19), we have

(25) $$h_n(t_1) - h_n(t_2) = \overline{D}(\theta_0 + \tilde{t}_1 n^{-1/2})t_1 - \overline{D}(\theta_0 + \tilde{t}_2 n^{-1/2})t_2$$

where $\tilde{t}_i$ lies in between t_i and 0 for $i = 1,2$ for all t_1 and t_2 on K. (Note that $\tilde{t}_i$'s do not necessarily belong to K.) Condition (C4) implies that the last expression converges in probability to $D \cdot (t_1 - t_2)$ uniformly over all t_1 and t_2 on K. Clearly $|D \cdot (t_1 - t_2)| \longrightarrow 0$ as $|t_1 - t_2| \longrightarrow 0$, and the stochastic equicontinuity of $h_n(\cdot)$ on K follows.

The above two lemmas mean that the sequence of stochastic processes $h_n(\cdot)$ converges weakly to the stochastic process $h(\cdot)$ on any compact subset K of $\mathbb{R}^p$. This result is conveyed to the sequence $d_n\big(h_n(\cdot)\big)$ in the next lemma as an application of a generalized version of the continuous mapping theorem by PROHOROV [1956].

LEMMA A.6: *Under the assumptions of Lemmas A.4 and A.5, and under assumption (E2), the stochastic processes C_n defined by*

(26) $$C_n(t) = d_n\big(h_n(t)\big) = d_n\big(n^{1/2}\overline{g}(\theta_0 + tn^{-1/2})\big)$$

converges weakly on every compact set K to the stochastic process C defined by $C(t) = d\big(h(t)\big) = d(\zeta + Dt)$.

PROOF: By Lemmas A.4 and A.5, we have the weak convergence of $h_n(t)$ to $h(t)$ on any given compact set $K \subset \mathbb{R}^p$. The functions $d_n(\cdot)$ are continuous on $\mathbb{R}^q$, and uniformly convergent on every compact set to the mapping $d(\cdot)$. Now apply Theorem 1.10 of PROHOROV [1956, p. 166].

PROOF OF THEOREM 2.5: First note that $C_n(t)$ is minimized by $\hat{t}_n = n^{1/2}(\hat{\theta} - \theta_0)$, which is $O_p(1)$ by Lemma A.3. By Lemma A.6, $C_n(t)$ converges weakly to $C(t)$ on every compact K. Clearly, $C(t)$ lives in a separable set of functions such that every sample path $t \longmapsto C(t)$ is locally bounded and continuous and diverges to ∞ if $|t| \longrightarrow \infty$ because of condition (E1). And finally every sample path $t \longmapsto C(t)$ achieves its minimum at a unique point by (E3). Now we can apply Theorem 2.7 of KIM and POLLARD [1990]. ▼

• References

ANDREWS D. W. K. (1994). – Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity, *Econometrica*, 62 (1), p. 43-72.

CHAMBERLAIN G. (1987). – Asymptotic Efficiency in Estimation with Conditional Moment Restric-tions, *Journal of Econometrics*, 34, p. 305-334.

DAVIDSON J. (2000). – *Econometric Theory*, Blackwell.

DE JONG R. M., HAN C. (2002). – The Properties of Lp-GMM Estimators, *Econometric Theory*, 18, p. 491-504.

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

KIM J., POLLARD D. (1990). – Cube Root Asymptotics, *The Annals of Statistics*, 18, p. 191-219.

MANSKI C. F. (1983). – Closest Empirical Distribution Estimation, *Econometrica*, 51 (2), p. 305-319.

NEWEY W. K. (1988). – Asymptotic Equivalence of Closest Moments and GMM estimators, *Econometric Theory*, 4, p. 336-340.

POLLARD D. (1985). – New Ways to Prove Central Limit Theorems, *Econometric Theory*, 1, p. 295-314.

PÖTSCHER B. M., PRUCHA I. R. (1997). – *Dynamic Nonlinear Econometric Models*, Springer.

PROHOROV Y. V. (1956). – Convergence of Random Processes and Limit Theorems in Probability Theory, *Theory of Probability and Its Applications*, 1 (2), p. 157-214.

VAN DER VAART A. W., WELLNER J. A. (1996). – *Weak Convergence and Empirical Processes*, Springer. 12