Closest Moment Estimation under General Conditions

Chirok Han and Robert de Jong*

January 28, 2002

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. Asymptotic normality is obtained as a special case when the distance function displays nonsingular quadratic behavior.

Keywords: GMM, generalized method of moments, closest moment, general distance function, asymptotics, root-n consistency, quadratic distance.

JEL Classfication Numbers: C00, C10, C13.

1 Introduction

WHEN ECONOMIC INFORMATION is given in the form of moment restriction, 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 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) gave a

*Chirok Han (corresponding author): School of Economics and Finance, Victoria University of Wellington, P.O.Box 600, Wellington, New Zealand, chirok.han@vuw.ac.nz; Robert de Jong: Department of Economics, Michigan State University, East Lansing, MI 48824, USA, dejongr@msu.edu.

[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.

direct treatment of 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, interesting distance functions such as $\|x\|_1 + \|x\|_2^2$, $\|x\|_1 + \|x\|_\infty$, $\sum_{j=1}^q \log(1 + |x_j|)$, and $|x_1| + x_2^2 + \cdots + x_q^2$, 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 its 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.

In what follows, section 2 presents the main result of this paper with some examples, section 3 proves the main theorem, and the last section contains concluding remarks.

2 Main Theorem

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

$$Eg(y_i, \theta_0) = 0 \quad \text{for all } i. \tag{2.1}$$

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

$$\delta[\bar{g}(\hat{\theta})] = \inf_{\theta \in \Theta} \delta[\bar{g}(\hat{\theta})], \tag{2.2}$$

where δ is a nonnegative real function on $\mathbb{R}^q$. When the existence (measurability) of $\hat{\theta}$ is at risk, we can change the above definition so that $\hat{\theta}$ *almost* minimizes the criterion function in the sense that $\hat{\theta}$ satisfies

$$\delta[\bar{g}(\hat{\theta})] \leq \inf_{\theta \in \Theta} \delta[\bar{g}(\hat{\theta})] + o_p(1). \tag{2.3}$$

A set of sufficient conditions for the existence of $\hat{\theta}$ satisfying (2.2) can be found in Lemma 2 of Jennrich (1969, p. 637), which requires the compactness of the parameter space, the measurability of the criterion function for each θ, and the continuity of the function in θ for each sample path. In this paper, we will assume that (2.2) is satisfied to make the exposition simpler and clearer, but extending the results to allow (2.3) would require only minor modification of the whole statements.

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

Assumption 2.1 (moment conditions).

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

(C2) $\bar{g}(\theta)$ *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) $\bar{D}(\theta)$ *exists and converges in probability to a nonrandom function* $D(\theta)$ *uniformly in a neighborhood of* θ_0*, and* $D(\theta_0)$ *has full column rank;*

(C5) $\bar{g}(\theta)$ *allows local Taylor expansion at* θ_0*, i.e., there exists* θ^* *in between* θ *and* θ_0 *such that*

$$\bar{g}(\theta) = \bar{g}(\theta_0) + \bar{D}(\theta^*)(\theta - \theta_0) \tag{2.4}$$

for θ *in a neighborhood of* θ_0*;*

(C6) $n^{1/2}\bar{g}(\theta_0) \xrightarrow{d} N(0, \Omega)$.

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 $\bar{g}(\theta)$. Sufficient conditions can be found in Davidson (2000). Similar remarks apply to condition (C4). A sufficient condition for (C5) is that the function $g(y, \theta)$ is twice continuously differentiable with respect to θ on Θ, and that Θ is convex under (C1). Condition (C6) can be regarded as the result of a central limit theorem.

The above conditions are standard, but the differentiability of $\bar{g}(\cdot)$ 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) and (C5) in terms of stochastic differentiability (see Pollard, 1985). However, we will not pursue that issue here, in order to keep us focused solely on distance functions.

The following conditions are assumed to hold for the distance function δ.

Assumption 2.2 (distance function).

(D1) $\delta(\cdot)$ *is continuous;*

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

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

(D4) δ *satisfies the triangular inequality up to a finite constant locally (in a neighborhood of 0), i.e., there exists* $\varepsilon > 0$ *such that if* $|x_1| < \varepsilon$ *and* $|x_2| < \varepsilon$*, then* $\delta(x_1 + x_2) \leq M[\delta(x_1) + \delta(x_2)]$ *for all* x_1 *and* x_2*, for some* $M < \infty$*.*

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 to $-g$. Condition (D4) is also satisfied for a wide class of functions. If δ 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 $\bar{g}(\hat{\theta})$ is asymptotically close to 0, due to the uniform convergence of $\bar{g}$. So the local behavior of δ around 0 naturally plays a key role in determining asymptotic distribution of the estimator. The key quantity in our analysis is the sequence of *localized distance* functions d_n corresponding to δ as

$$d_n(x) = \frac{\delta(n^{-1/2}x)}{\delta(n^{-1/2}1)} \text{ for } n = 1, 2, \ldots. \tag{2.5}$$

The $d_n(.)$ function can be interpreted as a means to "zoom in" on $\delta(.)$ at the origin. The factor $n^{-1/2}$ (or 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 of (2.5) 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 $d_n(.)$. For a $(r \times 1)$ vector $x = (x_1, \ldots, x_r)$, let $|x| = \max_j |x_j|$.

Assumption 2.3 (localized distance).

(E1) *There is a real function* $\phi(.)$ *on* $\mathbb{R}^q$ *such that* $\inf_n d_n(x) \geq \phi(x)$*, and* $\phi(x) \to \infty$ *if* $|x| \to \infty$*;*

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

(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) imposes that the local behavior of $\delta(.)$ is such that the model is well identified in infinitesimal neighborhoods of 0. Condition (E2) combined with (2.5) is key to bridge the gap between the central limit theorem for $n^{1/2}\bar{g}(\theta_0)$ and the limit distribution of the estimator. If $\delta(.)$ is a norm, this condition is automatically satisfied. (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 $\delta(0) = 0$. 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 are some examples that should enhance understanding.

Example 2.4. Suppose the distance function is

$$\delta(x) = |x_1| + x_2^2, \quad x = (x_1, x_2), \tag{2.6}$$

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). Then the corresponding localized distance is

$$d_n(x) = \frac{n^{-1/2}|x_1| + n^{-1}x_2^2}{n^{-1/2} + n^{-1}} = \frac{1}{1 + n^{-1/2}}|x_1| + \frac{1}{n^{1/2} + 1}x_2^2. \tag{2.7}$$

This d_n is bounded uniformly over n from below by

$$\phi(x) = |x_1|/2, \tag{2.8}$$

which satisfies (E1). And $d_n(x)$ converges uniformly on any compact subset of $\mathbb{R}^2$ to

$$d(x) = |x_1| \tag{2.9}$$

which satisfies (E3). □

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

Our main theorem is the following. Following convention, let $D = D(\theta_0)$.

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*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{t \in \mathbb{R}^p} d(\zeta + Dt) \text{ where } \zeta \sim N(0, \Omega). \tag{2.10}$$

The proof is provided in the next section.

In many cases, it is easier to define d_n as $d_n(x) = an^r\delta(n^{-1/2}x)$ for some constants $a \neq 0$ and r such that $an^r\delta(n^{-1/2}1) \to c$ with $0 < c < \infty$ than to define d_n as in (2.5). Indeed, for any such an^r used instead of $\delta(n^{-1/2}1)$, the results in Theorem 2.5 hold. The following examples use this simplified method to find the corresponding "d" functions.

Example 2.6. 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 $\delta(x) = \frac{1}{2}x'W(\tilde{x})x$ where $W(x) = \partial^2\delta(x)/\partial x\partial x'$ and $\tilde{x}$ is in between 0 and x. Thus, (using $a = 2$ and $r = 1$)

$$2n\delta(n^{-1/2}x) = x'Wx + x'[W(\tilde{x}_n) - W]x \tag{2.11}$$

where $W = W(0)$ and $\tilde{x}_n$ lies in between 0 and $n^{-1/2}x$. Clearly the second term disappears uniformly in any compact set containing 0. So condition (E2) is satisfied for $d(x) = x'Wx$ up to a finite positive constant. Condition (D4) is easy to check. And the quadratic form of d ensures that conditions (E1) and (E3) are satisfied. Finally $d(\zeta + Dt) = (\zeta + Dt)'W(\zeta + Dt)$ (using notations used in Theorem 2.5) is minimized by $t^* = -(D'WD)^{-1}D'W\zeta$, which has the distribution of GMM estimator (after centering and rescaling) using W as weighting matrix. This is the case of Manski (1983), Newey (1988), and Pötscher and Prucha (1997).

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 $d = \delta$ and all the conditions described in de Jong and Han (2002) are derived or assumed.

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{t \in \mathbb{R}^p} \|\zeta + Dt\|_\infty. \tag{2.12}$$

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

(*i*) Let $\delta(x) = \|x\|_1 + \|x\|_\infty$. We have $d = \delta$, and therefore, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \operatorname{argmin}_t \{\|\zeta + Dh\|_1 + \|\zeta + Dt\|_\infty\}$. The asymptotic distribution is different from both the L_1-GMM limit distribution and the L_∞-GMM limit distribution.

(*ii*) Let $\delta(x) = \|x\|_1 + \|x\|_2^2$. We have $d(x) = \|x\|_1$, and therefore the asymptotic distribution is the same as that of the L_1-GMM estimator. Note that the limiting distribution is different from $\operatorname{argmin}_h \{\|\zeta + Dt\|_1 + \|\zeta + Dt\|_2^2\}$.

(*iii*) For $\delta(x) = \sum_{j=1}^q \log(1+|x_j|)$, we have $n^{1/2} \log(1+|n^{-1/2} x_j|) \to |x_j|$, and $d(x) = \|x\|_1$. Therefore, the asymptotic distribution for the estimator with this distance function is the same as that of the L_1-GMM estimator.

(*iv*) Consider $\delta(x) = |x_1| + x_2^2$, as in Example 2.4. We have $d(x) = |x_1|$, and the estimator has the same asymptotic distribution as that of the method of moments estimator using the first moment condition only, and therefore, it is asymptotically normally distributed.

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 $EAg(y_i, \theta_0) = 0$, where A is a $(q \times q)$ matrix such that $A'A$ is equal to the weight. Extending this way of thinking to general CM estimation, we can define a "weighted" CM estimator as the minimizer of $\delta[A\bar{g}(\theta)]$, where A is a $(q \times q)$ matrix. Then, since $EAg(y_i, \theta) = 0$ is also a set of correct moment conditions, and since all the conditions in Assumption 2.1 are satisfied for $Ag(.,.)$, the results of Theorem 2.5 hold, but Ω and D should be replaced with with $A\Omega A'$ and AD, respectively. When a consistent estimate A_n of A is used in place of A, and therefore when $\hat{\theta}$ minimizes $\delta[A_n \bar{g}(\theta)]$, the conditions of Assumption 2.1 are still satisfied for $A_n \bar{g}(\theta)$ and $A_n \bar{D}(\theta)$. (Of course, the limiting quantities should change correpondingly.) Hence, using A_n instead of A does not affect the results.

3 Proof

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 limiting criterion function.

Theorem 3.1. *Under conditions (C1), (C2), (C3), (D1), and (D2),* $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. See Theorems 9.3.1 and 9.3.2 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 function is not continuous in general, and more restrictions should be imposed to the limit criterion function to make the argmin function continuous. (For more information, see Van der Vaart, 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 in distribution (in the sense of the weak convergence of probability measures) 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) $C(t) \to \infty$ as $|t| \to \infty$; 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 on $\mathbb{R}^p$ as

$$h_n(t) = \begin{cases} n^{1/2}\bar{g}(\theta_0 + tn^{-1/2}), & \text{if } \theta_o + tn^{-1/2} \in \Theta \\ n^{1/2}\bar{g}(\theta_n^0), & \text{otherwise} \end{cases} \tag{3.1}$$

where θ_n^0 maximizes $\delta[\bar{g}(\theta)]$ over Θ for $n = 1, 2, \ldots$. Noting that the minimizer $\hat{t}_n$ of $\delta[h_n(t)]$ is equal to $n^{1/2}(\hat{\theta} - \theta_0)$, Theorem 3.3 will first establish root-n rate of convergence for $\hat{t}_n$. The next two lemmas will show the convergence in distribution of h_n to the stochastic process h, defined on $\mathbb{R}^p$ as

$$h(t) = \zeta + Dt \text{ where } \zeta \sim N(0, \Omega), \tag{3.2}$$

on any compact, and Lemma 3.6 will establish weak convergence of $d_n(h_n(.))$. And finally all the facts and results are assembled to prove our main theorem.

To begin with, 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 3.2. *Suppose that a sequence of fuctions* d_n *on* $\mathbb{R}^q$ *converges to* d *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* $d_n(x_n) = d(x_n) + o_p(1)$.

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 $\sup_{|x| \le M} |d_n(x) - d(x)| < \frac{1}{2}\varepsilon$ 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 $n > n_0$,

$$\begin{aligned} P\{|d_n(x_n) - d(x_n)| > \varepsilon\} &= P\{|d_n(x_n) - d(x_n)| > \varepsilon, |x_n| \le M\} \\ &\quad + P\{|d_n(x_n) - d(x_n)| > \varepsilon, |x_n| > M\} \\ &\le 0 + P\{|x_n| > M\} < \eta, \end{aligned} \tag{3.3}$$

which gives the conclusion. □

Theorem 3.3. *Under (C1)–(C6), (D1)–(D4), and (E1)–(E2),* $n^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$.

Proof. By condition (C5), we have

$$\delta[\bar{g}(\hat{\theta}) - \bar{g}(\theta_0)] = \delta[\bar{D}(\tilde{\theta})(\hat{\theta} - \theta_0)] \tag{3.4}$$

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

$$\delta[\bar{g}(\hat{\theta}) - \bar{g}(\theta_0)] \leq M\{\delta[\bar{g}(\hat{\theta})] + \delta[\bar{g}(\theta_0)]\} \leq 2M\delta[\bar{g}(\theta_0)] \tag{3.5}$$

for some $M < \infty$, where the first inequality comes from (D4) and the second inequality is obtained from the definition of $\hat{\theta}$. Now, dividing (3.4) and (3.5) by $\delta(n^{-1/2}1)$, we get

$$d_n[\bar{D}(\tilde{\theta})n^{1/2}(\hat{\theta} - \theta_0)] \leq 2Md_n[n^{1/2}\bar{g}(\theta_0)] \tag{3.6}$$

The left hand side of (3.6) is bounded from below by $\phi[\bar{D}(\tilde{\theta})n^{1/2}(\hat{\theta} - \theta_0)]$ for ϕ possessing the properties described in (E1), and the right hand side is $2d[n^{1/2}\bar{g}(\theta_0)]M + o_p(1)$ under (E2) by Lemma 3.2. This last term is $O_p(1)$ because of (C6) and the continuity of $d(\cdot)$. Therefore, $\phi[\bar{D}(\tilde{\theta})n^{1/2}(\hat{\theta} - \theta_0)]$ is also bounded by an $O_p(1)$ sequence.

To conclude, this last result and the second part of (E1) imply that $\bar{D}(\tilde{\theta})\sqrt{n}(\hat{\theta}-\theta_0) = O_p(1)$. The desired result now follows because $\bar{D}(\tilde{\theta}) \xrightarrow{p} D$ by (C4) and the consistency of $\tilde{\theta}$, and because D has full column rank. □

The next two lemmas will prove that h_n converges in distribution to h on every compact set $K \subset \mathbb{R}^p$ by showing that the finite dimensional distributions of h_n converge to those of h (Lemma 3.4) and that h_n is stochastically equicontinous (Lemma 3.5).

Lemma 3.4. *Let* h_n *and* h *be defined by* (3.1) *and* (3.2)*. Under assumptions (C3), (C4), (C5) and (C6),* h_n *converges in finite-dimensional distribution to* h*, i.e., for any finite collection of* $(t_1, \ldots, t_r)$*, for any* $r = 1, 2, \ldots,$

$$[h_n(t_1), \ldots, h_n(t_r)] \xrightarrow{d} [h(t_1), \ldots, h(t_r)]. \tag{3.7}$$

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 assumption (C5),

$$h_n(t) = n^{1/2}\bar{g}(\theta_0) + \bar{D}(\theta_0 + \tilde{t}n^{-1/2})t, \tag{3.8}$$

with $\tilde{t}$ lying in between t and 0. So assumptions (C4) and (C6) imply that $h_n(t) \xrightarrow{d} \zeta + Dt = h(t)$. Finally, (3.7) follows from the Cramér-Wold device. □

Lemma 3.5. *Under assumptions (C4) and (C5), the processes* h_n *defined by* (3.1) *are stochastically equicontinuous on every compact set* $K \subset \mathbb{R}^p$.

Proof. By Taylor expansion (2.4), we have

$$h_n(t_1) - h_n(t_2) = \bar{D}(\theta_0 + \tilde{t}_1 n^{-1/2})t_1 - \bar{D}(\theta_0 + \tilde{t}_2 n^{-1/2})t_2 \tag{3.9}$$

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 $|D \cdot (t_1 - t_2)| \to 0$ as $|t_1 - t_2| \to 0$, and the stochastic equicontinuity of h_n on K follows. □

The above two lemmas mean that the sequence of stochastic processes h_n converges in distribution (in the sense of the weak convergence of probability measures) to the stochastic process h on any compact subset K of $\mathbb{R}^p$. This result is conveyed to the sequence $d_n[h_n(t)]$ in the next lemma.

Lemma 3.6. *Under the assumptions of Lemmas 3.4 and 3.5, and under assumption (E2), the stochastic processes C_n defined by*

$$C_n(t) = d_n[h_n(t)] = d_n[n^{1/2}\bar{g}(\theta_0 + tn^{-1/2})] \tag{3.10}$$

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

Proof. By Lemmas 3.4 and 3.5, we have the weak convergence of h_n to h on K. 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 (a generalized version of the continuous mapping theorem) of Prohorov (1956, p. 166) or Theorem 3.27 of Kallenberg (1997, p. 54). □

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 Theorem 3.3. By Lemma 3.6, C_n converges in distribution to C on every compact K. Clearly, $C(t)$ lives in a separable set of functions such that every sample path $t \mapsto C(t)$ is locally bounded and continuous and diverges to ∞ if $|t| \to \infty$ because of condition (E1) and the fact that D has full column rank. And finally every sample path $t \mapsto C(t)$ achieves its minimum at a unique point by (E3). Now we can apply Theorem 2.7 of Kim and Pollard (1990). □

4 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. Chamberlain (1987)'s efficiency result for optimal GMM estimators can again be verified following the same logic as in de Jong and Han (2002).

References

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

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

Davidson, J. (2000), *Econometric Theory*, Blackwell.

De Jong, R. M., and C. Han (2002), "The Properties of L_p-GMM Estimators," *Econometric Theory*, forthcoming.

Hansen, L. P. (1982), “Large Sample Properties of Generalized Method of Memonts Estimators,” *Econometrica*, 50, 1029–1054.

Jennrich, R. I. (1969), “Asymptotic Properties of Non-linear Least Squares Estimators,” *The Annals of Mathematical Statistics*, 40 (2), 633–643.

Kallenberg, O. (1997), *Foundations of Modern Probability*, Springer.

Kim, J., and D. Pollard (1990), “Cube Root Asymptotics,” *The Annals of Statistics*, 18, 191–219.

Manski, C. F. (1983), “Closest Empirical Distribution Estimation,” *Econometrica*, 51 (2), 305–319.

Newey, W. K. (1988), “Asymptotic Equivalence of Closest Moments and GMM estimators,” *Econometric Theory*, 4, 336–340.

Pollard, D. (1985), “New Ways to Prove Central Limit Theorems,” *Econometric Theory*, 1, 295–314.

Pötscher, B. M., and I. R. Prucha (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), 157–214.

Van der Vaart, A. W., and J. A. Wellner (1996), *Weak Convergence and Empirical Processes*, Springer.