The properties of L_p-GMM estimators

Robert de Jong and Chirok Han
Michigan State University

February 2000

Abstract

This paper considers Generalized Method of Moment-type estimators for which a criterion function is minimized that is not the "standard" quadratic distance measure, but instead is a general L_p distance measure. It is shown that the resulting estimators are root-n consistent, but not in general asymptotically normally distributed, and we derive the limit distribution of these estimators. In addition, we prove that it is not possible to obtain estimators that are more efficient than the "usual" L_2-GMM estimators by considering L_p-GMM estimators. We also consider the issue of the choice of the weight matrix for L_p-GMM estimators.

Keywords: L_p GMM, Generalized method of moment, L_p distance

1 Introduction

Since Lars Peter Hansen's (1982) original formulation, Generalized Method of Moment (GMM) estimation has become an extremely important and popular estimation technique in economics. This is due to the fact that economic theory usually implies moment conditions that are exploited in the GMM technique, while typically economic theory is uninformative about the exact stochastic structure of economic processes. GMM estimation provides an estimator when a certain set of moment conditions $Eg(y, \theta_0) = 0$ is *a priori* known to hold. When the number of moment conditions exceeds the number of parameters, we cannot hope to obtain an estimator by setting the empirical equivalent $\bar{g}(\theta)$ of our moment condition equal to zero, but instead we will need to make $\bar{g}(\theta)$ as close to zero as possible in some sense. The usual GMM formulation minimizes a quadratic measurement of distance. Hansen (1982) established the large sample properties of these GMM estimators under mild regularity conditions.

The above exposition raises the natural questions of what happens if distance measures other than a quadratic one is used and whether or not those other distance measures can give better estimators. The answer to the latter question is no, as Chamberlain (1987) has shown that the optimal GMM (in the usual sense) estimators attain the efficiency bound. Apart from this general remark on the efficiency of optimal GMM estimators, there have been attempts such as Manski (1983) and Newey (1988) to directly treat the use of non-quadratic measures of distance between population and empirical moments. In those articles results are stated that imply that under mild assumptions, estimators that minimize a general discrepancy function are consistent and asymptotically normally distributed. Based on these results, Newey (1988) concludes that (under regularity conditions) estimators using two different measures of distance are asymptotically equivalent if the corresponding Hessian matrices are asymptotically equal. This implies that it is impossible to obtain better estimators by modifying the quadratic criterion function, given the assumptions of that paper. This conclusion gives a direct argument for the use of quadratic distance measure beside Chamberlain's general argument.

However, when considering L_p-GMM as defined below, it turns out that only the L_2 norm satisfies the assumptions of Manski (1983) and Newey (1988), and other values of p in $[1, \infty)$ do not. The problems are the following. When $p = 1$, the L_p norm is not differentiable at 0; when $p \in (1, 2)$, it is continuously differentiable but is not twice differentiable at 0; when $p \in (2, \infty)$, it is continuously twice differentiable, but the Hessian matrix evaluated at the true parameter becomes zero (and therefore singular). Therefore, the papers by Newey and Manski have no implications for L_p-GMM for values of p other than 2. When considering L_p-GMM, it turns out that the "standard" asymptotic framework will fail. Also, the least absolute deviations type asymptotic framework also does not directly apply. Linton (1999) has recently pointed out in an example in *Econometric Theory* that the estimator minimizing the L_1 distance of the sample moments from zero can have a nonnormal limit distribution. In this paper, we will establish the limit distribution of general L_p-GMM estimators, and we show that L_p-GMM estimators are root-n consistent, but in general need not have an asymptotically normal distribution. In addition, we prove a theorem that shows that L_p-GMM estimators cannot be more efficient than L_2-GMM estimators, thereby strengthening Newey's conclusion to L_p-GMM estimators. Finally, we discuss the problem of finding the optimal weight matrix for L_p-GMM estimators.

Section 2 defines our estimator and gives the main theorem for consistency and asymptotic distribution, whose proof is given in the Mathematical Appendix. Section 3 discusses the efficiency of L_2-GMM among all L_p-GMM estimators. Section 4 describes the problem of the selection of the weight matrix. In addition, this section gives some interesting results for the case when $p = 1$ including Linton's (1999) example. The conclusions section (Section 5) is followed by a Mathematical Appendix in which all the proofs are gathered.

2 Main theorem

In this section, the main result of this paper on which the remainder of our discussion is based will be stated. Let $y_1, y_2, \ldots$ be a sequence of i.i.d. random vectors in $\mathbb{R}^m$. Let $g(y_i, \theta)$ be a set of q moment conditions with parameter $\theta \in \Theta \subset \mathbb{R}^k$, that is, let $g(y_i, \theta)$ be a random vector in $\mathbb{R}^q$ that satisfies

$$Eg(y_i, \theta_0) = 0. \tag{1}$$

Let $\bar{g}(\theta) = n^{-1}\sum_{i=1}^{n} g(y_i, \theta)$. The L_p norm $\|\cdot\|_p$ is defined as

$$\|x\|_p = (\sum_{j=1}^{q} |x_j|^p)^{1/p} \tag{2}$$

for $p \in [1, \infty)$. The L_p-GMM estimator $\hat{\theta}_n$ is assumed to satisfy

$$\|\bar{g}(\hat{\theta}_n)\|_p = \inf_{\theta \in \Theta} \|\bar{g}(\theta)\|_p. \tag{3}$$

Let $|x| = \max_{i,j} |x_{ij}|$ if x is a $k_1 \times k_2$ matrix. Let $\Omega = Eg(y_i, \theta_0)g(y_i, \theta_0)'$, and $D = E(\partial/\partial\theta')g(y_i, \theta_0)$. The regularity assumption below will be needed to establish our results:

Assumption 1.

(i) Θ *is a compact and convex subset in* $\mathbb{R}^k$*;*

(ii) θ_0 *is an interior point of* Θ*;*

(iii) $Eg(y_i, \theta) = 0$ *iff* $\theta = \theta_0$*, i.e.,* θ_0 *uniquely satisfies the moment conditions;*

(iv) $g(y, \theta)$ *is continuous in* θ *for each* $y \in \mathbb{R}^m$*, and is measurable for each* $\theta \in \Theta$*;*

(v) $E\sup_{\theta \in \Theta} |g(y_i, \theta)| < \infty$*;*

(vi) $\Omega = Eg(y_i, \theta_0)g(y_i, \theta_0)'$ *is finite;*

(vii) $g(y, \theta)$ *has first derivative with respect to* θ *which is continuous in* $\theta \in \Theta$ *for each* $y \in \mathbb{R}^m$ *and measurable for each* $\theta \in \Theta$*,* $E\sup_{\theta \in \Theta} |(\partial/\partial\theta')g(y_i, \theta)| < \infty$*, and* D *is of full column rank.*

(viii) $\|x + D\xi\|_p$ *achieves its minimum at a unique point of* ξ *in* $\mathbb{R}^k$ *for each* $x \in \mathbb{R}^q$*.*

The following theorem now summarizes the asymptotic properties of L_p-GMM estimators. Note that we do not yet explicitly consider weight matrices at this point, but such a treatment can be easily done with the result below at hand.

Theorem 1. *Let* Y *be a random vector in* $\mathbb{R}^q$ *distributed* $N(0, \Omega)$*. Then under Assumption 1,* $\hat{\theta}_n \to \theta_0$ *a.s., and*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{\xi \in \mathbb{R}^k} \|Y + D\xi\|_p. \tag{4}$$

The proof of this theorem, like all the proofs of this paper, can be found in the Appendix. As a special case of the above theorem, the usual L_2-GMM estimator can be considered. $\|Y + D\xi\|_2^2 = (Y + D\xi)'(Y + D\xi)$ is minimized by $\tilde{\xi} = -(D'D)^{-1}D'Y$, so applying Theorem 1, we get

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} -(D'D)^{-1}D'Y \sim N[0, (D'D)^{-1}D'\Omega D(D'D)^{-1}] \tag{5}$$

which coincides with usual analysis.

In examples below, we will show that for general values of p, normality need not result for the L_p-GMM estimator. We will be able to establish though that the limit distribution is symmetric around 0 and possesses finite second moments.

To conclude this section, note that part (*viii*) of Assumption 1 is far from innocent in the L_1 case. Consider for example sequences of random variables y_{i1} and y_{i2} that are independent of each other and are $N(0,1)$ distributed, and consider the L_p-GMM estimator that minimizes

$$|\bar{y_1} - \theta| + |\bar{y_2} - \theta|. \tag{6}$$

Part (*viii*) of Assumption 1 will not hold in this case, because any value of μ in the interval $[\min(\bar{y}_1, \bar{y}_2), \max(\bar{y}_1, \bar{y}_2)]$ will minimize the criterion function. Therefore, our result does not establish the limit distribution of the L_p-GMM estimator for this case. However, if we consider the weighted criterion function

$$|\bar{y_1} - \theta| + c|\bar{y_2} - \theta| \tag{7}$$

for any $c \in [0, \infty)$ except for $c = 1$, part (*viii*) of Assumption 1 will be satisfied.

3 Efficiency of L_2-GMM

In this section and in the remainder of this paper, we consider L_p-GMM estimation with a weight matrix W, i.e., L_p-GMM estimators that minimize the distance from zero of "weighted" average of moment conditions $\|W\bar{g}(\theta)\|_p$, where W is a $q \times q$ nonrandom matrix. It is straightforward to extend our analysis to the case of estimated matrices W, and we will not pursue that issue here. Clearly, whenever $Eg(y_i, \theta_0) = 0$ we will have $EWg(y_i, \theta_0) = 0$, and therefore our previous analysis applies. Below, we will keep using the notations Y, D, and Ω defined previously.

Let $\tilde{\xi}$ minimize $\|W(Y + D\xi)\|_p$. Applying Theorem 1, we see that $\hat{\theta}_n$, which minimizes $\|W\bar{g}(\theta)\|_p$, is strongly consistent (since $Wg(y_i, \theta_0)$ is also a set of legitimate moment conditions) and $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \tilde{\xi}$.

To facilitate the efficiency discussion, we need to show unbiasedness of L_p-GMM estimators. This is established by noting that the limiting distribution of $n^{1/2}(\hat{\theta}_n - \theta_0)$ is symmetric and has a finite second moment. The following theorem states the unbiasedness result:

Theorem 2. *Under Assumption 1, L_p-GMM estimators are asymptotically unbiased.*

Because of the asymptotic unbiasedness of our estimators, we can compare weighted L_p-GMM estimators by their asymptotic variances. This property is crucial to prove the following theorem. This result states that optimal L_2-GMM estimators are asymptotically efficient among the class of weighted L_p-GMM estimators.

Theorem 3. *Under Assumption 1, an optimal L_2-GMM estimator is asymptotically efficient among the class of weighted L_p-GMM estimators, i.e., the asymptotic variance of an optimal L_2-GMM estimator is less than that of any weighted L_p-GMM estimator.*

The above theorem provides us with the knowledge that the central message from the result by Newey (1988)—that there is no potential for efficiency improvement by considering discrepancy functions other than quadratic—can be extended towards L_p-GMM estimators. Basically, Theorem 3 is obtained by noting that the expression for the limit distribution can be viewed as a finite sample estimation problem in its own right, for which the Cramér-Rao underbound applies.

4 Further remarks on weight matrices

In this section, we will discuss various issues involving the choice of the weight matrix W and discuss several examples. We will not be able to prove optimality of a particular nonsingular weight matrix for general L_p-GMM, but instead we will sketch some of the issues below.

It is well-known that the optimal weight matrix W for $p = 2$ satisfies $W\Omega W' = I$ (or $W'W = \Omega^{-1}$). This result can be easily obtained using our first theorem too, for

$$\|W(Y + D\xi)\|_2^2 = (Y + D\xi)'W'W(Y + D\xi) \tag{8}$$

is minimized by

$$\tilde{\xi} = -(D'W'WD)^{-1}D'W'WY \tag{9}$$

and its variance is minimized when $W'W = \Omega^{-1}$. Therefore the optimal L_2-GMM estimator has the asymptotic distribution

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N[0, (D'\Omega^{-1}D)^{-1}]. \tag{10}$$

Can this efficiency be attained for p other than 2? In general, the answer is yes. It can be achieved for general p by weighting cleverly. Consider

$$W^* = (\Omega^{-1}D \quad W_2)' \tag{11}$$

where W_2 is of size $q \times (q-k)$, chosen to be orthogonal to D, i.e., $W_2'D = 0$ and chosen such that W^* is nonsingular. This weight matrix always exists when $q > k$.[1] Then

$$\|W^*(Y + D\xi)\|_p^p = \|D'\Omega^{-1}Y + D'\Omega^{-1}D\xi\|_p^p + \|W_2'Y\|_p^p \tag{12}$$

is minimized by $\tilde{\xi} = -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}Y \sim N[0, (D'\Omega^{-1}D)^{-1}]$ for any $p \geq 1$. So the W^*-weighted L_p-GMM estimator $\hat{\theta}_n$ with W^* chosen as in Equation (11) has the asymptotic distribution (10), and therefore the weight matrix W^* is optimal for any p.

For $p = 2$, there are two different types of optimal weights. One is given by (11) (say, $D'\Omega^{-1}$ type) and the other is characterized by $W\Omega W' = I$. In general, each of these neither implies nor is implied by the other, but they give one and the same asymptotic distribution. Furthermore, the optimal weight of the second type is not unique, since any orthogonal transformation of an optimal weight is again optimal. (When $W\Omega W' = I$, $V = HW$ also satisfies $V\Omega V' = I$ provided $H'H = HH' = I$.) This is, of course, because the W-weighted L_2 distance $\|Wx\|_2 = (x'W'Wx)^{1/2}$ depends only on the product $W'W$ but not W itself.

But when $p \neq 2$, two different orthogonal transformations W and V of $\Omega^{-1/2}$ are not expected to give equivalent asymptotic distribution, even though both $W\Omega W' = I$ and $V\Omega V' = I$ hold. Here are a few examples.

(i) First example is for $p = 1$, $q = 2$, and $k = 1$. Let $D = -(1 \quad 2)'$, $\Omega = I$, $W = I$, and

$$V = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}. \tag{13}$$

[1] This weight needs $W_2'D = 0$ and $|W^*| \neq 0$, so there are $(q-k)k+1$ restrictions. But W^* has $q(q-k)$ free parameters. The number of parameters is greater than or equal to the number of restrictions when $q > k$.

It can be seen that V is an optimal weight matrix here, since the same limit distribution as for optimal L_2-GMM is obtained using V. In the case $W = I$, this situation arises when when we are minimizing the criterion function

$$|\bar{y_1} - \theta| + |\bar{y_2} - 2\theta| \tag{14}$$

where the y_{ij} are independent across all i and j, have mean θ_0, and have variance 1. Then $W\Omega W' = V\Omega V' = I$, but the rescaled and centered W-weighted L_1-GMM estimator is asymptotically distributed as $N(0, 1/4)$, while the rescaled and centered V-weighted L_p-GMM estimator is asymptotically distributed $N(0, 1/5)$.

(ii) Here is a more interesting example for $p = 1$. Let $q = 3$, $k = 1$, $\Omega = I$, and $D = -(1\ 1\ 1)'$, and consider the two weights $W = I$ and

$$V = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ -\sqrt{2/3} & 1/\sqrt{6} & 1/\sqrt{6} \end{pmatrix}. \tag{15}$$

Again, V can be shown to be an optimal weight matrix in this example. In the case of $W = I$, this situation could result when we are minimizing the criterion function

$$|\bar{y_1} - \theta| + |\bar{y_2} - \theta| + |\bar{y_3} - \theta| \tag{16}$$

where the y_{ij} are independent across all i and j, have mean θ_0, and have variance 1. Note that both W and V are chosen to be orthogonal. The W-weighted L_1-GMM estimator (after centering and scaling) converges in distribution to $\operatorname{argmin}_\xi \|N(0, I_3) + D\xi\|_1$. The minimizing argument $\tilde{\xi}$, which is the (unique) median of three independent standard normal random variables, has distribution

$$P(\tilde{\xi} \le x) = 6 \int_{-\infty}^{x} \Phi(t)[1 - \Phi(t)]\phi(t)dt = \Phi(x)^2[3 - 2\Phi(x)] \tag{17}$$

(see Linton (1999)). This distribution is not normal, and simulations for three standard normals illustrates that the density of the median has sharper center and thicker tail than a (properly rescaled) normal ($N(0, 2/3)$). The result of using V as the weight matrix is different. We have $VD = (-\sqrt{3}\ 0\ 0)'$ and the V-weighted L_1-GMM estimator (after centered and scaled) converges in distribution to $\operatorname{argmin}_\xi\{\|Z + VD\xi\|_1 = |Z_1 - \sqrt{3}\xi| + |Z_2| + |Z_3|\}$ where $Z = (Z_1, Z_2, Z_3)' \sim N(0, I)$. Note that basically, this optimal weight matrix will eliminate two out of three absolute value elements of the criterion function of Equation (16). The solution $\tilde{\xi}$ is distributed $N(0, 1/3)$. This example shows that two weights W and V that satisfy $W\Omega W' = I$ and $V\Omega V' = I$ can give asymptotics different not only in variance but in the type of limit distribution, since the one distribution is non-normal, while the other is normal.

(iii) The only tractable example for $p > 2$ that we could find is the following. Consider the case of $q = 3$, $k = 1$, $\Omega = I$, and $D = -(1\ 1\ 1)'$. The weight V of (15) is again optimal and the V-weighted L_3 GMM estimator is asymptotically normal. In the case when the weight is $W = I$ so the objective function to be minimized is

$$|\bar{y_1} - \theta|^3 + |\bar{y_2} - \theta|^3 + |\bar{y_3} - \theta|^3, \tag{18}$$

the W-weighted L_3 GMM estimator (after centering and rescaling) converges in distribution to $\tilde{\xi} = \operatorname{argmin}_{\xi} \|Y + D\xi\|_3$ where $Y = (Y_1, Y_2, Y_3)' \sim N(0, I_3)$. Let $(Y_{(1)}, Y_{(2)}, Y_{(3)})$ be the order statistic of (Y_1, Y_2, Y_3), and $(\delta_{(1)}, \delta_{(2)}, \delta_{(3)})$ be the order statistic of $(|Y_1 - Y_2|, |Y_2 - Y_3|, |Y_3 - Y_1|)$. Then it turns out that

$$\begin{aligned} \tilde{\xi} &= \bar{Y} + \operatorname{sgn}(Y_{(1)} + Y_{(3)} - 2Y_{(2)})[(2/3)(\delta_{(3)} + \delta_{(2)}) - (2\delta_{(3)}\delta_{(2)})^{1/2}] \\ &= Y_{(2)} - \operatorname{sgn}(Y_{(1)} + Y_{(3)} - 2Y_{(2)})[\delta_{(3)} - (2\delta_{(3)}\delta_{(2)})^{1/2}] \qquad (19) \end{aligned}$$

where $\operatorname{sgn}(a) = 1\{a > 0\} - 1\{a < 0\}$ and $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$. In simulations, this distribution cannot be distinguished from a normal.

The natural question now arises whether we can get optimality by a nonsingular weight matrix W satisfying $W\Omega W' = I$. In short, the answer is yes provided $D'\Omega^{-1}D$ equals a scalar or a scalar matrix (a scalar times the identity matrix). The question here is whether we can construct $(\Omega^{-1}D \quad W_2)'$ (where $W_2'D = 0$) by an orthogonal transformation of $\Omega^{-1/2}$, that is, whether there exists an orthogonal matrix H of size $q \times q$ such that $H\Omega^{-1/2} = \lambda(\Omega^{-1}D \quad W_2)'$ and $W_2'D = 0$. If such a weight matrix H is exist, it will have the form $H = \lambda(\Omega^{-1/2}D \quad \Omega^{1/2}W_2)'$ for which (a) $W_2'D = 0$, (b) H is nonsingular, i.e., $|H| \neq 0$, and (c) $HH' = I$, i.e.,

$$HH' = \lambda^2 \begin{pmatrix} D'\Omega^{-1}D & D'W_2 \\ W_2'D & W_2'\Omega W_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \qquad (20)$$

(a) imposes $(q-1)k$ restrictions, and (b) imposes 1 extra restriction. When $k = 1$, (c) is equivalent to $W_2'\Omega W_2 = (D'\Omega^{-1}D)I_{q-1}$ due to (a), which imposes $(q-1)(q-2)/2$ more restrictions. Therefore, when $k = 1$, we have $q(q-1)+1$ free parameters (for W_2 and λ) and $(q-1)+1+(q-1)(q-2)/2 = q(q-1)/2+1$ restrictions. So the number of parameters to be set is greater than or equal to the number of restrictions, whence we conclude that we can find W_2 satisfying (a), (b), and (c). When $k > 1$, (c) can not be satisfied unless $D'\Omega^{-1}D$ is a scalar matrix, but if $D'\Omega^{-1}D$ is so, W_2 and λ satisfying (a), (b), and (c) can be found. The V matrices in the examples above are constructed in this way and are optimal for those problems.

Note that this rule does not depend on the specific value of p, and that the reason the weight W satisfying $W\Omega W' = I$ is optimal for $p = 2$ does not lie in that the optimal weight of type $D'\Omega^{-1}$ can be obtained by an orthogonal transformation of $\Omega^{-1/2}$, but in the specific properties of L_2 distance.

5 Conclusion

In this paper we derived an abstract expression for the limit distribution of estimators which minimizes the L_p distance between population moments an sample moments, as follows:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{\xi \in \mathbb{R}^k} \|Y + D\xi\|_p \qquad (21)$$

where $Y \sim N[0, Eg(y_i, \theta_0)g(y_i, \theta_0)']$ and $D = E(\partial/\partial\theta')g(y_i, \theta_0)$. This asymptotic representation allows a generalization of the well-known GMM framework of Hansen (1982) towards the L_p distance. As mentioned in the introduction, Manski (1983) and Newey

(1988) generalized GMM to allow arbitrary distance (or, more generally, discrepancy) function. But unfortunately they need the second order differentiability of the distance functions and the nonsingularity of a Hessian matrix evaluated at true parameter. Only the L_2 distance satisfies these conditions among all L_p distances.

However, our analysis can not give an explicit form for the asymptotic distribution, but only allows the above abstract representation in terms of the argmin functional. Nonetheless, our method directly supports the result of Chamberlain (1987) that the optimal L_2-GMM estimator is efficient among the class of L_p-GMM estimators. Interestingly, our analysis reduced the analysis of efficiency issues of L_p-GMM estimators to the analysis of the small sample properties of estimators minimizing the L_p distance between Y and $-D\xi$, i.e., $\operatorname{argmin}_{\xi \in \mathbb{R}^k} \|Y + D\xi\|_p$.

As a final remark, it will be interesting whether any robustness argument can be made for $p = 1$ as an analogy of the robustness of median. However, our method deals with asymptotics, for which all involved random elements are normal. We think the problem of robustness can be issued and arguments for L_1 GMM estimation over L_2 one can be made in the context of small sample (the meaning of 'small' should be specified in any way: Normalized sums have distribution far from normal, or the estimator is away from the true parameter with still high probability) properties of L_p GMM estimators, which is far beyond the scope of this paper, and to which we still have a long way to go.

Mathematical Appendix

In order to establish the theorems, we will need several results that will be stated as lemmas. Lemma 1 is used to prove Theorem 2 (the asymptotic unbiasedness of L_p-GMM estimators).

Lemma 1. *Let a random vector Y in $\mathbb{R}^q$ with finite q have a normal distribution. Let D be a real nonrandom matrix of size $q \times k$ $(q \geq k)$ with full column rank. Then for any $p \in [1, \infty)$,*

$$\tilde{\xi} = \underset{\xi \in \mathbb{R}^k}{\operatorname{argmin}} \|Y + D\xi\|_p \tag{22}$$

will have a well-defined finite covariance matrix.

Proof. First note that, because $D'D$ has full column rank under Assumptions 1,

$$\begin{aligned} \tilde{\xi}'\tilde{\xi} &\leq \tilde{\xi}'D'D\tilde{\xi}/\lambda_{min}(D'D) \\ &\leq (\|Y + D\tilde{\xi}\|_2 + \|Y\|_2)^2/\lambda_{min}(D'D) \\ &\leq c_{p/2}^{2/p}(\|Y + D\tilde{\xi}\|_p + \|Y\|_p)^2/\lambda_{min}(D'D) \\ &\leq c_{p/2}^{2/p}(2\|Y\|_p)^2/\lambda_{min}(D'D) \end{aligned} \tag{23}$$

where the first inequality follows from full column rank of D, the second inequality is the triangle inequality, the third is Loève's c_r inequality (see Davidson (1994), p. 140), and the fourth follows by the fact that $\|Y + D\xi\|_p$ is minimized at $\xi = \tilde{\xi}$. The result then follows because all moments of the normal distribution are finite. □

The first step towards the proof of Theorem 1 is the strong consistency proof for L_p-GMM estimators, which can be accomplished by invoking several theorems from Bierens (1994).

Lemma 2. *Under Assumption 1, the L_p-GMM estimator $\hat{\theta}_n$ is strongly consistent.*

Proof. First, conditions (i) and (iv) of Assumption 1 ensure the existence and measurability of $\hat{\theta}_n$ by Theorem 1.6.1 of Bierens (1994). The above conditions together with condition (v) of Assumption 1 imply that $\bar{g}(\theta)$ converges to $Eg(y_i, \theta)$ almost surely uniformly on Θ by Theorem 2.7.5 of Bierens (1994). Hence, $\|\bar{g}(\theta)\|_p \to \|Eg(y_i, \theta)\|_p$ a.s. uniformly on Θ since $\|\cdot\|_p$ is continuous. Finally, this uniform convergence result and the uniqueness of θ_0 by condition (iii) of Assumption 1 give the stated result by Theorem 4.2.1 of Bierens (1994). □

To prove the main assertion of Theorem 1, we will use Theorem 2.7 of Kim and Pollard (1990). We restate Kim and Pollard's theorem as our next lemma.

Lemma 3. *Let $Q, Q_1, Q_2, \cdots$ be real-valued random processes on $\mathbb{R}^k$ with continuous paths, and $\hat{\xi}_n$ be random vector in $\mathbb{R}^k$, such that*

(i) *$Q(\xi) \to \infty$ as $|\xi| \to \infty$;*

(ii) *$Q(\cdot)$ achieves its minimum at a unique point in $\mathbb{R}^k$;*

(iii) *Q_n converges weakly to Q on any set $\Xi = [-M, M]^k$;*

(iv) $\hat{\xi}_n = O_P(1)$*;*

(v) $\hat{\xi}_n$ *minimizes* $Q_n(\xi)$*.*

Then $\hat{\xi}_n \xrightarrow{d} \operatorname{argmin}_{\xi \in \mathbb{R}^k} Q(\xi)$*.*

Proof. See Theorem 2.7 of Kim and Pollard (1990). □

To apply Kim and Pollard's theorem and show that its conditions are satisfied in our situation, we need the following three lemmas. For these lemmas, we need the following definitions. Define

$$\hat{\xi}_n = n^{1/2}(\hat{\theta}_n - \theta_0), \quad \text{for } n = 1, 2, \ldots, \tag{24}$$

where $\hat{\theta}_n$ is L_p-GMM estimator. Define $\mathbb{R}^q$-valued random functions $h, h_1, h_2, \ldots$ by

$$h_n(\xi) := \begin{cases} n^{1/2}\bar{g}(\theta_0 + \xi n^{-1/2}), & \text{if } \theta_0 + \xi n^{-1/2} \in \Theta \\ n^{1/2}\bar{g}(\theta^0), & \text{otherwise} \end{cases} \tag{25}$$

where $\theta^0 = \operatorname{argmax}_{\theta \in \Theta} \|\bar{g}(\theta)\|_p$, for $n = 1, 2, \ldots$, and $h(\xi) = Y + D\xi$ where Y is a $\mathbb{R}^q$-valued random vector distributed $N(0, \Omega)$. Let

$$Q_n(\xi) = \|h_n(\xi)\|_p \quad \text{and} \quad Q(\xi) = \|h(\xi)\|_p. \tag{26}$$

The lemmas that we need for the proof of our central result are then the following:

Lemma 4. *Suppose the conditions of Assumption 1 are satisfied. Then* $n^{1/2}(\hat{\theta}_n - \theta_0) = O_P(1)$*.*

Proof. The Taylor expansion of $\bar{g}(\theta)$ around $\theta = \theta_0$ implies that $\bar{g}(\hat{\theta}_n) = \bar{g}(\theta_0) + (\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$, where $\tilde{\theta}_n$ is a mean value in between $\hat{\theta}_n$ and θ_0. From the above Taylor series, from the triangular inequality for the L_p norm, and from the fact that $\hat{\theta}_n$ minimizes $\|\bar{g}(\theta)\|_p$, we have

$$\|n^{1/2}(\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)\|_p \leq \|n^{1/2}\bar{g}(\hat{\theta}_n)\|_p + \|n^{1/2}\bar{g}(\theta_0)\|_p \leq 2\|n^{1/2}\bar{g}(\theta_0)\|_p. \tag{27}$$

But condition *(vi)* of Theorem 1 implies that $n^{1/2}\bar{g}(\theta_0)$ converges in distribution by central limit theorem, and therefore is $O_P(1)$. Therefore,

$$\|n^{1/2}(\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)\|_p = O_P(1). \tag{28}$$

Condition *(vii)* of Theorem 1 implies that $(\partial/\partial\theta')\bar{g}(\theta)$ follows a strong uniform law of large numbers, which combined with the consistency of $\hat{\theta}$ implies that $(\partial/\partial\theta')\bar{g}(\tilde{\theta}_n) \xrightarrow{\text{a.s.}} E(\partial/\partial\theta')g(y_i, \theta_0) = D$. Now let $\tilde{D} = (\partial/\partial\theta')\bar{g}(\tilde{\theta}_n)$. Then it follows that $\tilde{D}'\tilde{D} \xrightarrow{\text{a.s.}} D'D$. Since $D'D$ is strictly positive definite, $\tilde{D}'\tilde{D}$ becomes strictly positive definite for n large enough. Therefore, for n large enough,

$$n(\hat{\theta}_n - \theta_0)'(\hat{\theta}_n - \theta_0) \leq n(\hat{\theta}_n - \theta_0)'\tilde{D}'\tilde{D}(\hat{\theta}_n - \theta_0)/\tilde{\lambda}_{min}, \tag{29}$$

where $\tilde{\lambda}_{min}$ is the smallest eigenvalue of $\tilde{D}'\tilde{D}$. Because $\tilde{\lambda}_{min} \xrightarrow{\text{a.s.}} \lambda_{min} > 0$ where λ_{min} is the smallest eigenvalue of $D'D$, we get $\tilde{\lambda}_{min} \geq 0.5\lambda_{min}$ eventually (for n large enough) almost surely. Therefore, as n increases, the right hand side of (29) eventually becomes less than $4n(\hat{\theta}_n - \theta_0)'\tilde{D}'\tilde{D}(\hat{\theta}_n - \theta_0)/\lambda_{min}$. By Equation (28) and because of the equivalence of L_p and L_q norms for $p, q \in [1, \infty)$, this expression is $O_P(1)$, which completes the proof. □

Lemma 5. *Consider random functions $Q, Q_1, Q_2, \dots$ defined by (26). Under Assumption 1, the finite-dimensional distributions of Q_n converge to the finite-dimensional distributions of Q.*

Proof. With fixed ξ, condition (*ii*) of Assumption 1 (θ_0 is an interior point of Θ) ensures that $\theta_0 + \xi n^{-1/2}$ belongs to Θ for n large enough. When this happens, by the Taylor expansion,

$$h_n(\xi) = n^{1/2}\bar{g}(\theta_0 + \xi n^{-1/2}) = n^{1/2}\bar{g}(\theta_0) + (\partial/\partial\theta')\bar{g}(\theta_0 + \tilde{\xi}n^{-1/2})\xi, \tag{30}$$

with $\tilde{\xi}$ lying in between ξ and 0. Condition (*vi*) of Assumption 1 (finiteness of the second moment of $g(y_i, \theta_0)$) implies that $n^{1/2}\bar{g}(\theta_0) \xrightarrow{d} Y$, and condition (*vii*) of Theorem 1 implies that $(\partial/\partial\theta)\bar{g}(\theta_0 + \tilde{\xi}n^{-1/2})\xi \to D\xi$ a.s. similar to the proof of Lemma 4.

To conclude the proof and show the convergence of the finite-dimensional distributions of h_n to h, we can use the Cramér-Wold device (see for example Billingsley (1968), Theorem 7.7), which states that

$$(h_n(\xi_1)' \cdots h_n(\xi_r)')' \xrightarrow{d} (h(\xi_1)' \cdots h(\xi_r)')' \tag{31}$$

if and only if

$$\sum_{j=1}^{r} \lambda_j' h_n(\xi_j) \xrightarrow{d} \sum_{j=1}^{r} \lambda_j' h(\xi_j) \tag{32}$$

for each $\lambda_1 \in \mathbb{R}^q, \dots, \lambda_r \in \mathbb{R}^q$. And (32) is to be easily shown using the result of the first part of this proof.

Finally, note that since $\|\cdot\|_p$ is continuous, the finite-dimensional distribution of $Q_n = \|h_n\|_p$ converge to those of $Q = \|h\|_p$ by the continuous mapping theorem. □

Lemma 6. *Under Assumption 1, $Q_n(.)$ defined by Equation (26) is stochastically equicontinuous on any set $\Xi = [-M, M]^k$.*

Proof. Using the triangular inequality for $\|\cdot\|_p$, we have

$$\begin{aligned} &|Q_n(\xi_1) - Q_n(\xi_2)| \\ &= |\,\|h_n(\xi_1)\|_p - \|h_n(\xi_2)\|_p| \\ &\le \|h_n(\xi_1) - h_n(\xi_2)\|_p \\ &= \|(\partial/\partial\theta')\bar{g}(\theta_0 + \tilde{\xi}_1 n^{-1/2})\xi_1 - (\partial/\partial\theta')\bar{g}(\theta_0 + \tilde{\xi}_2 n^{-1/2})\xi_2\|_p \end{aligned} \tag{33}$$

where $\tilde{\xi}_i$ lies in between ξ_i and 0 for $i = 1, 2$. By the strong uniform law of large numbers for $(\partial/\partial\theta')\bar{g}(\theta)$ and the convergence to zero of $\tilde{\xi}_i n^{-1/2}$ uniformly over all ξ_1 and ξ_2,

$$\sup_{\xi_1 \in \Xi, |\xi_1 - \xi_2| < \delta} |Q_n(\xi_1) - Q_n(\xi_2)| \xrightarrow{\text{a.s.}} \sup_{\xi_1 \in \Xi, |\xi_1 - \xi_2| < \delta} \|D(\xi_1 - \xi_2)\|_p \tag{34}$$

under the conditions of Assumption 1. Therefore, by nonsingularity of $D'D$, it follows that for all $\eta > 0$

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(\sup_{\xi_1 \in \Xi, |\xi_1 - \xi_2| < \delta} |Q_n(\xi_1) - Q_n(\xi_2)| > \eta) = 0, \tag{35}$$

which is the stochastic equicontinuity condition. □

Proof of Theorem 1. The strong consistency result of this theorem is proven in Lemma 2. For the proof of the main assertion of this theorem, we will show that for the Q_n, Q and $\hat{\xi}_n$ as defined above, all the conditions of Lemma 3 are implied by the conditions of Theorem 1. First, note that Q_n, Q, and $\hat{\xi}_n$, defined by (24) and (26), satisfy conditions (i)–(v) of Lemma 3 under the conditions of Theorem 1. Condition (v) of Theorem 3 is guaranteed by the definitions of $\hat{\theta}_n$, $\hat{\xi}_n$, and Q_n. It is also not difficult to notice that condition (i) of Theorem 3 is trivially satisfied since D is of full column rank. And condition (ii) of Theorem 3 is just supposed by condition $(viii)$ of Theorem 1. The weak convergence condition is verified by showing stochastic equicontinuity and finite-dimensional convergence, which together with compactness of the parameter space is well-known to imply weak convergence. Lemmas 4, 5, and 6 therefore ensure that the conditions of Theorem 3 are all implied by the conditions of Theorem 1, and therefore convergence in distribution of our estimator is proven by invoking Theorem 3. □

Proof of Theorem 2. By Lemma 1, $\tilde{\xi}$ has a finite mean. And by Theorem 1, $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \tilde{\xi} = \operatorname{argmin}_\xi \|Y + D\xi\|_p$ where $Y \sim N(0, \Omega)$ and $D = E(\partial/\partial\theta')g(y_i, \theta_0)$. From the symmetry of Y, it follows that $\|Y+D\xi\|_p$ is distributed identically to $\|Y+D(-\xi)\|_p$, which implies identical distributions of $\tilde{\xi}$ and $-\tilde{\xi}$. Therefore, the mean of $\tilde{\xi}$ is 0. □

Proof of Theorem 3. Let $\hat{\theta}_n$ be the W-weighted L_p-GMM estimator. By Theorem 1,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \operatorname*{argmin}_{\xi} \|W(Y + D\xi)\|_p. \tag{36}$$

So the problem here is to show that $\tilde{\xi}_2 = \operatorname{argmin}_\xi \|\Omega^{-1/2}(Y+D\xi)\|_2$ has smaller variance than any other $\tilde{\xi}_p = \operatorname{argmin}_\xi \|W(Y + D\xi)\|_p$. Now, let us view the minimization problem $\min_\xi \|Y + D\xi\|_p$ as generating estimators $\tilde{\xi}_p$ of the unknown parameter ξ, where $Y \sim N(-D\xi, \Omega)$ with known Ω. The result of Theorem 2 now states that all L_p-GMM estimators will be asymptotically unbiased, and the argument can be easily extended to show global unbiasedness of $\tilde{\xi}_p$ for ξ (as required for the application of the Cramér-Rao lower variance bound). The likelihood function

$$L(Y, D; \xi) = (2\pi)^{-q/2}|\Omega|^{-1/2} \exp(-(1/2)(Y + D\xi)'\Omega^{-1}(Y + D\xi)) \tag{37}$$

satisfies all the required regularity conditions for Cramér-Rao inequality (see Theil (1971), p.384). And it now follows that the asymptotic distribution of the optimal L_2-GMM estimator

$$\operatorname*{argmin}_{\xi\in\mathbb{R}^k} \|\Omega^{-1/2}(Y + D\xi)\|_2 = -(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}Y \tag{38}$$

attains the Cramér-Rao variance lower bound of $(D'\Omega^{-1}D)^{-1}$, since it equals the maximum likelihood estimator. The result then follows. □

References

Bierens, Herman J. (1994) *Topics in Advanced Econometrics* (Cambridge University Press)

Billingsley, Patrick (1968) *Convergence of Probability Measures* (John Wiley & Sons, Inc.)

Chamberlain, Gary (1987) 'Asymptotic efficiency in estimation with conditional moment restrictions.' *Journal of Econometrics* 34, 305–334

Davidson, J. (1994) *Stochastic limit theory* (Oxford: Oxford University Press)

Hansen, L. P. (1982) 'Large sample properties of generalized method of moments estimators.' *Econometrica* 50, 1029–1054

Kim, J., and D. Pollard (1990) 'Cube root asymptotics.' *The Annals of Statistics* 18, 191–219

Linton, Oliver (1999) 'Problem.' *Econometric Theory* 15, 151

Manski, Charles F. (1983) 'Closest empirical distribution estimation.' *Econometrica* 51, 305–319

Newey, Whitney K. (1988) 'Asymptotic equivalence of closest moments and GMM estimators.' *Econometric Theory* 4, 336–340

Theil, Henri (1971) *Principles of Econometrics* (John Wiley & Sons, Inc.)