

Lecture Note 3: Extremum Estimators and GMM, Part II

(Based on Newey and McFadden 1994)

Verifying Consistency: Here are some examples of ways to verify the key assumptions (i), (iii), and (iv) of the main consistency theorem stated in LN2. Condition (i) turns out to be closely related to identification, and conditions (iii) and (iv) will usually follow from the uniform law of large numbers.

Condition (ii), compactness of the parameter space, is usually assumed directly. In some cases consistency can be shown without requiring compactness, see Newey and McFadden.

Maximum Likelihood

First, start with conditions (iii) and (iv). Since

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(z_i|\theta),$$

we can apply the ULLN directly. Suppose that $\log f(z_i|\theta)$ is continuous at each $\theta \in \Theta$ with probability one, and

$$E[\sup_{\theta \in \Theta} |\log f(z_i|\theta)|] < \infty.$$

The, by the ULLN,

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \log f(z_i|\theta) - E[\log f(z_i|\theta)] \right\| \xrightarrow{p} 0,$$

and $Q(\theta) = E[\log f(z_i|\theta)]$ is continuous.

Next, turn to condition (i).

We say a likelihood model is *identified* if the distribution of z_i at θ_0 is different from the distribution at any other θ . Formally, for any $\theta \neq \theta_0$, let

$$\Delta = \{z : f(z|\theta) \neq f(z|\theta_0)\}.$$

Then we require

$$Pr_{\theta_0}(z_i \in \Delta) > 0.$$

An implication is that the ratio $f(z_i|\theta)/f(z_i|\theta_0)$, regarded as a random variable, is not degenerate at 1.

Now, consider

$$\begin{aligned} Q(\theta) - Q(\theta_0) &= E[\log f(z_i|\theta)] - E[\log f(z_i|\theta_0)] \\ &= E \left[\log \frac{f(z_i|\theta)}{f(z_i|\theta_0)} \right] \end{aligned}$$

$$\begin{aligned}
&< \log E \left[\frac{f(z_i|\theta)}{f(z_i|\theta_0)} \right] \\
&= \log \int \frac{f(z|\theta)}{f(z|\theta_0)} f(z|\theta_0) dz \\
&= \log \int f(z|\theta) dz \\
&= \log 1 = 0.
\end{aligned}$$

Here, the strict inequality follows from Jensen's inequality, which holds strictly when the random variable is nonconstant and positive.

So it follows that for any $\theta \neq \theta_0$, $Q(\theta) < Q(\theta_0)$, verifying condition (i).

Consistency for GMM:

Let

$$\begin{aligned}
\hat{g}_n(\theta) &:= \frac{1}{n} \sum_{i=1}^n g(z_i, \theta), \\
g_0(\theta) &= E[g(z_i, \theta)] \quad (= 0 \text{ at } \theta = \theta_0).
\end{aligned}$$

Then

$$\hat{Q}_n(\theta) = -\hat{g}_n(\theta)' \hat{W} \hat{g}_n(\theta).$$

Suppose that $g(z_i, \theta)$ satisfies the conditions for the ULLN. Then

$$\sup_{\theta \in \Theta} \|\hat{g}_n(\theta) - g_0(\theta)\| \xrightarrow{p} 0,$$

and $g_0(\theta)$ is continuous.

Assume that $\hat{W} \xrightarrow{p} W$, where W is positive semidefinite and finite. Define

$$Q(\theta) = -g_0(\theta)' W g_0(\theta).$$

This is a continuous function of θ , since g_0 is continuous and W is positive semidefinite.

Can then show (see NM, Theorem 2.6) that:

$$\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q(\theta)| \xrightarrow{p} 0.$$

So we've verified (iii) and (iv).

Next, consider condition (i): by assumption

$$g_0(\theta_0) = E[g(z_i, \theta_0)] = 0.$$

Suppose we can also show that

$$Wg_0(\theta) \neq 0 \quad \forall \theta \neq \theta_0.$$

Let $R'R = W$. Then the previous display implies

$$Rg_0(\theta) \neq 0 \quad \forall \theta \neq \theta_0.$$

Therefore

$$Q(\theta) = -(Rg_0(\theta))'(Rg_0(\theta)) < 0 \quad \forall \theta \neq \theta_0.$$

So $Q(\theta)$ is uniquely maximized at $\theta = \theta_0$.

Asymptotic Normality of Extremum Estimators

Idea: in large samples, estimators are often approximately equal to sample averages, so a CLT gives asymptotic normality.

Example: MLE

Assume the log likelihood $\log f(z_i|\theta)$ is twice continuously differentiable and that $\hat{\theta}$ is an interior solution to the maximum likelihood problem. Then

$$0 = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log f(z_i|\hat{\theta}).$$

Use the mean-value version of Taylor's theorem to write each individual term in the sum as

$$\nabla_{\theta} \log f(z_i|\hat{\theta}) = \nabla_{\theta} \log f(z_i|\theta_0) + \nabla_{\theta\theta} \log f(z_i|\bar{\theta})(\hat{\theta} - \theta_0),$$

where $\bar{\theta}$ is between θ_0 and $\hat{\theta}$.

So we can write

$$0 = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log f(z_i|\theta_0) + \left[\frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta} \log f(z_i|\bar{\theta}) \right] (\hat{\theta} - \theta_0).$$

Rearrange:

$$\sqrt{n}(\hat{\theta} - \theta_0) = - \left[\frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta} \log f(z_i|\bar{\theta}) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} \log f(z_i|\theta_0).$$

Let

$$H = E[\nabla_{\theta\theta} \log f(z_i|\theta_0)],$$

$$J = E [(\nabla_{\theta} \log f(z_i|\theta_0))(\nabla_{\theta} \log f(z_i|\theta_0))'].$$

Then, as long as

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta} \log f(z_i|\theta)$$

converges in probability to

$$E[\nabla_{\theta\theta} \log f(z_i|\theta)],$$

uniformly in a neighborhood of θ_0 , and since $\bar{\theta} \xrightarrow{p} \theta_0$ (it is “between” $\hat{\theta}$ and θ_0), we have

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta} \log f(z_i|\bar{\theta}) \xrightarrow{p} H.$$

Note: in reading the last few equations, be sure to distinguish between $\bar{\theta}$, θ_0 , and arbitrary θ .

Also, $\nabla_{\theta} \log f(z_i|\theta_0)$ has mean 0 and variance J , so by the CLT,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} \log f(z_i|\theta_0) \xrightarrow{d} N(0, J).$$

So by Slutsky’s lemma, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to the product of $-H^{-1}$ and a $N(0, J)$ random variable, in other words

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1} J H^{-1}).$$

Recall the information equality:

$$-E[\nabla_{\theta\theta} \log f(z_i|\theta_0)] = E [(\nabla_{\theta} \log f(z_i|\theta_0))(\nabla_{\theta} \log f(z_i|\theta_0))'],$$

or

$$-H = J.$$

So the limit distribution simplifies to $N(0, J^{-1})$. Here J is the “Fisher information matrix.”

More generally, for an estimator $\hat{\theta}$, we can often figure out a (vector-valued) function $\psi(z_i)$ such that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(z_i) + o_p(1),$$

where

$$E[\psi(z_i)] = 0,$$

$$E[\psi(z_i)\psi(z_i)'] \text{ is finite.}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(z_i) \xrightarrow{d} N(0, E[\psi(z_i)\psi(z_i)']).$$

Terminology: if $\hat{\theta}$ has such a representation, we say that it is asymptotically linear, with influence function ψ . The idea is that $\pi(z_i)$ gives the “influence” of a single observation on the resulting estimate.

For the MLE, the influence function is:

$$\psi(z) = -H^{-1}\nabla_{\theta} \log f(z|\theta_0).$$

Asymptotic Normality for Extremum Estimators:

Theorem: Suppose $\hat{\theta}$ maximizes $\hat{Q}_n(\theta)$ over Θ , and $\hat{\theta} \xrightarrow{p} \theta_0$, and the following hold:

- (i) $\theta_0 \in \text{int}(\Theta)$.
- (ii) $\hat{Q}_n(\theta)$ is twice continuously differentiable in a neighborhood $\mathcal{N}(\theta_0)$ of θ_0 .
- (iii) $\sqrt{n}\nabla_{\theta}\hat{Q}_n(\theta) \xrightarrow{d} N(0, \Sigma)$.
- (iv) There is a $H(\theta)$, continuous at θ_0 , such that

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \|\nabla_{\theta\theta}\hat{Q}_n(\theta) - H(\theta)\| \xrightarrow{p} 0.$$

- (v) $H = H(\theta_0)$ is nonsingular.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1}).$$

Proof is similar to our reasoning for the special case of MLE.

For GMM, can give a slightly more useful version of the theorem, which only requires once-differentiability:

Theorem: Consider a GMM estimator with moment function $g(z_i, \theta)$, and weighting matrix \hat{W} . Suppose $\hat{\theta} \xrightarrow{p} \theta_0$, and $\hat{W} \xrightarrow{p} W$, and assume:

- (i) $\theta_0 \in \text{int}(\Theta)$.
- (ii) $g(z_i, \theta)$ is continuously differentiable in a neighborhood $\mathcal{N}(\theta_0)$ of θ_0 .
- (iii) $E[g(z_i, \theta_0)] = 0$.
- (iv) $E[\|g(z_i, \theta_0)\|^2]$ is finite.
- (v)

$$E\left[\sup_{\theta \in \mathcal{N}(\theta_0)} \|\nabla_{\theta} g(z_i, \theta)\|\right] < \infty.$$

- (vi) $G'WG$ is nonsingular, where $G = E[\nabla_{\theta} g(z_i, \theta)]$.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

where

$$\begin{aligned}\Omega &= E[g(z_i, \theta_0)g(z_i, \theta_0)'], \\ V &= (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}.\end{aligned}$$

Here is the intuition for the result and the variance formula:

In the GMM problem, the first order condition for a minimum is

$$\left[\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} g(z_i, \hat{\theta}) \right]' \hat{W} \left[\frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) \right] = 0.$$

Expand the function $g(z, \theta)$:

$$g(z_i, \hat{\theta}) = g(z_i, \theta_0) + \nabla_{\theta} g(z_i, \bar{\theta})(\hat{\theta} - \theta_0).$$

Substitute this into the first order condition:

$$\left[\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} g(z_i, \hat{\theta}) \right]' \hat{W} \left[\frac{1}{n} \sum_{i=1}^n g(z_i, \theta_0) + \left[\frac{1}{n} \sum_{i=1}^n \nabla_{\theta} g(z_i, \bar{\theta}) \right] (\hat{\theta} - \theta_0) \right] = 0.$$

Let

$$\hat{G}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} g(z_i, \theta).$$

Rearrange the first order condition to get

$$\sqrt{n}(\hat{\theta} - \theta_0) = -(\hat{G}_n(\hat{\theta})' \hat{W} \hat{G}_n(\bar{\theta}))^{-1} \hat{G}_n(\hat{\theta})' \hat{W} \frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0).$$

Then, under the conditions of the theorem,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0) \xrightarrow{d} N(0, \Omega),$$

$$\hat{W} \xrightarrow{p} W,$$

$$\hat{G}_n(\hat{\theta}) \xrightarrow{p} G,$$

$$\hat{G}_n(\bar{\theta}) \xrightarrow{p} G.$$

So

$$-(\hat{G}_n(\hat{\theta})' \hat{W} \hat{G}_n(\bar{\theta}))^{-1} \hat{G}_n(\hat{\theta})' \hat{W} \xrightarrow{p} (G'WG)^{-1}G'W,$$

and by the Slutsky lemma,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}).$$

Optimal weighting matrix: Suppose $W = \Omega^{-1}$. (That is, $\hat{W} \xrightarrow{p} \Omega^{-1}$.) Then

$$V = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1} = (G'\Omega^{-1}G)^{-1}.$$

This can be shown to be the best variance, in the sense that the difference between any other feasible V and this variance is a positive semidefinite matrix.

Estimating Variances

Basic idea is to replace expectations by sample averages, and if necessary, replace θ_0 by any consistent estimator $\hat{\theta}$.

MLE: recall $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1})$, where

$$\begin{aligned} J &= E[(\nabla_{\theta} \log f(z_i, \theta_0))(\nabla_{\theta} \log f(z_i, \theta_0))'] \\ &= -E[\nabla_{\theta\theta} \log f(z_i, \theta_0)]. \end{aligned}$$

This suggest either forming:

$$\hat{J} = \frac{1}{n} \sum_{i=1}^n (\nabla_{\theta} \log f(z_i, \hat{\theta}))(\nabla_{\theta} \log f(z_i, \hat{\theta}))',$$

or

$$\hat{H} = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta\theta} \log f(z_i, \hat{\theta}),$$

and then forming \hat{J}^{-1} or $(-\hat{H})^{-1}$ as the estimate of the asymptotic variance of the MLE. These will differ slightly, but both should be consistent under appropriate conditions.

GMM: Recall the asymptotic variance of GMM is

$$(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1},$$

where

$$G = E[\nabla_{\theta} g(z_i, \theta_0)],$$

$$W = \text{plim } \hat{W},$$

$$\Omega = E[g(z_i, \theta_0)g(z_i, \theta_0)'].$$

We can estimate

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}) g(z_i, \hat{\theta})',$$

$$\hat{G} = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} g(z_i, \hat{\theta}).$$

So the estimated variance is

$$\hat{V} = (\hat{G}' \hat{W} \hat{G})^{-1} \hat{G}' \hat{W} \hat{\Omega} \hat{W} \hat{G} (\hat{G}' \hat{W} \hat{G})^{-1}.$$

Optimal GMM: Recall that the ideal choice of the weight matrix is

$$\hat{W} \xrightarrow{p} \Omega^{-1} = (E[g(z_i, \theta_0) g(z_i, \theta_0)'])^{-1}.$$

But since θ_0 is not known, how do we implement this?

One solution (Hansen):

First, use a suboptimal \hat{W}_1 , solve the GMM problem, to get an initial estimator $\hat{\theta}_1 \xrightarrow{p} \theta_0$.

Calculate

$$\hat{\Omega}_1 = \frac{1}{n} \sum_{i=1}^n g(z_i, \hat{\theta}_1) g(z_i, \hat{\theta}_1)',$$

$$\hat{W}_2 = (\hat{\Omega}_1)^{-1}.$$

Then re-estimate θ using the new weight matrix \hat{W}_2 :

$$\hat{\theta}_2 = \arg \min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n g(z_i, \theta) \right]' \hat{W}_2 \left[\frac{1}{n} \sum_{i=1}^n g(z_i, \theta) \right].$$

Since $\hat{W}_2 \xrightarrow{p} \Omega^{-1}$, the final estimator $\hat{\theta}_2$ will have asymptotic variance $(G' \Omega^{-1} G)^{-1}$.