FINC 9311-21 Financial Econometrics Handout

Jialin Yu

1 Extremum Estimators

Let θ_0 be a vector of $k \times 1$ unknown parameters. Extremum estimators: estimators obtained by maximizing or minimizing some objective functions. Why does this make sense? You want to find the best parameter which naturally involves maximization (of gain) or minimization (of loss). Examples:

1. Maximum Likelihood: Maximizing the log-likelihood function;
2. Minimum Distance Estimators: LS, GMM, etc.

Suppose the objective function to be maximized is $Q_T(\theta)$, then the extremum estimator is defined by

$$\widehat{\theta}_T = \arg\max_{\theta \in \Theta} Q_T(\theta)$$

For well behaved objective functions, it is equivalent to say that the extremum estimator is defined by

$$\frac{\partial}{\partial \theta} Q_T(\theta)\Big|_{\theta=\widehat{\theta}_T} = 0$$

Example: OLS. What is the objective function? What is the first order condition (FOC)?

We will prove consistency/asymptotic distribution at the general level for all extremum estimators. Then, we will specialize the argument to two (very important) special cases of MLE (maximum likelihood estimation) and GMM (generalized method of moments).

1.1 Consistency of $\widehat{\theta}$

What is the issue here: you want to maximize $Q_0(\theta)$ but can maximize only its sample counterpart $Q_T(\theta)$. Under what conditions is the maximizer

of the sample counterpart "close" to the maximizer of the true objective function? For concreteness, think the OLS example.

Definition 1 *(Uniform Convergence in Probability)* $Q_T(\theta)$ *converges uniformly in probability to* $Q_0(\theta)$ *if*

$$\sup_{\theta\in\Theta}|Q_T(\theta)-Q_0(\theta)|\xrightarrow{p}0$$

Theorem 2 *If there is a function* $Q_0(\theta)$ *such that (i)* $Q_0(\theta)$ *is uniquely maximized at* θ_0 *(ii)* Θ *is compact (iii)* $Q_0(\theta)$ *is continuous (iv)* $Q_T(\theta)$ *converges uniformly in probability to* $Q_0(\theta)$*, then* $\widehat{\theta}_T\xrightarrow{p}\theta_0$*.*

Proof: For any $\varepsilon>0$, we have with probability approaching 1 that (a) $Q_T\left(\widehat{\theta}_T\right)>Q_T(\theta_0)-\varepsilon/3$ because $\widehat{\theta}_T$ is the maximizer of Q_T; (b) $Q_0\left(\widehat{\theta}_T\right)>Q_T\left(\widehat{\theta}_T\right)-\varepsilon/3$ by the uniform convergence of $Q_T(\theta)$ to $Q_0(\theta)$ (c) $Q_T(\theta_0)>Q_0(\theta_0)-\varepsilon/3$ again by the uniform convergence. Therefore, with probability approaching 1,

$$\begin{aligned}Q_0\left(\widehat{\theta}_T\right)&>Q_T\left(\widehat{\theta}_T\right)-\varepsilon/3\\&>Q_T(\theta_0)-2\varepsilon/3\\&>Q_0(\theta_0)-\varepsilon\end{aligned}$$

Let N be an open neighborhood of θ_0. θ_0 is the unique maximizer of $Q_0(\theta)$ implies $\sup_{\theta\in\Theta\cap N^C}Q_0(\theta)=Q_0(\theta^*)<Q_0(\theta_0)$ for some $\theta^*\in\Theta\cap N^C$. Choose $\varepsilon=Q_0(\theta_0)-\sup_{\theta\in\Theta\cap N^C}Q_0(\theta)$, it follows that with probability approaching 1, $Q_0\left(\widehat{\theta}_T\right)>\sup_{\theta\in\Theta\cap N^C}Q_0(\theta)$ and therefore $\widehat{\theta}_T\in N$.

Counterexample:

1. if Θ is not compact, e.g. $\Theta=[0,1)\cup\{2\}$, and assuming the objective function to be maximized is $f(x)=x$ if $x\in[0,1)$, $f(x)=1$ if $x=2$ (by the way, is this objective function continuous?);

2. If $Q_0(\theta)$ is not continuous, $f(x)=Cos(x)$ if $x\in[0,2\pi)$, $f(x)=0$ if $x=2\pi$. $\Theta=[0,2\pi]$

3. What about non-uniqueness if the maximum? This relates also to the issue of identification. What is non-identification? E.g., if you are to estimate $\beta = u - v$, can you separately identify u and v (even with infinite number of observations)?

The following lemma helps to easily verify the uniform convergence condition (iv) in practice.

Lemma 3 *(Uniform Law of Large Numbers) If the data are i.i.d., Θ is compact, $a(x_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one, and there is $d(x)$ such that $\|a(x, \theta)\| \leq d(x)$ for all $\theta \in \Theta$ and $E[d(x)] < \infty$. then $E[a(x, \theta)]$ is continuous and*

$$\sup_{\theta \in \Theta} \left\| n^{-1} \sum_{i=1}^{n} a(x_i, \theta) - E[a(x, \theta)] \right\| \xrightarrow{p} 0$$

Verify OLS consistency using both classical and this method.

1.2 Asymptotic Normality of $\widehat{\theta}$

Taylor expand $\frac{\partial}{\partial \theta} Q_T(\theta)\big|_{\theta = \widehat{\theta}_T} = 0$ around θ_0

$$0 = Q_T'\left(\widehat{\theta}_T\right) = Q_T'(\theta_0) + Q_T''(\theta_T^*)\left(\widehat{\theta}_T - \theta_0\right)$$

for some θ_T^* in between θ_0 and $\widehat{\theta}_T$ (recall the implicit assumptions involved here in this expansion).

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) = -\left[Q_T''(\theta_T^*)\right]^{-1} \sqrt{T} Q_T'(\theta_0)$$

By CLT

$$\sqrt{T} Q_T'(\theta_0) \xrightarrow{d} N(0, V(\theta_0))$$

$\widehat{\theta}_T \xrightarrow{p} \theta_0$ and θ_T^* is in between θ_0 and $\widehat{\theta}_T$ imply $\theta_T^* \xrightarrow{p} \theta_0$. Therefore,

$$-\left[Q_T''(\theta_T^*)\right]^{-1} \xrightarrow{p} -\left[Q_T''(\theta_0)\right]^{-1} \equiv -S(\theta_0)^{-1} \tag{1}$$

(this step is almost correct). Asymptotic normality follows from Slutsky

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) \xrightarrow{d} N\left(0, S\left(\theta_0\right)^{-1} V\left(\theta_0\right)\left[S\left(\theta_0\right)^T\right]^{-1}\right)$$

Theorem 4 *If the estimator satisfies* $\widehat{\theta} \xrightarrow{p} \theta$ *and (i)* $\theta_0 \in Interior\left(\Theta\right)$ *(ii)* $Q_T\left(\theta\right)$ *is twice continuously differentiable in a neighborhood* N *of* θ_0 *(iii)* $\sqrt{T}Q_T'\left(\theta_0\right) \xrightarrow{d} N\left(0, \Sigma\right)$ *(or equivalently replace this assumption with your favorite central limit theorem) (iv) there is* $H\left(\theta\right)$ *continuous at* θ_0 *such that* $\sup_{\theta \in N} \left\|Q_T''\left(\theta\right) - H\left(\theta\right)\right\| \xrightarrow{p} 0$ *(this makes the almost correct step correct); (v)* $H = H\left(\theta_0\right)$ *is nonsingular. Then*

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, H^{-1}\Sigma H^{-1}\right)$$

Condition (iv) makes (1) valid. Condition (iv) can be verified simiar to verifying uniform law of large numbers (Lemma 3).

Example: consistency and normality of OLS

2 Maximum Likelihood Estimator

Let $f(x_i, \theta)$ be the probability density of observation x_i. The maximum likelihood estimator (MLE) is defined as

$$\widehat{\theta}_T = \max_{\theta \in \Theta} \sum_{i=1}^{T} \log f(x_i, \theta)$$

I.e., you choose the parameter which is most likely to generate the observations. It is not obvious now why this is optimal. But we will show that MLE has a number of optimality properties.

Let $s_i(\theta) = \frac{\partial}{\partial \theta} \log f(x_i, \theta)$.

$$S_T(\theta) = \sum_{i=1}^{T} s_i(\theta)$$

is called the Score. The MLE estimator sets the score to 0.

$$1 = \int f(x, \theta)\, dx$$

Differentiate once,

$$\begin{aligned} 0 &= \int \frac{\partial f(x, \theta)}{\partial \theta} dx \\ &= \int \frac{\partial \log f(x, \theta)}{\partial \theta} f(x, \theta)\, dx \\ &= \int S(\theta) f(x, \theta)\, dx \end{aligned}$$

Differentiate again,

$$\begin{aligned} 0 &= \int S'(\theta) f(x, \theta)\, dx + \int S(\theta) \frac{\partial f(x, \theta)}{\partial \theta} dx \\ &= \int S'(\theta) f(x, \theta)\, dx + \int S(\theta) \frac{\partial \log f(x, \theta)}{\partial \theta} f(x, \theta)\, dx \\ &= \int S'(\theta) f(x, \theta)\, dx + \int S(\theta)^2 f(x, \theta)\, dx \end{aligned}$$

Therefore,

$$I(\theta) = -E\left[S'(\theta)\right] = E\left[S(\theta)^2\right]$$

where $I(\theta)$ is the information matrix.

$$I_i(\theta) = -E\left[s_i'(\theta)\right] = E\left[s_i(\theta)^2\right]$$

denotes the information in the i-th observation.

$$I(\theta) = E\left[S(\theta)^2\right]$$

denotes the information in the sample, and

$$\overline{I}_T = T^{-1}\sum I_i(\theta)$$

denotes the average information.

With independent sampling, the s_i' are independent of each other and $I(\theta) = \sum I_i(\theta)$. With i.i.d. sampling $I_i(\theta) = I_j(\theta) = \overline{I}_T(\theta) = T^{-1}I(\theta) \equiv \overline{I}(\theta)$.

2.1 Consistency of the MLE estimator

A feature of MLE estimator is that identification is sufficient to guarantee the log-likelihood function has a unique maximum at the true parameter θ_0.

Lemma 5 *If θ_0 is identified ($\theta \neq \theta_0$ implies $f(x,\theta) \neq f(x,\theta_0)$ with positive probability) and $E[|\log f(x,\theta)|] < \infty$ for all θ then $Q_0(\theta) = E[\log f(x,\theta)]$ has a unique maximum at θ_0.*

Proof: By the strict Jensen's inequality

$$\begin{aligned} Q_0(\theta_0) - Q_0(\theta) &= E\left[-\log \frac{f(x,\theta)}{f(x,\theta_0)}\right] \\ &> -\log E\left[\frac{f(x,\theta)}{f(x,\theta_0)}\right] \\ &= -\log \int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0)\, dx \\ &= 0 \end{aligned}$$

Theorem 6 *Suppose the observations are i.i.d. with p.d.f.* $f(x_i,\theta_0)$ *and (i)* $f(x,\theta) \neq f(x,\theta_0)$ *with positive probability if* $\theta \neq \theta_0$ *(ii)* $\theta_0 \in \Theta$ *compact (iii)* $logf(x,\theta)$ *is continuous at each* θ *with probability one (iv)* $E\left[\sup_{\theta\in\Theta} |\log f(x,\theta)|\right] < \infty$*. Then* $\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0$*.*

Proof: The theorem is proved by verifying the conditions of theorem 2 using the uniform law of large numbers.

2.2 Asymptotic Normality of the MLE estimator

Theorem 7 *Assume the conditions for theorem 6 are satisfied and (i)* $\theta_0 \in Interior(\Theta)$ *(ii)* $f(x,\theta)$ *is twice continuously differentiable with respect to* θ *and* $f(x,\theta) > 0$ *in a neighborhood* N *of* θ_0 *(iii)* $\int \sup_{\theta\in N} \left\|\frac{\partial}{\partial\theta} f(x,\theta)\right\| dx < \infty$ *and* $\int \sup_{\theta\in N} \left\|\frac{\partial^2}{\partial\theta^2} f(x,\theta)\right\| dx < \infty$ *(iv)* $I_i(\theta)$ *(the information matrix of an individual observation) is not singular (v)* $E\left[\sup_{\theta\in N} \left\|\frac{\partial^2}{\partial\theta^2} \log f(x,\theta)\right\|\right] < \infty$*. Then*

$$\sqrt{T}\left(\widehat{\theta}_{MLE} - \theta_0\right) \xrightarrow{d} N\left(0, I_i(\theta)^{-1}\right)$$