Econometrics I
TA Session 10

Jukina HATAKEYAMA[*]

June 18, 2024

Contents

[*] E-mail: u868710a@ecs.osaka-u.ac.jp

1 M–Estimation

This topic is challenging for you. If you interested in m-estimation or large sample test or so, read Newey and Mcfadden (1994). You can find this pdf in the internet.

1.1 Definition of the M–Estimator

Let $m\colon \mathbb{R}^k \times \Theta \to \mathbb{R}$ be a real–valued function of the random vector $X_i \in \mathbb{R}^k$ for $i \in \{1, \dots, n\}$ and the parameter vector $\theta \in \Theta$ where Θ denotes the **parameter space** and is a subset of $\mathbb{R}^p$. An **M–estimator** of the parameter θ solves the problem

$$\arg\max_{\theta\in\Theta} M_n(\theta) := \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta), \tag{1}$$

assuming that a solution, called $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, exists.

Remark 1.1 (Dependency of θ on X_i for $i \in \{1, \dots, n\}$)**.** What is really important is that the parameter θ which we estimate via the M–estimation depends on the observed data (and its size). However, the devision by n, while needed for the theoretical development, does not affect the maximisation problem. Also, The maximisation problem can be transformed into the minimisation one without loss of generality.

Then the parameter vector θ_0 is assumed to uniquely solve the population level problem

$$\arg\max_{\theta\in\Theta} \mathbb{E}[m(X, \theta)]. \tag{2}$$

We will focus on "**how we can translate the fact that θ_0 solves (2) into the consistency of the M–estimator $\hat{\theta}$ which solves (1)**." Since for each $\theta \in \Theta$, $m(X_i, \theta)$ for $i \in \{1, \dots, n\}$ is just an i.i.d. sequence, the **weak law of large numbers** implies that

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta) \xrightarrow[n\to\infty]{p} \mathbb{E}[m(X, \theta)], \tag{3}$$

under very weak finite moment assumptions. Thus, from the fact that $\hat{\theta}$ maximises the function on the L.H.S. of (3) and θ_0 on the R.H.S. of (3), it seems plausible that $\hat{\theta} \xrightarrow[n\to\infty]{p} \theta_0$, which means the **consistency** of $\hat{\theta}$.

To make this informal argument correct, there are essentially two issues to address. The first is identifiability of θ_0, which is purely a population level issue. The second is the sense in which the convergence in (3) happens in different values of θ in Θ.

1.2 Identifiability

We set the value θ_0 so that it solves (2). However, we do not argue that θ_0 is always the *unique* solution of (2). To obtain the unique solution, the identification requires that for all $\theta \in \Theta$, $\theta \neq \theta_0$, θ_0 be the unique solution:

$$\mathbb{E}[m(X, \theta_0)] > \mathbb{E}[m(X, \theta)] \tag{4}$$

1.3 Uniformly Convergence

(3) indicates the **pointwise convergence in probability**. However, this convergence is not sufficient for consistency, which means that it is not enough to simply invoke the usual weak law of large numbers at each $\theta \in \Theta$. Instead, **uniform convergence in probability** is sufficient.

Definition 1.1 (Uniform Weak Law of Large Numbers)**.** $\frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta)$ converges uniformly in probability to $\mathbb{E}[m(X, \theta)]$ means

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} m(X_i, \theta) - \mathbb{E}[m(X, \theta)] \right| \xrightarrow[n\to\infty]{p} 0. \tag{5}$$

We can now state a theorem concerning uniform convergence appropriate for the random sampling environment.

Theorem 1.1. Let X be a random vector taking values in $\mathbb{R}^k$, let Θ be a subset on $\mathbb{R}^d$, and let $m \colon \mathbb{R}^k \times \Theta \to \mathbb{R}$ be a real–valued function. Assume that

(a) Θ is compact;

(b) For each $\theta \in \Theta$, $m(\cdot, \theta)$ is Borel measurable on $\mathbb{R}^k$;

(c) For each $X \in \mathbb{R}^k$, $m(X, \cdot)$ is continuous on Θ;

(d) $|m(X, \theta)| \leq b(X)$ for all $\theta \in \Theta$, where b is a nonnegative function on $\mathbb{R}^k$ such that $\mathbb{E}[b(X)] < \infty$.

Then (5) holds.

1.4 Consistency for M–Estimators

According to the above setting, we have the following results.

Theorem 1.2 (Consistency of M–Estimators)**.** Under the assumptions of *Uniform Weak Law of Large Numbers*, assume that the identification assumption holds. Then a random vector, denoted as θ, solves (1), and $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. Let $\hat{\theta} = \arg\max_{\theta\in\Theta} M_n(\theta)$. Then,

$$M_n(\hat{\theta}) \geq M_n(\theta_0),$$

by definition. Also,

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}) &= M_n(\theta_0) - M(\hat{\theta}) + M(\theta_0) - M_n(\theta_0) \\ &\leq M_n(\hat{\theta}) - M(\hat{\theta}) + M(\theta_0) - M_n(\theta_0) \\ &\leq \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + M(\theta_0) - M_n(\theta_0) \end{aligned}$$

Therefore, for all $\delta > 0$,

$$M(\theta_0) - M(\hat{\theta}) > \delta \xrightarrow[n\to\infty]{p} 0 \iff \mathbb{P}\left(M(\hat{\theta}) < M(\theta_0) - \delta\right) \xrightarrow[n\to\infty]{p} 0.$$

By the assumption that for each $X \in \mathbb{R}^k$, $m(X, \cdot)$ is continuous on Θ, $\forall \varepsilon > 0, \exists \delta > 0$, such that

$$\|\theta - \theta_0\| \geq \varepsilon \implies M(\theta) < M(\theta_0) - \delta$$

Hence,

$$\mathbb{P}\left(\|\hat{\theta} - \theta_0\| \geq \varepsilon\right) = \mathbb{P}\left(M(\hat{\theta}) < M(\theta_0) - \delta\right) \leq \mathbb{P}\left(M_n(\hat{\theta}) < M(\theta_0) - \delta\right) \xrightarrow[n\to\infty]{p} 0$$

which proves the consistency of M–estimators. □

We end this section with a lemma which we use in the following proof of the asymptotic normality of an M–estimator.

Lemma 1.1. Suppose that $\hat{\theta} \xrightarrow[n\to\infty]{p} \theta_0$, and assume that a function $r\colon \mathbb{R}^k \times \Theta \to \mathbb{R}^q$ satisfies the same assumptions on $m(X, \theta)$ in Theorem 2.1. Then,

$$\frac{1}{n}\sum_{i=1}^{n} r(X_i, \hat{\theta}) \xrightarrow[n\to\infty]{p} \mathbb{E}[r(X, \theta)]. \qquad (6)$$

That is, $\frac{1}{n}\sum_{i=1}^{n} r(X_i, \hat{\theta})$ is a consistent estimator of $\mathbb{E}[r(X, \theta_0)]$.

1.5 Asymptotic Normality of M–Estimators

Under additional assumptions on the objective function, we can also show that M–estimators are asymptotically normally distributed (and converge at the rate $\sqrt{n}$. It turns out that continuity over the parameter space does not ensure asymptotic normality.

Theorem 1.3 (Asymptotic Normality of M–Estimators)**.** In addition to the assumptions in Theorem 2.2, assume

(a) θ_0 is in the interior of Θ;

(b) $s(X, \cdot)$ is continuously differentiable on the interior of Θ for all $X \in \mathbb{R}^k$;

(c) Each element of $H(X, \theta)$ is bounded in absolute value by a function $b(X)$, where $\mathbb{E}[b(X)] < \infty$;

(d) $H \equiv \mathbb{E}[H(X, \theta_0)]$ is positive definite;

(e) $\mathbb{E}[s(X, \theta_0)] = \mathbf{0}$;

(f) Each element of $s(X, \theta_0)$ has finite second moment.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, H^{-1}JH^{-1}\right) \qquad (7)$$

where $H \equiv \mathbb{E}[H(X, \theta_0)]$ and $J \equiv \mathbb{E}[s(X, \theta_0)s(X, \theta_0)'] = \mathbb{V}[s(X, \theta_0)]$.

Proof. Assume that θ_0 is in the interior of Θ, which means that Θ must have nonempty interior; this assumption is true in most applications. Then, since $\hat{\theta} \xrightarrow[n\to\infty]{p} \theta_0$, $\hat{\theta}$ is in the interior of Θ with probability approaching one. If $m(X, \cdot)$ is continuously differentiable on the interior Θ, then (with probability approaching one) $\hat{\theta}$ solves the first order condition

$$\sum_{i=1}^{n} s(x_i, \hat{\theta}) = \mathbf{0}, \tag{8}$$

where $s(X, \theta)$ is the $p \times 1$ vector of partial derivatives of $m(X, \theta)$ with respect to θ:

$$s(X,\theta)' = \nabla_\theta m(X,\theta) \equiv \left[\frac{\partial m(X,\theta)}{\partial \theta_1}, \frac{\partial m(X,\theta)}{\partial \theta_2}, \ldots, \frac{\partial m(X,\theta)}{\partial \theta_p}\right]$$

(That is, $s(X, \theta)$ is the transpose of the gradient of $m(X, \theta)$.) We call $s(X, \theta)$ the **score of the objective function** $m(X, \theta)$.

If $m(X, \cdot)$ is twice continuously differentiable (with respect to θ), then each row of the left–hand side of (8) can be expanded about θ_0 in a mean–value expansion:

$$\sum_{i=1}^{n} s(X_i, \hat{\theta}) = \sum_{i=1}^{n} s(X_i, \theta_0) + \left(\sum_{i=1}^{n} \ddot{H}_i\right)(\hat{\theta} - \theta_0) \tag{9}$$

Here we note the following theorem related to the mean–value expansion.

Theorem 1.4. If f is a real continuous function on $[a, b]$ which is differentiable in (a, b), then there is a point $x \in (a, b)$ at which

$$f(b) - f(a) = (b - a)f'(x).$$

The notation $\ddot{H}_i$ denotes the $p \times p$ **Hessian of the objective function** $m(X_i, \cdot)$ with respect to θ, but with each row of $H_i \equiv H(X_i, \theta) = \partial^2 m(X_i, \theta)/\partial\theta\partial\theta' \equiv \nabla_\theta^2 m(X_i, \theta)$ evaluated at a different mean value. Each of the p mean values is on the line segment between θ_0 and $\hat{\theta}$, say $\tilde{\theta}_i$, and we know that each must converge in probability to θ_0 (since each is "trapped" between θ_0 and $\hat{\theta}$).

Conbining (8) and (9) and multiplying through by $1/\sqrt{n}$ gives

$$\mathbf{0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i, \theta_0) + \left(\frac{1}{n}\sum_{i=1}^{n} \ddot{H}(X_i, \tilde{\theta}_i)\right)\sqrt{n}(\hat{\theta} - \theta_0)$$

Now, we can apply Lemma 2.3 to get

$$\frac{1}{n}\sum_{i=1}^{n} \ddot{H}(X_i, \tilde{\theta}_i) \xrightarrow[n\to\infty]{p} \mathbb{E}[H(X, \theta_0)]$$

(under some moment conditions). If $H \equiv \mathbb{E}[H(X, \theta_0)]$ is nonsingular, then $n^{-1}\sum_{i=1}^{n} \ddot{H}_i$ is non–singular with probability approaching one and $[n^{-1}\sum_{i=1}^{n} \ddot{H}(X_i, \tilde{\theta}_i)]^{-1} \xrightarrow{p} H^{-1}$ (by **continuous mapping theorem**). Therefore, we can write

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left(\frac{1}{n}\sum_{i=1}^{n} \ddot{H}(X_i, \tilde{\theta}_i)\right)^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(X_i, \theta_0)\right]$$

As we will show, $\mathbb{E}[s(X_i,\theta_0)] = \mathbf{0}$. Therefore, $n^{-1/2}\sum_{i=1}^n s(X_i,\theta_0)$ generally satisfies the **Lindeberg-Lévy central limit theorem**, because it is the average of i.i.d. random vectors with zero mean, multiplied by the usual $\sqrt{n}$. Since $o_p(1)\cdot O_p(1) = o_p(1)$, we have

$$\sqrt{n}(\hat{\theta}-\theta_0) = \left(\frac{1}{n}\sum_{i=1}^n \ddot{H}(X_i,\tilde{\theta}_i)\right)^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta_0)\right] + o_p(1). \tag{10}$$

This is an important equation. It shows that $\sqrt{n}(\hat{\theta}-\theta_0)$ inherits its limiting distribution from the average of the scores, evaluate at θ_0. The matrix H^{-1} simply acts as a linear transformation.

(10) allows us to derive the **first–order asymptotic distribution** of $\hat{\theta}$. Higher order representations attempt to reduce the error in the $o_p(1)$ term in (10); such derivations are much more complicated than (10) and are beyond the scope of this course. We have essentially proved the Theorem 2.3. □

A key component of Theorem 12.3 is that the score evaluated at θ_0 has expected value zero. In many applications, including NLS, we can show this result directly. But it is also useful to know that it holds in the abstract M-estimation framework, at least if we can interchange the expectation and the derivative. To see this point, note that, if θ_0 is in the interior of Θ, and $\mathbb{E}[m(X,\theta)]$ is differentiable for $\theta \in \operatorname{int}\Theta$, then

$$\nabla_\theta \mathbb{E}[m(X,\theta)]|_{\theta=\theta_0} = \mathbf{0}. \tag{11}$$

where ∇_θ denotes the gradient with respect to θ. Now, if the derivative and expectations operator can be interchanged (which is the case quite generally), then (11) implies

$$\mathbb{E}[\nabla_\theta m(X,\theta_0)] = \mathbb{E}[s(X,\theta_0)] = \mathbf{0}. \tag{12}$$

A similar argument shows that, in general, $\mathbb{E}[m(X,\theta_0)]$ is positive semidefinite. If θ_0 is identified, $\mathbb{E}[H(X,\theta_0)]$ is positive definite.

2 Consistency and Asymptotic Normality for the Maximum Likelihood Estimator

If we set

- setting $m(X_i,\theta) := \log p_\theta(X_i)$, that is,

$$M_n(\theta) := L_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i);$$

- $\mathbb{L}(\theta) = \mathbb{E}\left[\mathbb{L}_n(\theta)\right] = \mathbb{E}\left[\log p_\theta(X)\right]$,

in the above setting, then we can state the same theorem as mentioned above.

2.1 Consistency for the Maximum Likelihood Estimator

Suppose

$$\mathbb{L}(\theta) = \mathbb{E}\left[\mathbb{L}_n(\theta)\right] = \mathbb{E}\left[\log p_\theta(X)\right]$$

exists for $\theta \in \mathbb{R}^d$. Then we obtain the sollowing theorem.

Theorem 2.1. Suppose

(i) $\mathbb{L}(\theta)$ is uniquely maximised at θ_0, idest

$$\forall\epsilon > 0, \qquad \sup_{\theta:\, \|\theta-\theta_0\|\geq 0} \mathbb{L}(\theta) < \mathbb{L}(\theta_0);$$

(ii) Θ is compact;

(iii) $\mathbb{L}(\theta)$ is continuous;

(iv) $\sup_{\theta\in\Theta}|\mathbb{L}_n(\theta) - \mathbb{L}(\theta)| \xrightarrow[n\to\infty]{\mathbb{P}} 0$,

then $\hat{\theta} \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0$.

Under the same assumption, we can derive the following theorem.

Theorem 2.2. Suppose $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ and

(i) θ_0 belongs to the interior of Θ;

(ii) $\mathbb{L}_n(\theta)$ is twice continuously differentiable;

(iii) $\sqrt{n}\nabla_\theta\mathbb{L}_n(\theta_0) \xrightarrow[n\to\infty]{d} N_{\mathbb{R}^d}\left(\mathbf{0}, \Sigma\right)$;

(iv) $H(\theta) = \mathbb{E}[\nabla^2_{\theta\theta'}\log p_\theta(X)]$ is continuous at θ_0 and

$$\sup_{\theta:\, \|\theta-\theta_0\|\leq\delta} \left|\nabla^2_{\theta\theta'}\mathbb{L}_n(\theta) - H(\theta)\right| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \quad \text{with } \delta > 0;$$

(v) $H = H(\theta)$ is nonsingular,

then

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow[n\to\infty]{d} N_{\mathbb{R}^d}\left(\mathbf{0}, H^{-1}JH^{-1}\right). \tag{13}$$

3 Non–linear Optimization Procedure

In this section, we review some concepts related to the non–linear optimization problem. There are situations where the solution can not be obtained in closed form. In such a situation, we solve the optimal solution derived from the principal problem by means of

iterative algorithm instead of searching for the analytic solution. Here we describe the Newton–Raphson method and scoring method, which is widely used in common.

3.1 Newton–Raphson Method

From the first–order Taylor series expansion around $\beta = \beta^*$, we have

$$0 = \nabla_\beta \log L(\beta) \approx \nabla_\beta \log L(\beta^*) + \nabla^2_{\beta\beta'} \log L(\beta^*)(\beta - \beta^*)$$

Then, by the **mean–value theorem (expansion)**,

$$\nabla^2_{\beta\beta'} \log L(\overline{\beta})(\beta - \beta^*) = -\nabla_\beta \log L(\beta^*)$$

holds. Thus, assuming that the Hessian matrix is positive definite yields

$$\beta - \beta^* = -\left(\nabla^2_{\beta\beta'} \log L(\overline{\beta})\right)^{-1} \nabla_\beta \log L(\beta^*)$$

This equation yields the following algorithm called **Newton–Raphson Method**.

Algorithm (Newton–Raphson Method)**.**

$$\beta^{(j+1)} = \beta^{(j)} - \left(\nabla^2_{\beta\beta'} \log L(\beta^{(j)})\right)^{-1} \nabla_\beta \log L(\beta^{(j)})$$

3.2 Scoring Method

If we take expectation on second derivative of likelihood function, the method is known as the **method of scoring**.

Algorithm (Scoring Method)**.**

$$\beta^{(j+1)} = \beta^{(j)} - \left(\mathbb{E}\left[\nabla^2_{\beta\beta'} \log L(\beta^{(j)})\right]\right)^{-1} \nabla_\beta \log L(\beta^{(j)})$$

Note that

$$I(\theta) := -\mathbb{E}\left[\nabla^2_{\beta\beta'} log L(\beta)\right]$$

is the **Fisher information** matrix.

Appendix

A Pointwise Convergence and Uniform Convergence

A.1 Pointwise Convergence

The **pointwise convergence** is defined as below.

Definition A.1 (Pointwise Convergence)**.** Suppose $\{f_n\}$ for $n \in \{1, 2, \ldots\}$ is a sequence os funcitons defined on a set E, and suppose that the sequence of numbers $\{f_n(x)\}$ converges for every $x \in E$. We can then define a function f by

$$\lim_{n\to\infty} f_n(x) = f(x), \tag{14}$$

for every $x \in E$.

Under these circumstances we say that $\{f_n\}$ converges on E and that f is the *limit*, or the *limit function*, of $\{f_n\}$. Sometimes we shall use a more descriptive terminology and shall say that "$\{f_n\}$ converges to f **pointwise** on E" if (14) holds.

A.2 Uniform Convergence

The definition of the **uniform convergence** is given as follows.

Definition A.2. We say that a sequence functions f_n for $n \in \{1, 2, \ldots\}$ **converges uniformly** on E to a function f if for every $\varepsilon > 0$, there is an integer N such that $n \geq N$ implies

$$|f_n(x) - f(x)| < \varepsilon, \text{ for any } x \in \mathbb{R}$$

for all $x \in E$.

Here we exhibit a well–known fact.

Theorem A.1. Suppose

$$\lim_{n\to\infty} f_n(x) = f(x).$$

for every $x \in E$. Put

$$M_n = \sup_{x\in E} |f_n(x) - f(x)|.$$

Then $f_n \to f$ uniformly on E if and only if $M_n \to 0$ as $n \to \infty$.

Example A.1. Consider the following sequence of functions:

$$f_n(x) = x\mathrm{e}^{-nx}$$

on $I = [0, \infty)$. First, we look at pointwise convergence: $f_n(0) = 0$ and for $x > 0$ we have that $f_n(x) \to 0$ as $n \to \infty$. Thus $f_n(x) \to 0$ pointwise on I. We now need to investigate uniform convergence. Since the limit function $f(x) = 0$ we have

$$\sup_{x\in[0,\infty)} \left|x\mathrm{e}^{-nx}\right| = \sup_{x\in[0,\infty)} x\mathrm{e}^{-nx}$$

since $f_n(x) \geq 0$ for $x \geq 0$. Now we have $f_n'(x) = (1 - nx)\mathrm{e}^{-nx}$ for $x > 0$ and we see that $f_n'(x) = 0$ when $x = 1/n$. Further, $f_n'(x) < 0$ when $0 < x < 1$, and $f_n'(x) > 0$ when $x > 1$, so we conclude that $f_n(x)$ has a maximum at $x = 1/n$ and hence

$$\sup_{x\in[0,\infty)} x\mathrm{e}^{-nx} = f_n(\frac{1}{n}) = \frac{1}{n}\mathrm{e}^{-1} = \frac{1}{n\mathrm{e}} \to 0$$

as $n \to \infty$. So we see that the sequence of functions $f_n(x) = x\mathrm{e}^{-nx}$ converges uniformly to 0 on the interval $I = [0, \infty)$. □

Example A.2. Consider the following sequence of functions:

$$f_n(x) = \frac{nx}{1 + n^2x^2}$$

on $I = (0, \infty)$. This function converges pointwise to zero since

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{nx}{1 + n^2x^2} = \lim_{n\to\infty} \frac{x}{2nx^2} = \lim_{n\to\infty} \frac{1}{2nx} = 0$$

as $n \to \infty$. We now need to investigate the uniform convergence. For any $\varepsilon < 1/2$, when $x = 1/n$,

$$\left| f_n\left(\frac{1}{n}\right) - f\left(\frac{1}{n}\right) \right| = \frac{1}{2} - 0 > \varepsilon$$

So we see that the sequence of functions $\{f_n(x)\}$ does not converge uniformly to 0 on the interval $I = (0, \infty)$. □

B Interior point

Here we mention some concepts concerned with basic set theory which we have used in the above sections.

Definition B.1. Let X be a metric space. All points and sets mentioned below are understood to be elements and subsets of X.

(a) A **neighborhood** of a point p is a set $N_r(p)$ consisting of all points q such that $d(p, q) < r$. The number r is called the **radius** of $N_r(p)$.

(b) A point p is an **interior** point of E if there is a neighborhood N of p such that $N \subset E$.