LECTURE 8 NOTES

1. Consistency of GMM estimators. We turn our attention to the consistency of GMM estimators. Since GMM estimators are extremum estimators, it is possible to adapt the proof of Lecture 7, Theorem 2.5 to establish their consistency. Recall a GMM estimator minimizes

$$Q_n(\theta) := \big\|\tfrac{1}{n}\textstyle\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta)\big\|_{\widehat{W}_n}^2,$$

where $\widehat{W}_n$ converges in probability to some $W \succ 0$. We expect it to converge to the minimizer of the population criterion

$$Q(\theta) := \big\|\mathbf{E}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta)\big]\big\|_W^2.$$

We skip the technical part that establishes the convergence of the minimizer from the convergence of the minimum.

Theorem 1.1. *Let $\{\mathbf{x}_i\}$ be* i.i.d. *random variables, and $f_{\theta^*}(x)$ be their density for some θ^* in a compact parameter space Θ. If*

1. $\mathbf{E}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta)\big] = 0$ *if and only if* $\theta = \theta^*$,
2. Q_n *is continuous,*
3. $\sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$,

then the GMM estimator is consistent; i.e. $\arg\max_{\theta\in\Theta} Q_n(\theta) \xrightarrow{p} \theta^*$.

Proof. By an argument similar to that in the proof of Lecture 7, Theorem 2.5,

$$\begin{aligned} Q(\hat{\theta}_n) &= Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + \underbrace{Q_n(\hat{\theta}_n) - Q_n(\theta^*)}_{\le 0 \text{ since } \hat{\theta}_n \text{ minimizes } Q_n(\theta)} + Q_n(\theta^*) \\ &\le \big|Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n)\big| + Q_n(\theta^*), \end{aligned}$$

The first term is $o_P(1)$ by the uniform convergence assumption. The second term is also $o_P(1)$ because

1. $Q(\theta^*) = \big\|\mathbf{E}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta^*)\big]\big\|_W^2 = 0$ by assumption,
2. $Q_n(\theta^*) \xrightarrow{p} Q(\theta^*)$ by the LLN.

Thus $Q(\hat{\theta}_n) \xrightarrow{p} 0$ as claimed. □

The first assumption is an identifiability assumption. It ensures the minimizer of $Q(\theta)$ is unique. The second assumption is a uniform convergence assumption. It is implied by a (multivariate) ULLN.

LEMMA 1.2. *As long as*

1. $\sup_{\theta\in\Theta}\big\|\frac{1}{n}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta) - \mathbf{E}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta)\big]\big\|_2 \xrightarrow{p} 0$,
2. $\widehat{W}_n \xrightarrow{p} W$ *for some* $W \succ 0$,
3. $\sup_{\theta\in\Theta}\mathbf{E}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta)\big] < \infty$,

the quadratic distance

$$\sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0 \text{ for any } W \succ 0. \tag{1.1}$$

2. Asymptotic normality of the MLE. Consistency ensures a sequence of estimators converges to the target as the sample size grows. However, the rate of convergence is unknown. As we shall see, in the usual asymptotic setup, the rate is typically $\frac{1}{\sqrt{n}}$. Thus $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges to a non-degenerate random variable.

THEOREM 2.1. *Let* $\{\mathbf{x}_i\}$ *be* i.i.d. *random variables with density* $f_{\theta^*}(x)$ *for some* $\theta^* \in \text{int}(\Theta)$. *Assume*

1. $\ell_x(\theta)$ *is twice-continuously differentiable for any* $x \in \mathcal{X}$.
2. $\frac{1}{\sqrt{n}}\sum_{i\in[n]}\nabla\ell_{\mathbf{x}_i}(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, I(\theta^*)\big)$, *where* $I(\theta) := \mathbf{var}_{\theta^*}\big[\nabla\ell_{\mathbf{x}_1}(\theta)\big]$.
3. $\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]$ *is non-singular.*

If the MLE is consistent, i.e. $\hat{\theta}_n \xrightarrow{p} \theta^*$, *then it is asymptotically normal:*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1} I(\theta^*)\,\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1}\big),$$

where $I(\theta^*) = \mathbf{E}_{\theta^*}\big[\nabla\ell_{\mathbf{x}_1}(\theta^*)\nabla\ell_{\mathbf{x}_1}(\theta^*)^T\big]$ *is the Fisher information.*

Before we prove Theorem 2.1, we elaborate on its statement. The gradient of the log-likelihood function

$$\nabla_\theta\big[\log f_\theta(\mathbf{x}_1)\big]$$

is called the *score function*, and its variance

$$I(\theta) := \mathbf{var}_\theta\big[\nabla_\theta\big[\log f_\theta(\mathbf{x}_1)\big]\big]$$

is the Fisher Information. As the Fisher information increases, the asymptotic variance of the MLE decreases.

The assumption that the score function vanishes in expectation is usually justified by exchanging differentiation and expectation:

$$\begin{aligned}
\mathbf{E}_\theta\big[\nabla_\theta \log f_\theta(\mathbf{x}_1)\big] &= \int_{\mathcal{X}} \nabla_\theta \log f_\theta(x_1) f_\theta(x_1) dx_1 \\
&= \int_{\mathcal{X}} \nabla_\theta f_\theta(x_1) dx_1 && (\nabla_\theta \log f_\theta(x_1) = \tfrac{\nabla_\theta f_\theta(x_1)}{f_\theta(x_1)}) \\
&= \nabla_\theta\Big[\textstyle\int_{\mathcal{X}} f_\theta(x_1) dx_1\Big] \\
&= \nabla_\theta 1 = 0.
\end{aligned}$$

Although it is a valid assumption for "most" densities, there are non-pathological counterexamples; e.g. it is possible to show the $\mathrm{unif}(0,\theta)$ density does not satisfy the assumption.

Proof. Since $\hat{\theta}_n$ maximizes $\frac{1}{n}\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\theta)$, we necessarily have

$$0 = \tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla \ell_{\mathbf{x}_i}(\hat{\theta}_n),$$

which, by a Taylor expansion of $\frac{1}{n}\sum_{i\in[n]} \ell_{\mathbf{x}_i}(\hat{\theta}_n)$, is

$$= \tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla \ell_{\mathbf{x}_i}(\theta^*) + \tfrac{1}{n}\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\tilde{\theta}_n)(\hat{\theta}_n - \theta^*)$$

for some $\tilde{\theta}_n$ on the segment between θ^* and $\hat{\theta}_n$. Rearranging,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \big(-\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\tilde{\theta}_n)\big)^{-1}\big(\tfrac{1}{\sqrt{n}}\sum_{i\in[n]} \nabla \ell_{\mathbf{x}_i}(\theta^*)\big).$$

Since $\theta^* \in \mathrm{int}(\Theta)$, $\mathbf{E}_P\big[\nabla \ell_{\mathbf{x}_1}(\theta^*)\big] = 0$. By the CLT,

$$\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in[n]} \nabla \ell_{\mathbf{x}_i}(\theta^*) \xrightarrow{d} \mathcal{N}(0, I(\theta^*)).$$

By Taylor's theorem,

$$\begin{aligned}
&-\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\tilde{\theta}_n) \\
&= -\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\theta^*) + \Big(\tfrac{1}{n}\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\theta^*) - \tfrac{1}{n}\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\tilde{\theta}_n)\Big) \\
&= -\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\theta^*) + o\big(\|\tilde{\theta}_n - \theta^*\|_2\big).
\end{aligned}$$

Since $\hat{\theta}_n \xrightarrow{p} \theta^*$, $\|\tilde{\theta}_n - \theta^*\|_2 \sim o_P(1)$, which implies

$$\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\tilde{\theta}_n) - \tfrac{1}{n}\sum_{i\in[n]} \nabla^2 \ell_{\mathbf{x}_i}(\theta^*) \sim o_P(1).$$

By the LLN,

$$\tfrac{1}{n}\textstyle\sum_{i\in[n]}\nabla^2\ell_{\mathbf{x}_i}(\theta^*)\xrightarrow{p}\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big].$$

We put the pieces together to deduce

$$-\tfrac{1}{n}\textstyle\sum_{i\in[n]}\nabla^2\ell_{\mathbf{x}_i}(\tilde\theta_n)\xrightarrow{p}-\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big],$$

by the continuity of $\nabla^2\ell_{\mathbf{x}_i}$ and the LLN. Since $\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]$ is nonsingular, we appeal to the continuous mapping theorem to deduce

$$-\Big(\tfrac{1}{n}\textstyle\sum_{i\in[n]}\nabla^2\ell_{\mathbf{x}_i}(\tilde\theta_n)\Big)^{-1}\xrightarrow{p}-\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1},$$

We combine the two limits by Slutsky's theorem to deduce

$$\sqrt{n}(\hat\theta_n-\theta^*)\xrightarrow{d}\mathcal{N}\big(0,\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1}I(\theta^*)\,\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1}\big).$$

□

Under additional regularity conditions on ℓ_x, we are free to interchange differentiation and expectation:

$$\nabla_{\theta_i}\mathbf{E}_\theta\big[\nabla_{\theta_j}\big[\log f_\theta(\mathbf{x}_1)\big]\big]=\mathbf{E}_\theta\big[\nabla_{\theta_i,\theta_j}\big[\log f_\theta(\mathbf{x}_1)\big]\big].$$

Under additional regularity conditions, it is possible to show that

$$I(\theta)=-\mathbf{E}_\theta\big[\nabla_\theta^2\big[\log f_\theta(\mathbf{x})\big]\big].$$

The additional regularity conditions essentially allow us to interchange differentiation and integration.

Lemma 2.2. *If*

$$\begin{aligned}&\nabla_i\mathbf{E}_\theta\big[\nabla_j\big[\log f_\theta(\mathbf{x})\big]\big]\\&\quad=\mathbf{E}_\theta\big[\nabla_{i,j}\log f_\theta(\mathbf{x})\big]+\int_{\mathcal{X}}\nabla_j\big[\log f_\theta(\mathbf{x})\big]\nabla_i f_\theta(x)dx\end{aligned}$$

for any $i,j\in[p]$, the Fisher Information is also given by

$$I(\theta)=-\mathbf{E}_\theta\big[\nabla_\theta^2\big[\log f_\theta(\mathbf{x})\big]\big].$$

Proof. We differentiate $\mathbf{E}_\theta\big[\nabla_j\log f_\theta(\mathbf{x})\big]=0$ to obtain

$$0=\nabla_i\mathbf{E}_\theta\big[\nabla_j\log f_\theta(\mathbf{x})\big],$$

which, by assumption, is

$$= \mathbf{E}_\theta\big[\nabla_{i,j}\log f_\theta(\mathbf{x})\big] + \int_{\mathcal{X}} \nabla_j\big[\log f_\theta(\mathbf{x})\big]\nabla_i f_\theta(x)dx. \tag{2.1}$$

Since $\nabla_i \log f_\theta(x) = \frac{\nabla_i f_\theta(x)}{f_\theta(x)}$, the second term is

$$\int_{\mathcal{X}} \nabla_j\big[\log f_\theta(\mathbf{x})\big]\nabla_i f_\theta(x)dx = \mathbf{E}_\theta\big[\nabla_i\big[\log f_\theta(\mathbf{x})\big]\nabla_j\big[\log f_\theta(\mathbf{x})\big]\big]. \tag{2.2}$$

We plug (2.2) into (2.1) and rearrange to conclude

$$-\mathbf{E}_\theta\big[\nabla_{i,j}\log f_\theta(\mathbf{x})\big] = \mathbf{E}_\theta\big[\nabla_i\big[\log f_\theta(\mathbf{x})\big]\nabla_j\big[\log f_\theta(\mathbf{x})\big]\big] = \big[I(\theta)\big]_{i,j}.$$

□

Under the conditions of Lemma 2.2, the asymptotic variance of the MLE simplifies to

$$\begin{aligned}&\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1} I(\theta^*)\,\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1} \\ &\quad = -\,\mathbf{E}_{\theta^*}\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1} \\ &\quad = I(\theta^*)^{-1}.\end{aligned}$$

We observe that the key ingredients of Theorem 2.1 are

1. the consistency of the MLE,
2. the asymptotic normality of the re-normalized *score function*:

$$\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in[n]}\nabla\ell_{\mathbf{x}_i}(\theta) \xrightarrow{d} \mathcal{N}\big(0, I(\theta^*)\big]\big),$$

3. the validity of the second-order Taylor expansion of $\ell_{\mathbf{x}}(\theta)$ at θ^*.

The preceding proof is classical. Modern proofs replace the laundry list of regularity conditions on $\ell_x(\theta)$ by a *stochastic equicontinuity* condition. The modern approach allows us to weaken the regularity conditions on $\ell_x(\theta)$ to include non-smooth log-likelihoods. We refer to Chapter 5 of Van der Vaart (2000) for an account of the modern approach.

EXAMPLE 2.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$. *We showed that the log-likelihood is*

$$\begin{aligned}\ell_{\mathbf{x}}(p) &= \sum_{i\in[n]}\mathbf{x}_i\log\frac{p}{1-p} + n\log(1-p) \\ &= \mathbf{t}\log\frac{p}{1-p} + n\log(1-p),\end{aligned}$$

and the MLE is $\bar{\mathbf{x}}$*. The Fisher information is*

$$\begin{aligned}\mathbf{var}_p\big[\nabla\ell_{\mathbf{x}_1}(p)\big] &= \mathbf{var}_p\big[\mathbf{x}_1\big(\tfrac{1}{p}+\tfrac{1}{1-p}\big)-\tfrac{1}{1-p}\big]\\ &= \big(\tfrac{1}{p}+\tfrac{1}{1-p}\big)^2\,\mathbf{var}_p\big[\mathbf{x}_i\big]\\ &= \big(\tfrac{1}{p}+\tfrac{1}{1-p}\big)^2 p(1-p)\\ &= \tfrac{1}{p(1-p)}.\end{aligned}$$

Thus the asymptotic distribution of the MLE is

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}\big(0, p(1-p)\big).$$

Since $\hat{p}_n = \frac{1}{n}\sum_{i\in[n]}\mathbf{x}_i$*, by the CLT,*

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} \mathcal{N}\big(0, p(1-p)\big).$$

If we wish to estimate the odds ratio $\log\frac{p}{1-p}$ *instead, the MLE, by equivariance, is* $\log\frac{\hat{p}}{1-\hat{p}}$*. By the delta method, its asymptotic distribution is*

$$\sqrt{n}\big(\log\tfrac{\hat{p}}{1-\hat{p}} - \log\tfrac{p}{1-p}\big) \xrightarrow{d} \mathcal{N}\big(0, \underbrace{\big(\tfrac{1}{p}+\tfrac{1}{1-p}\big)^2 p(1-p)}_{\frac{1}{p(1-p)}}\big).$$

The variance of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is the *asymptotic variance* of $\hat{\theta}_n$. We remark that it is generally not equal to

$$\lim_{n\to\infty} n\,\mathbf{var}\big[\hat{\theta}\big].$$

In fact, it is possible to show that the latter is generally at least the asymptotic variance.

Taking a step back, we remark that the proofs in the preceding section obscures the intuition behind Theorem 2.1. The crux of the argument is, in asymptopia, the MLE is essentially a sum of *i.i.d.* random variables:

$$\hat{\theta}_n = \theta^* - \frac{1}{\sqrt{n}}\sum_{i\in[n]} I(\theta^*)\nabla\ell_{\mathbf{x}_i}(\theta^*) + o_P(n^{-1/2})$$

By the CLT, we expect it to be asymptotically normal. Most of the technical details is in showing the remainder term is negligible.

To wrap up, we study the asymptotic distribution of the MLE when the model is misspecified. Surprisingly, it remains asymptotically normal.

THEOREM 2.4. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} P$ *and* $\theta^* := \arg\max_{\theta\in\Theta} \mathbf{E}_P\big[\ell_{\mathbf{x}}(\theta)\big]$, *where* $\ell_x(\theta) := \log f_\theta(x)$ *is the log-likelihood of a parametric model. Assume*

1. ℓ_x *is twice-continuously differentiable for any* $x \in \mathcal{X}$,
2. $\frac{1}{\sqrt{n}}\sum_{i\in[n]} \nabla\ell_{\mathbf{x}_i}(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{var}_P\big[\nabla\ell_{\mathbf{x}_1}(\theta^*)\big]\big)$,
3. $\mathbf{E}_P\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]$ *is non-singular.*

If the MLE is consistent, i.e. $\hat{\theta}_n \xrightarrow{p} \theta^*$, *then it is asymptotically normal:*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{E}_P\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1} I(\theta^*)\, \mathbf{E}_P\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]^{-1}\big),$$

where $I(\theta) := \mathbf{var}_P\big[\nabla\ell_{\mathbf{x}}(\theta)\big]$.

We hasten to remark that θ^* is no longer the true parameter; it corresponds to the best approximation of P in the parametric model. Since the model is mis-specified, $\mathbf{E}_P\big[\nabla^2\ell_{\mathbf{x}_1}(\theta^*)\big]$ is not the Fisher information, so the asymptotic variance does not simplify to $I(\theta^*)^{-1}$.

References.

VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

YUEKAI SUN
BERKELEY, CALIFORNIA
DECEMBER 4, 2015