LECTURE 9 NOTES

1. Asymptotic normality of GMM estimators. Recall GMM estimators minimize the quadratic form

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

Its first two derivatives are

$$\begin{aligned}\nabla Q_n(\theta) &= \big(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla g_{\mathbf{x}_i}(\theta)\big)\widehat{W}_n\big(\tfrac{1}{n}\textstyle\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta)\big)\\ \nabla^2 Q_n(\theta) &= \big(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla^2 g_{\mathbf{x}_i}(\theta^*)\big)\widehat{W}_n\big(\tfrac{1}{n}\textstyle\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta)\big)\\ &\quad + \big(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla g_{\mathbf{x}_i}(\theta)\big)\widehat{W}_n\big(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \nabla g_{\mathbf{x}_i}(\theta)\big)^T.\end{aligned}$$

There is a similar asymptotic normality result for GMM estimators. The key ingredients are

1. the consistency of GMM estimators,
2. the asymptotic normality of $\frac{1}{\sqrt{n}}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta^*)$,
3. the validity of the second-order Taylor expansion of g_x at θ^*.

The proof even mimics the proof of Lecture 8, Theorem 2.1.

THEOREM 1.1. *Let* $\{\mathbf{x}_i\}$ *be* i.i.d. *random variables distributed according to* F_{θ^*}. *If*

1. $\frac{1}{\sqrt{n}}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{var}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta^*)\big]\big)$
2. g_x *is continuously differentiable,*
3. $\mathbf{E}_{\theta^*}\big[\nabla^2 Q(\theta^*)\big]$ *is non-singular.*

If the GMM estimator is consistent, then

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\big(0, (GWG^T)^{-1}GWJ(GW)^T(GWG^T)^{-1}\big),$$

where $G := \mathbf{E}_{\theta^*}\big[\nabla g_{\mathbf{x}_1}(\theta^*)\big]$ *and* $J := \mathbf{var}_{\theta^*}\big[g_{\mathbf{x}_1}(\theta^*)\big]$.

PROOF SKETCH. Since $\hat{\theta}_n$ minimizes Q_n,

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

which, by Taylor's theorem, is

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

Rearranging,

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

The technical part of the proof consists of showing

1. $\big(\frac{1}{n}\sum_{i\in[n]} \nabla g_{\mathbf{x}_i}(\hat{\theta}_n)\big)\widehat{W}_n\big(\frac{1}{\sqrt{n}}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta^*)\big) \xrightarrow{d} \mathcal{N}\big(0, (GW)J(GW)^T\big)$, and
2. $\Big(\big(\frac{1}{n}\sum_{i\in[n]} \nabla g_{\mathbf{x}_i}(\hat{\theta}_n)\big)\widehat{W}_n\big(\frac{1}{n}\sum_{i\in[n]} \nabla g_{\mathbf{x}_i}(\tilde{\theta}_n)\big)^T \Big)^{-1} \xrightarrow{p} (GWG^T)^{-1}$

by appealing to the consistency of $\hat{\theta}_n$ and the usual barrage of convergence theorems. We skip the technical details here. We combine the two limits by Slutsky's theorem to deduce

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\big(0, (GWG^T)^{-1}GWJ(GW)^T(GWG^T)^{-1}\big),$$

where we plugged in GWG^T for $\nabla^2 Q(\theta^*)$. □

Although the choice of W does not affect the consistency of GMM estimators, it affects the asymptotic variance of the estimator. As we shall see, choosing $W = J^{-1}$ minimizes the asymptotic variance.

LEMMA 1.2. *For any $W \succ 0$,*

$$(GWG^T)^{-1}GWJ(GW)^T(GWG^T)^{-1} \succeq \big(GJ^{-1}G^T\big)^{-1}.$$

Of course, $\mathbf{var}_P\big[g_{\mathbf{x}}(\theta^*)\big]$ is typically unknown in practice. To obtain an efficient GMM estimator, i.e. a GMM estimator with minimum asymptotic variance, practitioners usually resort to a two-step procedure:

1. obtain a consistent pilot estimator $\tilde{\theta}_n$ e.g. by an inefficient GMM estimator.
2. obtain $\hat{\theta}_n$ by $\arg\min_{\theta\in\Theta} \big\|\frac{1}{n}\sum_{i\in[n]} g_{\mathbf{x}_i}(\theta)\big\|^2_{\widehat{W}_n}$, where

$$\widehat{W}_n \leftarrow \tfrac{1}{n}\textstyle\sum_{i\in[n]} g_{\mathbf{x}_i}(\tilde{\theta}_n)g_{\mathbf{x}_i}(\tilde{\theta}_n)^T.$$

Since $\widehat{W}_n \xrightarrow{p} \mathbf{var}_P\big[g_{\mathbf{x}_1}(\theta^*)\big]$ by the LLN, it is possible to adapt the proof of Theorem 1.1 to show that $\{\hat{\theta}_n\}$ is asymptotically normal, and its asymptotic variance is $\big(GJ^{-1}G^T\big)^{-1}$.

2. Asymptotic efficiency. In the notes on point estimation, we considered a decision theoretic approach to evaluating point estimators. Here, we consider an asymptotic approach.

The appeal of the asymptotic approach is the possibility of *exactly* characterizing the risk (typically MSE) of a point estimator. For any sequence of consistent estimators, the asymptotic bias vanishes. Thus the asymptotic MSE is the asymptotic variance.

THEOREM 2.1 (Cramér-Rao inequality). *Assume the density of* $\mathbf{x}$ *obeys*

$$\mathbf{E}_\theta\big[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big]\big] = 0.$$

Let $\tau(\theta) := \mathbf{E}_\theta\big[\delta(\mathbf{x})\big]$. *For any estimator* $\delta(\mathbf{x})$ *that obeys*

$$\nabla_j \tau(\theta) = \nabla_j \, \mathbf{E}_\theta\big[\delta(\mathbf{x})\big] = \int_{\mathcal{X}} \delta(\mathbf{x}) \, \nabla_j f_\theta(x) dx, \tag{2.1}$$

for any $j \in [p]$, *we have*

$$\mathbf{var}_\theta\big[\delta(\mathbf{x})\big] \succeq \nabla\tau(\theta) \, \mathbf{var}_\theta\big[\nabla_\theta \log f_\theta(\mathbf{x})\big]^{-1} \nabla\tau(\theta),$$

where $A \succeq B$ *means* $A - B$ *is positive semi-definite.*

PROOF. Consider

$$\mathbf{var}_\theta\left[\begin{bmatrix}\delta(\mathbf{x}) \\ \nabla_\theta \log f_\theta(\mathbf{x})\end{bmatrix}\right] = \begin{bmatrix}\mathbf{var}_\theta\big[\delta(\mathbf{x})\big] & \mathbf{cov}_\theta\left[\delta(\mathbf{x}) \, \nabla_\theta\big[\log f_\theta(\mathbf{x})\big]^T\right] \\ \mathbf{cov}_\theta\left[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big] \, \delta(\mathbf{x})^T\right] & \mathbf{var}_\theta\big[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big]\big]\end{bmatrix}.$$

It is, by definition, positive semi-definite, which implies the Schur complement of the lower-right block

$$\mathbf{var}_\theta\big[\delta(\mathbf{x})\big] - \mathbf{cov}_\theta\left[\delta(\mathbf{x}) \, \nabla_\theta\big[\log f_\theta(\mathbf{x})\big]^T\right] \mathbf{var}_\theta\big[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big]\big]^{-1} \mathbf{cov}_\theta\left[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big] \, \delta(\mathbf{x})^T\right]$$

is positive semi-definite. Rearranging, we obtain

$$\mathbf{var}_\theta\big[\delta(\mathbf{x})\big] \succeq \mathbf{cov}_\theta\left[\delta(\mathbf{x}) \, \nabla_\theta\big[\log f_\theta(\mathbf{x})\big]^T\right] \mathbf{var}_\theta\big[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big]\big]^{-1} \mathbf{cov}_\theta\left[\nabla_\theta\big[\log f_\theta(\mathbf{x})\big] \, \delta(\mathbf{x})^T\right].$$

To complete the proof, we must show that

$$\nabla_\theta \, \mathbf{E}_\theta \big[\delta(\mathbf{x})\big] = \mathbf{cov}_\theta \Big[\, \delta(\mathbf{x}) \, \nabla_\theta \big[\log f_\theta(\mathbf{x})\big]^T \Big] \, .$$

Indeed, by differentiating under the integral sign, we have

$$\begin{aligned} \nabla_j \, \mathbf{E}_\theta \big[\delta(\mathbf{x})\big] &= \int_{\mathcal{X}} \delta(\mathbf{x}) \, \nabla_j f_\theta(x) dx \\ &= \int_{\mathcal{X}} \delta(\mathbf{x}) \, \frac{\nabla_j f_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\ &= \int_{\mathcal{X}} \delta(\mathbf{x}) \, \nabla_j \big[\log f_\theta(x)\big] f_\theta(x) dx \\ &= \mathbf{E}_\theta \Big[\, \delta(\mathbf{x}) \, \nabla_j \big[\log f_\theta(\mathbf{x})\big] \Big], \end{aligned}$$

which, by the assumption $\mathbf{E}_\theta \big[\nabla_\theta \big[\log f_\theta(\mathbf{x})\big]\big] = 0$, is

$$= \mathbf{cov}_\theta \Big[\, \delta(\mathbf{x}) \, \nabla_j \big[\log f_\theta(\mathbf{x})\big] \Big].$$

for any $j \in [p]$. □

As the Fisher Information increases, the Cramér-Rao inequality decreases. When $\mathbf{x}$ consists of *i.i.d.* random variables $\mathbf{x}_i$, $i \in [n]$, the Fisher Information increases linearly as the sample size grows:

$$\begin{aligned} I(\theta) &= \mathbf{var}_\theta \big[\nabla_\theta \big[\log f_\theta(\mathbf{x})\big]\big] \\ &= \mathbf{var}_\theta \Big[\textstyle\sum_{i\in[n]} \nabla_\theta \big[\log f_\theta(\mathbf{x}_1)\big]\Big] \\ &= n \, \mathbf{var}_\theta \big[\nabla_\theta \big[\log f_\theta(\mathbf{x}_1)\big]\big], \end{aligned}$$

where $f_\theta(x_1)$ is the density of $\mathbf{x}_1$.

We hasten to mention that the Cramér-Rao lower bound (as stated) is valid for any estimator that obeys (2.1). If $\delta(\mathbf{x})$ is unbiased, then

$$\nabla_\theta \, \mathbf{E}_\theta \big[\delta(\mathbf{x})\big] = \nabla_\theta \, \theta = I_p,$$

and the Cramér-Rao lower bound simplifies to

$$\mathbf{var}_\theta \big[\delta(\mathbf{x})\big] \succeq \mathbf{var}_\theta \big[\nabla_\theta \big[\log f_\theta(\mathbf{x})\big]\big]^{-1}. \tag{2.2}$$

Some sources refer to (2.2) as the Cramér-Rao lower bound.

Finally, we remark that the Cramér-Rao inequality is non-asymptotic: it is valid for finite sample sizes. It is generally not sharp; i.e. there may not

be any estimator that saturates the inequality. Even in the usually favorable setting of a one-dimensional exponential family, it is only known that $\phi(\mathbf{x})$ is an unbiased estimator of $\mathbf{E}_\theta\big[\phi(\mathbf{x})\big]$ that attains the lower bound. It is not known whether there are estimators of other functions of θ that attain the lower bound. However, as we shall see, the MLE saturates the inequality in asymptopia.

DEFINITION 2.2. *A sequence of estimators $\{\hat\theta_n\}$ is asymptotically efficient at a parameter $\tau(\theta) \in \mathbf{R}^p$ if*

1. *$\{\hat\theta_n\}$ is consistent: $\hat\theta_n \xrightarrow{p} \theta$*
2. *$\{\hat\theta_n\}$ is asymptotically normal: $\sqrt{n}(\hat\theta_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$,*
3. *the asymptotic variance of $\{\hat\theta_n\}$ attains the Cramér-Rao lower bound:*

$$V(\theta) = I(\theta^*)^{-1}.$$

By comparing the asymptotic distribution of the MLE with the Cramér-Rao lower bound, we see that the MLE is asymptotically efficient. The ratio between the asymptotic variance of the MLE and that of another estimator is the *asymptotic relative efficiency* (ARE) of the other estimator.

EXAMPLE 2.3. *Let $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Poi}(\lambda)$. We wish to estimate the probability of observing no events $\mathbf{P}_\lambda(\mathbf{x} = 0)$, which, by the Poisson density, is $e^{-\lambda}$. A naive estimator is*

$$\hat\theta_{\text{Ber},n} = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}\{\mathbf{x}_i = 0\}.$$

By the CLT, its asymptotic distribution is

$$\sqrt{n}\big(\hat\theta_{\text{Ber},n} - e^{-\lambda}\big) \xrightarrow{d} \mathcal{N}\big(0, e^{-\lambda}\big(1 - e^{-\lambda}\big)\big).$$

The MLE of $e^{-\lambda}$ is $e^{-\bar{\mathbf{x}}}$, where $\bar{\mathbf{x}}$ is the MLE of λ. By the CLT, the asymptotic variance of $\bar{\mathbf{x}}_n$ is λ. By the delta method,

$$\sqrt{n}\big(e^{-\bar{\mathbf{x}_n}} - e^{-\lambda}\big) \xrightarrow{d} \mathcal{N}\big(0, \lambda e^{-2\lambda}\big)$$

We see that the asymptotic relative efficiency of $\hat\theta_{\text{Ber}}$ to the MLE $e^{-\bar{\mathbf{x}}_n}$ is

$$\frac{\lambda e^{-2\lambda}}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda}{e^\lambda - 1},$$

which decreases exponentially as λ grows.

YUEKAI SUN
BERKELEY, CALIFORNIA
DECEMBER 4, 2015