STAT 460/560 Class 20: Asymptotic Normality of of M- and Z-estimators

Ben Bloem-Reddy

Reading: Chapter 5.3, [van98].

Last time, we established consistency for M- and Z-estimators under a couple of different conditions. Today, we'll look at asymptotic normality. Following van der Vaart, we'll start with Z-estimators. Recall that a Z-estimator $\hat{\theta}_n$ solves

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = \hat{P}_n \psi_\theta = 0 \; .$$

We'll assume that $P\psi_{\theta_0} = 0$, so that θ_0 is (asymptotically) the value of θ to which $\hat{\theta}_n$ converges in probability, i.e., $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$.

Classically, one assumes that $\theta \mapsto \Psi_n(\theta)$ has two derivatives, at which point the proof of asymptotic normality proceeds pretty much the same as how we proved asymptotic normality for MLEs a few weeks ago. First, perform a second-order Taylor expansion of Ψ_n around θ_0 (and for simplicity assuming that $\theta \in \mathbb{R}$),

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \ddot{\Psi}_n(\tilde{\theta}_n) \; ,$$

where is a point between $\hat{\theta}_n$ and θ_0, and the dots above a function indicate derivatives with respect to θ. Rewriting, we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ddot{\Psi}_n(\tilde{\theta}_n)} \; .$$

Under appropriate conditions (think about what they are), this should converge in distribution to a normal random variable with mean zero and variance

$$\frac{P\psi_{\theta_0}^2}{(P\dot{\psi}_{\theta_0})^2} \; .$$

Actually, we can conclude something more general, in which we won't assume the existence of a second derivative, instead replacing it with a Lipschitz continuity condition: assume there is a measurable function $\bar{\psi}$ with $P(\bar{\psi}^2) < \infty$ such that for every θ_1, θ_2 in a neighborhood of θ_0, and each x,

$$\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| \leq \bar{\psi}(x)\|\theta_1 - \theta_2\| \; . \tag{20.1}$$

Theorem 20.1. *For each θ in an open subset of $\mathbb{R}^k$, let $x \mapsto \psi_\theta(x)$ be a measurable vector-valued function satisfying* (20.1)*. Assume the following of the map $\theta \mapsto P\psi_\theta$: it has a zero at θ_0, that $P\|\psi_{\theta_0}\|^2 < \infty$, and it is differentiable at θ_0, with invertible derivative matrix V_{θ_0}. If $\hat{P}_n \psi_{\hat{\theta}_n} = o_P(n^{-1/2})$ and $\hat{\theta}_n \xrightarrow{p} \theta_0$, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_P(1) \; , \tag{20.2}$$

which implies that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}_k(0, V_{\theta_0}^{-1} P(\psi_{\theta_0}\psi_{\theta_0}^\top)(V_{\theta_0}^{-1})^\top) \; . \tag{20.3}$$

Proof. First, let's establish something easy:

$$\sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0}) \rightsquigarrow \mathcal{N}_k(0, P(\psi_{\theta_0}\psi_{\theta_0}^\top)) \; .$$

This follows from the fact that $\hat{P}_n\psi_{\theta_0} \xrightarrow{\text{P}} P\psi_{\theta_0} = 0$ (by the LLN) and the CLT. We'll use this later.

Next, van der Vaart tells us that the consistency of $\hat{\theta}_n$ and the Lipschitz condition (20.1) imply that

$$\sqrt{n}(\hat{P}_n\psi_{\hat{\theta}_n} - P\psi_{\hat{\theta}_n}) - \sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0}) \xrightarrow{\text{P}} 0 \; . \tag{20.4}$$

We have to take his word for this because establishing it requires (again) tools from Chapter 19.[1] But we can sort of see how it might work in the case that we have a nonrandom sequence $\theta_n \to \theta_0$. First, note that for fixed nonrandom θ_n,

$$E[\sqrt{n}(\hat{P}_n\psi_{\theta_n} - P\psi_{\theta_n})] = E[\sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0})] = 0 \; , \quad \text{and} \quad P\|\psi_{\theta_n} - \psi_{\theta_0}\|^2 \leq P\bar{\psi}^2\|\theta_n - \theta_0\|^2 \to 0 \; .$$

The second term bounds the variances, which therefore converge to zero.

Going back to (20.4), consider the term $\sqrt{n}(\hat{P}_n\psi_{\hat{\theta}_n} - P\psi_{\hat{\theta}_n})$. By assumptions, $\hat{P}_n\psi_{\hat{\theta}_n} = o_P(n^{-1/2})$, which is the same as saying that $\sqrt{n}\hat{P}_n\psi_{\hat{\theta}_n} = o_P(1)$. Moreover, since $P\psi_{\theta_0} = 0$, we have $\sqrt{n}\hat{P}_n\psi_{\hat{\theta}_n} = 0 + o_P(1) = P\psi_{\theta_0} + o_P(1)$. Therefore, we can write

$$\sqrt{n}(\hat{P}_n\psi_{\hat{\theta}_n} - P\psi_{\hat{\theta}_n}) = \sqrt{n}(P\psi_{\theta_0} - P\psi_{\hat{\theta}_n}) + o_P(1) \; .$$

Since $P\psi_\theta$ is differentiable at θ_0 (recall the definition of differentiability from Class 5), this becomes

$$\sqrt{n}(\hat{P}_n\psi_{\hat{\theta}_n} - P\psi_{\hat{\theta}_n}) = \sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + o_P(1 + \sqrt{n}\|\theta_0 - \hat{\theta}_n\|) \; .$$

On the other hand, (20.4) implies that $\sqrt{n}(\hat{P}_n\psi_{\hat{\theta}_n} - P\psi_{\hat{\theta}_n}) = \sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0}) + o_P(1)$, so the previous equation becomes

$$\sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0}) + o_p(1) = \sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + o_P(1 + \sqrt{n}\|\theta_0 - \hat{\theta}_n\|) \; . \tag{20.5}$$

We're almost there, but we need to show that the error term $o_P(\sqrt{n}\|\theta_0 - \hat{\theta}_n\|)$ doesn't blow up. We can do so by showing that $\sqrt{n}\|\theta_0 - \hat{\theta}_n\| = O_P(1)$ (i.e., it is bounded in probability). (See the activity below.) This has a name: $\hat{\theta}_n$ is $\sqrt{n}$*-consistent*. One of van der Vaart's rules of calculus for o_P and O_P is that $o_P(O_P(1)) = o_P(1)$, from which our result will follow.

Going back to (20.5), multiplying by $V_{\theta_0}^{-1}$ yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0}) + o_P(1) \; ,$$

which is (20.2). □

Activity 20.1. Finish the proof by showing that $\sqrt{n}\|\theta_0 - \hat{\theta}_n\| = O_P(1)$.

Hint: Recall that if a sequence X_n converges in distribution then it is bounded in probability.

Solution: We established above that $\sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0})$ converges in distribution, which implies that it is bounded in probability. Taking the norm of both sides of the previous equation yields

$$\begin{aligned}
\sqrt{n}\|\theta_0 - \hat{\theta}_n\| &= \sqrt{n}\|V_{\theta_0}^{-1}V_{\theta_0}(\theta_0 - \hat{\theta}_n)\| \\
&\leq \|V_{\theta_0}^{-1}\|\sqrt{n}\|V_{\theta_0}(\theta_0 - \hat{\theta}_n)\| \\
&= \|V_{\theta_0}^{-1}\|\big(\|\sqrt{n}(\hat{P}_n\psi_{\theta_0} - P\psi_{\theta_0})\| + o_P(1) + o_P(\sqrt{n}\|\theta_0 - \hat{\theta}_n\|)\big) \\
&= O_P(1) + o_P(\sqrt{n}\|\theta_0 - \hat{\theta}_n\|) \; .
\end{aligned}$$

[1] For the interested, the main challenge is establishing that the functions ψ_θ form something known as a *Donsker class*.

Hence, $\sqrt{n}\|\theta_0 - \hat{\theta}_n\| = O_P(1)$.

1. Asymptotic normality of M-estimators

Recall that $\hat{\theta}_n$ is an M-estimator if it maximizes

$$\hat{P}_n m_\theta \ ,$$

which in the limit Pm_θ is assume to be maximized at θ_0. For the next theorem, we assume that $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion

$$Pm_\theta = Pm_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^\top V_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2) \ , \tag{20.6}$$

where V_θ is the second derivative matrix.

Theorem 20.2. *For each θ in an open subset of $\mathbb{R}^k$, let $x \mapsto m_\theta$ be a measurable function. Let $\theta \mapsto m_\theta(x)$ be differentiable at θ_0 for P-almost every x, with derivative $\dot{m}_\theta(x)$, and such that $\theta \mapsto m_\theta$ satisfies the Lipschitz condition* (20.1) *for some bounding function $\bar{m}(x)$. Moreover, assume that $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion* (20.6) *at a point of maximum θ_0, with invertible symmetric second derivative matrix V_{θ_0}. If $\hat{P}_n m_{\hat{\theta}} \geq \sup_\theta \hat{P}_n m_\theta - o_P(n^{-1})$ and $\hat{\theta} \xrightarrow{p} \theta_0$, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i) + o_P(1) \ , \tag{20.7}$$

which implies that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}_k(0, V_{\theta_0}^{-1} P(\dot{m}_{\theta_0} \dot{m}_{\theta_0}^\top)(V_{\theta_0}^{-1})) \ . \tag{20.8}$$

Proof. The proof relies on two technical lemmas proved elsewhere (one in Chapter 19 (again!) and one near the end of Chapter 5). The first is that for every random sequence h_n that is bounded in probability,

$$\sqrt{n}(\hat{P}_n \left[\sqrt{n}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - h_n^\top \dot{m}_{\theta_0}\right] - P\left[\sqrt{n}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - h_n^\top \dot{m}_{\theta_0}\right]) \xrightarrow{\text{p}} 0 \ .$$

Secondly, $\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_P(1)$. With these in hand, we can complete the proof.

First, using the second-order Taylor expansion of Pm_θ, we can rearrange the previous equation as

$$n\hat{P}_n(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2} h_n^\top V_{\theta_0} h_n + \sqrt{n}(\hat{P}_n h_n^\top \dot{m}_{\theta_0} - P h_n^\top \dot{m}_{\theta_0}) + o_P(1) \ .$$

Because $\hat{h}_n := \sqrt{n}\hat{\theta}_n - \theta_0$ is bounded in probability and $\tilde{h}_n := -V_{\theta_0}^{-1}\sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0})$ converges in distribution (and therefore is also bounded in probability), this holds for each of them. Note that $\theta_0 + \hat{h}_n/\sqrt{n} = \hat{\theta}_n$. Plugging these in, we get

$$n\hat{P}_n(m_{\hat{\theta}_n} - m_{\theta_0}) = \frac{1}{2}\hat{h}_n^\top V_{\theta_0} \hat{h}_n + \sqrt{n}(\hat{P}_n \hat{h}_n^\top \dot{m}_{\theta_0} - P\hat{h}_n^\top \dot{m}_{\theta_0}) + o_P(1)$$

$$n\hat{P}_n(m_{\theta_0 + \tilde{h}_n/\sqrt{n}} - m_{\theta_0}) = -\frac{1}{2}\sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0})^\top V_{\theta_0}^{-1} \sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0}) + o_P(1)$$

By assumption, $\hat{\theta}_n$ approximately maximizes $\theta \mapsto \hat{P}_n m_\theta$, so the LHS of the first equation is greater than the LHS of the second, up to error of $o_P(1)$, and therefore the same holds for the RHS. Taking the difference and completing the square, we get

$$\frac{1}{2}(\hat{h}_n + V_{\theta_0}^{-1}\sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0}))^\top V_{\theta_0}(\hat{h}_n + V_{\theta_0}^{-1}\sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0})) + o_P(1) \geq 0 \ .$$

Since θ_0 maximizes Pm_θ, and the matrix of second derivatives V_{θ_0} is invertible, it must be strictly negative definite. Therefore, the quadratic form must converge to zero in probability, and the same must be true for $\|\sqrt{n}(\hat{\theta}_n - \theta_0) + V_{\theta_0}^{-1}\sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0})\|$.

Summing up, $\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\sqrt{n}(\hat{P}_n \dot{m}_{\theta_0} - P\dot{m}_{\theta_0}) + o_P(1)$. Since $P\dot{m}_{\theta_0} = 0$, the CLT and delta method yield the asymptotic normality. □

Exercise 20.1. Apply the previous theorem to the sample median of $X_1, \ldots, X_n$ with CDF F and PDF f to show that it is asymptotically normal with variance $1/(2f(\theta_0))^2$.

Hint: The sample median is also the M-estimator for $m_\theta(x) = |x - \theta| - |x|$.

References

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.