Asymptotic normality for M-estimators based on non-smooth criterion functions

John Duchi

March 1, 2018

In this note, we will describe a few consequences of the convergence theory for metric-space-valued random variables that we have developed. We will describe a few convergence results, and then we will use them to prove asymptotic normality of M-estimators based on possibly non-smooth criterion functions.

As motivation for this exercise, we may consider the loss function

$$\ell_\theta(x) = |\theta - x|$$

or, for some $\alpha \in (0, 1)$,

$$\ell_\theta(x) = (1 - \alpha) \left[\theta - x\right]_+ + \alpha \left[x - \theta\right]_+ .$$

In both cases, the loss $\ell_\theta(x)$ is not differentiable in θ everywhere, though it is differentiable for almost every θ (or, conversely, for all $x \neq \theta$). In the former case, the risk $R(\theta) := \mathbb{E}[\ell_\theta(X)]$ is minimized by the median of X, while the latter yields the α-quantile of X as its minimizers. In each case, we might expect that if the random variable has a density near $\theta_0 = \operatorname{argmin}_\theta R(\theta)$, then we we should have some type of differentiability. More generally, it is of course of interest to understand the asymptotics of median and quantile estimators, as they are important quantities.

Let us begin with the less-basic theorem that guarantees asymptotic normality of M-estimators. For this theorem, we require a few conditions on the loss ℓ_θ and the risk $R(\theta) := \mathbb{E}[\ell_\theta(X)]$. First, we require that $\theta_0 = \operatorname{argmin}_\theta R(\theta)$ is unique, and that the risk $R(\theta)$ be twice differentiable at θ_0 with positive-definite Hessian $\nabla^2 R(\theta_0) \succ 0$. We then require a few conditions on our losses. First, we require that near θ_0, they are locally Lipschitz: there exists a function $\dot{\ell} : \mathcal{X} \to \mathbb{R}_+$ with $P\dot{\ell}^2 < \infty$ such that for all θ_1, θ_2 in a (small enough) ball around θ_0,

$$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \leq \dot{\ell}(x) \left\|\theta_1 - \theta_2\right\| \quad \text{for all } x \in \mathcal{X}.$$

We also require that for P-almost all $x \in \mathcal{X}$, the function $\theta \mapsto \ell_\theta(x)$ is differentiable at $\theta = \theta_0$ with derivative $\dot{\ell}_{\theta_0}(x) = \frac{\partial}{\partial \theta} \ell_\theta(x)|_{\theta = \theta_0}$.

Theorem 1 (Theorem 5.23 [1])**.** *Let the above conditions hold, and in addition, assume that*

$$P_n \ell_{\widehat{\theta}_n} \leq \inf_\theta P_n \ell_\theta + o_P(1/n)$$

and that $\widehat{\theta}_n$ is weakly consistent for θ_0, meaning $\widehat{\theta}_n \xrightarrow{p} \theta_0$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\nabla^2 R(\theta_0)^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_P(1).$$

Let us give one example of this theorem as applied to the quantile estimator.

Example 1 (Quantile estimation)**:** Consider the loss function

$$\ell_\theta(x) := (1-\alpha)\left[\theta - x\right]_+ + \alpha\left[x - \theta\right]_+,$$

where $\alpha \in (0,1)$. This loss is a convex function, and moreover, we see that for all θ_1, θ_2, we have

$$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \le |\theta_1 - \theta_2|,$$

so under the conditions of the theorem, $\dot{\ell}(x) \equiv 1$ is sufficient. (The loss is certainly 1-Lipschitz continuous.) A calculation, which you should do for yourself, shows that $R(\theta) = \mathbb{E}[\ell_\theta(X)]$ is minimized by the α-quantile of X, that is,

$$\theta_0 = Q_P(\alpha) := \inf\left\{\theta \in \mathbb{R} \mid \alpha \le P(X \le \theta)\right\}.$$

Now, let $F(t) = P(X \le t)$ denote the CDF of X, and assume that X has a density $f(x)$ in a neighborhood of θ_0, the α quantile. In this case, the α-quantile is unique, and the derivative

$$\dot{\ell}_{\theta_0}(x) = (1-\alpha)1\left\{\theta_0 \ge x\right\} - \alpha 1\left\{\theta_0 \le x\right\}$$

exists with P-probability 1 (that is, for P-almost every x). Then, as the density f of X exists near θ_0, we have that

$$R'(\theta) = P\dot{\ell}_\theta = (1-\alpha)P(\theta \ge X) - \alpha P(\theta \le X) = P(X \le \theta) - \alpha = F(\theta) - \alpha$$

for all θ near θ_0. In this case, we may take a second derivative to obtain

$$R''(\theta_0) = F'(\theta_0) = f(\theta_0) > 0$$

and so we have the conditions of Theorem 1.

The consistency of $\widehat{\theta}_n = \operatorname{argmin}_\theta P_n \ell_\theta$ may follow from many avenues, but Exercise 7.10 in our collection of exercises shows one reasonably easy-to-apply possibility when the loss is convex in θ. Thus, applying Theorem 1 yields

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) &= -\frac{1}{f(\theta_0)} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[(1-\alpha)1\left\{\theta_0 \ge X_i\right\} - \alpha 1\left\{\theta_0 \le X_i\right\}\right] \\ &\stackrel{d}{\to} \mathsf{N}\left(0, \frac{\alpha(1-\alpha)}{f(\theta_0)^2}\right).\end{aligned}$$

When $\alpha = \frac{1}{2}$, which corresponds to the median, we obtain the asymptotic variance $\frac{1}{4f(\theta_0)^2}$. ♣

There are many other ways to discover the asymptotics of quantile-like estimators, including infinite-dimensional analogues of the delta method, as in Chapters 20 & 21 of van der Vaart [1].

It is often interesting to compare median estimators of location, especially for symmetric distributions, to other location estimators, such as mean estimators. Let us consider first the Gaussian distribution, where $f(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$. In this case, the median of a sample $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1)$ satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \stackrel{d}{\to} \mathsf{N}\left(0, \frac{\pi}{2}\right),$$

which is less efficient than the mean, which of course satisfies $\sqrt{n}(\overline{X}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, 1)$. For the Laplace (or double-exponential) distribution, the story is somewhat different. In this case, we have density $f(x) = \frac{1}{2}e^{-|x|}$, and the median estimator $\widehat{\theta}_n$ satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, 1),$$

while the variance of a Laplace random variable is 2, so that the sample mean satisfies $\sqrt{n}(\overline{X}_n - \theta_0) \xrightarrow{d} \mathsf{N}(0, 2)$, which is half as efficient as the median.

References

[1] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.