STAT6203: Theoretical Statistics III

Robust Statistics

Instructed by *Prof. Marco Avella Medina*
Transcribed by *Satvik Saha*

Department of Statistics, Columbia University

Table of Contents

1. Location Estimation

Consider the problem of estimating an unknown location parameter ξ from a sample

$$X_1, ..., X_n \overset{\text{iid}}{\sim} F(\cdot - \xi).$$

In general, we search for suitably *efficient* estimators. Huber (1964) seeks to obtain *robust* (in a certain sense) estimators which perform well when F is only being approximately known; in particular, suppose that F is a slightly 'contaminated' version of a known distribution F_0.

Definition 1.1 (Tukey-Huber Contamination). Given $\varepsilon \in [0, 1]$ and a distribution F_0, define

$$\mathcal{P}_\varepsilon(F_0) := \{(1 - \varepsilon)F_0 + \varepsilon H : \text{arbitrary distribution } H\}.$$

Remark. $\mathcal{P}_\varepsilon(F_0)$ is contained within the ε-Kolmogorov-Smirnov neighborhood of F_0: for any $F = (1 - \varepsilon)F_0 + \varepsilon H$,

$$\|F - F_0\|_\infty = \|(1 - \varepsilon)F_0 + \varepsilon H - F_0\|_\infty = \varepsilon\|H - F_0\|_\infty \leq \varepsilon.$$

With this, we may examine how an estimator based for a parameter ofF_0 suffers under the contaminated sample from $F \in \mathcal{P}_\varepsilon(F_0)$. As a first step, we may look at some simple population-level functionals.

Example 1.2. Consider the mean functional $\mu(G) := \mathbb{E}_G[X]$. The 'bias' of μ on $F := (1 - \varepsilon)F_0 + \varepsilon H$ against F_0 is

$$\mu(F) - \mu(F_0) = \varepsilon\mu(H) - \varepsilon\mu(F_0).$$

Note that this is completely unrestricted over $F \in \mathcal{P}_\varepsilon(F_0)$ even if we fix ε, owing to the arbitrary nature of H (which in principle may not even have a mean!).

Example 1.3. Consider the (left) median functional $m(G) := G^-\left(\frac{1}{2}\right) = \inf\left\{x : G(x) \geq \frac{1}{2}\right\}$, and let $\varepsilon < \frac{1}{2}$. For $F := (1 - \varepsilon)F_0 + \varepsilon H$, note that

$$m(F) = \inf\left\{x : (1 - \varepsilon)F_0(x) + \varepsilon H(x) \geq \frac{1}{2}\right\}.$$

Since $0 \leq H(x) \leq 1$, we can bound

$$\begin{aligned} F_0^-\left(\frac{\frac{1}{2} - \varepsilon}{1 - \varepsilon}\right) &= \inf\left\{x : (1 - \varepsilon)F_0(x) + \varepsilon \geq \frac{1}{2}\right\} \\ &\leq m(F) \\ &\leq \inf\left\{x : (1 - \varepsilon)F_0(x) \geq \frac{1}{2}\right\} = F_0^-\left(\frac{\frac{1}{2}}{1 - \varepsilon}\right). \end{aligned}$$

Unlike the previous example, the median $m(F)$ cannot deviate arbitrarily over $F \in \mathcal{P}_\varepsilon(F_0)$.

1.1. Minimax Asymptotic Bias

For the purposes of location estimation, it is natural to restrict ourselves to *translation invariant* estimators.

Definition 1.4 (Translation Invariant Estimators). We say that an estimator T is translation invariant if $T(X + a\mathbf{1}) = T(X) + a$ for all samples X and $a \in \mathbb{R}$. We denote the class of all translation invariant estimators by $\mathcal{T}$.

Definition 1.5 (Asymptotic Bias). Let $\{T_n\}$ be a sequence of estimators for $T(F_0)$. Its asymptotic bias is defined as

$$b(\{T_n\}, F) := \limsup_{n\to\infty} |\mathbb{E}_F[T_n] - T(F_0)|.$$

We can now choose our 'best' estimator on the basis of asymptotic bias; fix $\varepsilon > 0$ and consider the minimax problem

$$b_* := \min_{\{T_n\}\subset\mathcal{T}} \max_{F\in\mathcal{P}_\varepsilon(F_0)} b(\{T_n\}, F).$$

Theorem 1.6. *Let F_0 have a density that is symmetric about zero and decreasing on $\mathbb{R}_+$, and let $T(F_0) = 0$. The sample medians $\{M_n\}$ solve the minimax asymptotic bias problem, with*

$$\min_{\{T_n\}\subset\mathcal{T}} \max_{F\in\mathcal{P}_\varepsilon(F_0)} b(\{T_n\}, F) = F_0^{-1}\left(\frac{1/2}{1-\varepsilon}\right).$$

Proof. First note that the sample median estimator is consistent, and $\lim_{n\to\infty} \mathbb{E}_F[M_n] = F^-\left(\frac{1}{2}\right)$ under some mild conditions on F. We have already demonstrated in Example 1.3 that $|F^-\left(\frac{1}{2}\right)| \le F_0^{-1}\left(\frac{1/2}{1-\varepsilon}\right) =: m_*$ for all $F \in \mathcal{P}_\varepsilon(F_0)$, which gives us the upper bound $b_* \le m_*$.

Next, we will construct two distributions $F_+, F_- \in \mathcal{P}_\varepsilon(F_0)$, and use

$$b_* \ge \min_{\{T_n\}\subset\mathcal{T}} \max\big(b(\{T_n\}, F_+), b(\{T_n\}, F_-)\big) \tag{$\star$}$$

to obtain the matching lower bound $b_* \ge m_*$. Define

$$F_+(x) := \begin{cases} (1-\varepsilon)F_0(x) & \text{if } x < m_*, \\ 1-(1-\varepsilon)F_0(2m_* - x) & \text{if } x \ge m_*. \end{cases}$$

This describes a distribution which has the same shape as F_0 on $(-\infty, m_*)$, and is symmetric about m_*. Further set $F_-(x) = F_+(x + 2m_*)$, which is symmetric about $-m_*$. We can verify that F_+, F_- are indeed members of $\mathcal{P}_\varepsilon(F_0)$; solving for $F_+ = (1-\varepsilon)F_0 + \varepsilon H_+$ reveals that

$$H_+(x) = \frac{F_+(x) - (1-\varepsilon)F_0(x)}{\varepsilon} = \frac{1-(1-\varepsilon)(F_0(x) + F_0(2m_* - x))}{\varepsilon}\mathbf{1}(x \ge m_*).$$

This is continuous with $H(-\infty) = 0$, $H(\infty) = 1$, and is nondecreasing since

$$F_0(x) + F_0(2m_* - x) = F_0(x) - F_0(x - 2m_*) + 1$$

is decreasing on $x \geq m_*$, which makes it a valid *cdf*. The argument for F_-, which is a reflection of F_+ about zero, is similar.

Now, if $b_* < m_*$, then m_* must be strictly greater than the right hand side of ($\star$), hence there exists a sequence $\{T_n\} \subset \mathcal{T}$ such that

$$\max\left(\limsup_{n\to\infty} |\mathbb{E}_{F_+}[T_n]|, \quad \limsup_{n\to\infty} |\mathbb{E}_{F_-}[T_n]|\right) < m_*.$$

However, translation invariance of T_n forces $\mathbb{E}_{F_+}[T_n] = \mathbb{E}_{F_-}[T_n] + 2m_*$, hence

$$2m_* = \limsup_{n\to\infty} |\mathbb{E}_{F_+}[T_n] - \mathbb{E}_{F_-}[T_n]| \leq \limsup_{n\to\infty} |\mathbb{E}_{F_+}[T_n]| + \limsup_{n\to\infty} |\mathbb{E}_{F_-}[T_n]| < 2m_*,$$

a contradiction! □

Remark. Some further technical conditions are required to ensure that $\mathbb{E}_F[M_n] \to F^-\left(\frac{1}{2}\right)$, which we omit here.

1.2. Minimax Asymptotic Variance

Instead of relying on asymptotic bias, we will now examine the asymptotic variance of a special class of estimators.

Definition 1.7 (M-estimator). We say that T_n is an M-estimator of location if

$$T_n \in \arg\min_t \sum_{i=1}^n \rho(X_i - t)$$

for some function ρ. When ρ is differentiable, we denote $\psi := \rho'$, and T_n equivalently satisfies

$$\sum_{i=1}^n \psi(X_i - T_n) = 0.$$

Note that M-estimators are automatically translation invariant. We will generally assume that the loss function ρ is symmetric and convex from now on.

Lemma 1.8 (Asymptotic Normality of M-estimators). *Let $\{T_n\}$ be the sequence of M estimators associated with ψ. Define the maps $\lambda(t) := \mathbb{E}_F[\psi(X - t)]$, $\sigma^2(t) := \mathrm{var}_F[\psi(X - t)]$, and suppose that there exists $t_0 \in \mathbb{R}$ such that the following are satisfied.*

1. *λ is continuous in a neighborhood of t_0, with $\lambda(t_0) = 0$.*
2. *λ is differentiable at t_0, with $\lambda'(t_0) < 0$.*
3. *σ^2 is continuous, finite, non-zero in a neighborhood of t_0.*

Then $\sqrt{n}(T_n - t_0) \xrightarrow{d} \mathcal{N}(0, V(\psi, F))$, where

$$V(\psi, F) := \frac{\sigma^2(t_0)}{(\lambda'(t_0))^2}$$

is the asymptotic variance of $\{T_n\}$.

Example 1.9. The M-estimator associated with $\rho(x) = x^2$ is the sample mean, with asymptotic variance $\operatorname{var}_F(X)$.

Example 1.10. The M-estimator associated with $\rho(x) = |x|$ is the sample median, with asymptotic variance $1/4\left(f\left(F^{-}\left(\frac{1}{2}\right)\right)\right)^2$.

Example 1.11. The M-estimator associated with $\rho(x) = -\log f(x)$ is the MLE of ξ in the location family $F(\cdot - \xi)$.

From now on, we will restrict ourselves to the contamination neighborhood

$$\mathcal{P}_\varepsilon^{\text{sym}}(F_0) := \{(1-\varepsilon)F_0 + \varepsilon H : \text{symmetric } H\}$$

to ensure that $F \in \mathcal{P}_\varepsilon^{\text{sym}}(F_0)$ is symmetric (about zero), and further select $F_0 = \Phi$.

Consider the problem

$$V_* := \min_\psi \max_{F \in \mathcal{P}_\varepsilon^{\text{sym}}(\Phi)} V(\psi, F),$$

where ψ ranges over functions satisfying the requirements of Lemma 1.8, along with symmetry and convexity so that $t_0 = 0$.

Theorem 1.12. *Define the Huber loss*

$$\rho_c(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < c, \\ c|x| - \frac{1}{2}c^2 & \text{if } |x| \geq c. \end{cases}$$

The sequence of M-estimators associated with ρ_c solve the minimax asymptotic variance problem when c satisfies

$$2\left(\frac{\phi(c)}{c} - \Phi(c)\right) = \frac{\varepsilon}{1-\varepsilon}.$$

Remark. This corresponds to

$$\psi_c(x) := \rho_c'(x) = \begin{cases} x & \text{if } |x| < c, \\ c \operatorname{sign}(x) & \text{if } |x| \geq c. \end{cases}$$

Lemma 1.13. *Suppose that $F_* \in \mathcal{P}_\varepsilon^{\text{sym}}(\Phi)$ has density f_* such that*

$$F_* \in \operatorname*{arg\,max}_{F \in \mathcal{P}_\varepsilon^{\text{sym}}(\Phi)} V(\psi_*, F)$$

where $\psi_* = -f_*'/f_*$. Then ψ_* solves the minimax asymptotic variance solution, and

$$\min_\psi \max_{F \in \mathcal{P}_\varepsilon^{\text{sym}}(\Phi)} V(\psi, F) = V(\psi_*, F_*).$$

Proof. We already have

$$\min_{\psi} \max_{F \in \mathcal{P}_{\varepsilon}^{\mathrm{sym}}(\Phi)} V(\psi, F) \leq \max_{F \in \mathcal{P}_{\varepsilon}^{\mathrm{sym}}(\Phi)} V(\psi_*, F) = V(\psi_*, F_*).$$

To establish the other direction, we will show that $\psi_* \in \arg\min_{\psi} V(\psi, F_*)$, whence

$$\min_{\psi} \max_{F \in \mathcal{P}_{\varepsilon}^{\mathrm{sym}}(\Phi)} V(\psi, F) \geq \min_{\psi} V(\psi, F_*) = V(\psi_*, F_*).$$

To this end, compute

$$\mathbb{E}_{F_*}[\psi'] = \int \psi' f_* = -\int \psi f_*' = \int \psi \psi_* f_* = \mathbb{E}_{F_*}[\psi \psi_*] \leq \sqrt{\mathbb{E}_{F_*}[\psi^2] \mathbb{E}_{F_*}[\psi_*^2]}.$$

In particular, $\mathbb{E}_{F_*}[\psi_*'] = \mathbb{E}_{F_*}[\psi_*^2]$. Thus,

$$V(\psi, F_*) = \frac{\mathbb{E}_{F_*}[\psi^2]}{\left(\mathbb{E}_{F_*}[\psi']\right)^2} \geq \frac{\cancel{\mathbb{E}_{F_*}[\psi^2]}}{\cancel{\mathbb{E}_{F_*}[\psi^2]}\mathbb{E}_{F_*}[\psi_*^2]} = \frac{1}{\mathbb{E}_{F_*}[\psi_*^2]} = V(\psi_*, F_*).$$ □

Proof of Theorem 1.12. Set

$$f_c(x) := \frac{1-\varepsilon}{\sqrt{2\pi}} \exp(-\rho_c(x)) = \begin{cases} (1-\varepsilon)\phi(x) & \text{if } |x| < c, \\ \frac{1-\varepsilon}{\sqrt{2\pi}} \exp\left(-c|x| + \frac{1}{2}c^2\right) & \text{if } |x| \geq c, \end{cases}$$

and check that our choice of c ensures that f_c is a probability density function. The corresponding distribution F_c can be shown to belong to $\mathcal{P}_{\varepsilon}^{\mathrm{sym}}(\Phi)$, by isolating

$$h_c(x) := \frac{f_c(x) - (1-\varepsilon)\phi}{\varepsilon} \geq 0$$

and checking that this is also a probability density function. Now, for any $F := (1-\varepsilon)\Phi + \varepsilon H \in \mathcal{P}_{\varepsilon}^{\mathrm{sym}}(\Phi)$, we have

$$V(\psi_c, F) = \frac{\mathbb{E}_F[\psi_c^2]}{\left(\mathbb{E}_F[\psi_c']\right)^2} = \frac{(1-\varepsilon)\mathbb{E}_{\Phi}[\psi_c^2] + \varepsilon \mathbb{E}_H[\psi_c^2]}{\left((1-\varepsilon)\mathbb{E}_{\Phi}[\psi_c'] + \varepsilon \mathbb{E}_H[\psi_c']\right)^2}.$$

which attains its maximum precisely when $\mathbb{E}_H[\psi_c^2] = c^2$ and $\mathbb{E}_H[\psi_c'] = 0$. However, these are satisfied by H_c (supported on $\{|x| \geq c\}$, where $\psi_c(x) = c\,\mathrm{sign}(x)$), whence $V(\psi_c, F) \leq V(\psi_c, F_c)$. We have shown that F_c satisfies the conditions of Lemma 1.13, which directly gives the result. □

2. Quantification of Robustness

2.1. Influence Functions

Definition 2.1 (Influence Function). The influence function of a functional T at x with respect to a direibution F is defined by

$$\mathrm{IF}(x; T, F) = \lim_{\varepsilon \to 0^+} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon} = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} T((1-\varepsilon)F + \varepsilon\delta_x)\Big|_{\varepsilon=0}.$$

Remark. The influence function $\mathrm{IF}(x; T, F)$ is the Gateaux derivative of T at F in the direction $\delta_x - F$.

The influence function $\mathrm{IF}(x; T, F)$ quantifies infinitesimal effect on $T(F)$ of adding a new point x to F. We will call functionals with bounded influence functions *infinitesimally robust.*

Example 2.2. Let T be a linear functional. Denoting $F_t := (1-t)F + t\delta_x$, we have

$$T(F_t) = (1-t)T(F) + tT(\delta_x) \quad \Longrightarrow \quad \mathrm{IF}(x; T, F) = T(\delta_x) - T(F).$$

For instance, the influence function of the mean functional $\mu(F) := \mathbb{E}_F[X]$ is described by

$$\mathrm{IF}(\mu(F) + x; \mu, F) = x.$$

Note that this is unbounded!

Example 2.3. Let $m(G) := G^{-1}\left(\frac{1}{2}\right)$ be the median functional. Denoting $F_t := (1-t)F + t\delta_x$, note that we must have $F_t(m(F_t)) = \frac{1}{2}$, which expands to

$$(1-t)F(m(F_t)) + t\mathbf{1}(m(F_t) \geq x).$$

Differentiating this implicit equation and setting $t = 0$ gives

$$-\frac{1}{2} + f(m(F))\,\mathrm{IF}(x; m, F) + \mathbf{1}(m(F_t) \geq z) = 0 \quad \Longrightarrow \quad \mathrm{IF}(z; m, F) = \frac{\mathrm{sign}(z - m(F))}{2f(m(F))}.$$

Example 2.4. Let T be an M-estimator associated with ψ, obeying $\mathbb{E}_F[\psi(X, T(F))] = 0$. Plugging in $F_t := (1-t)F + t\delta_x$ gives

$$(1-t)\mathbb{E}_F[\psi(X, T(F_t))] + t\psi(x, T(F_t)) = 0.$$

Differentiating at $t = 0$ gives

$$\begin{aligned} -\cancel{\mathbb{E}_F[\psi(X, T(F))]} &+ (1-t)\mathbb{E}_F[\psi'(X, T(F))]\,\mathrm{IF}(x; T, F) \\ &+ \psi(x, T(F)) + t\psi'(x, T(F))\,\mathrm{IF}(x; T, F) = 0, \end{aligned}$$

hence

$$\mathrm{IF}(x; T, F) = M^{-1}\mathbb{E}_F[\psi(x, T(F))], \qquad M := -\mathbb{E}_F[\psi'(x, T(F))].$$

The influence function can be interpreted as a certain limit of the *sensitivity curve.* For $X_i \overset{\text{iid}}{\sim} F$, with estimators $T_n \xrightarrow{p} T(F)$, and under under suitable regularity conditions, we have

$$\mathrm{SC}(x, T) := n(T_{n+1}(X_1, ..., X_n, x) - T_n(X_1, ..., X_n)) \longrightarrow \mathrm{IF}(x; T, F).$$

2.2. Breakdown Points

Definition 2.5 (Breakdown Point, Asymptotic)**.** The asymptotic breakdown point of T is

$$\varepsilon^*(T, F) := \sup\left\{\varepsilon : \sup_H \|T((1-\varepsilon)F + \varepsilon H) - T(F)\| < \infty\right\}.$$

Example 2.6. Consider the mean functional $\mu(F) := \mathbb{E}_F[X]$. Then,

$$\mu((1-\varepsilon)F + \varepsilon H) - \mu(F) = \varepsilon(\mu(H) - \mu(F))$$

is unbounded over all distributions H whenever $\varepsilon > 0$. This means that $\varepsilon^*(\mu, F) = 0$.

Theorem 2.7. *Let T be an M-estimator of location associated with ψ, with ψ nondecreasing, $k_1 := -\psi(-\infty)$, $k_2 := \psi(\infty)$ both finite. Then,*

$$\varepsilon^*(T, F) = \frac{\min(k_1, k_2)}{k_1 + k_2}.$$

In particular, $\varepsilon^*(T, F) = \frac{1}{2}$ when $\psi(-\infty) + \psi(\infty) = 0$.

Definition 2.8 (Breakdown Point, Finite Sample)**.** The finite sample breakdown point of a statistic T_n at X is

$$\varepsilon^*(T_n, X) := \sup\left\{\frac{m}{n} : \sup_{d_H(X, X') = m} \|T_n(X') - T_n(X)\| < \infty\right\}.$$

Remark. Here, d_H denotes the *Hamming distance*, with

$$d_H(X, X') := \sum_{i=1}^{n} \mathbf{1}(X_i \neq X_i').$$

We may call X' an *m-replacement of X* when $d_H(X, X') = m$.

Bibliography

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732