STAT992: Robustness Theory Fall 2018

Lecture 1: Location Estimation

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, September 6*

1.1 Introduction to robustness

- Topics: this course will focus on classical notions of robustness (1960-1980).
- Be able to handle deviations from an ideal model and quantitively analyze effect of these deviations.
- Goal: develop procedures that are still reliable, reasonably efficient under small deviations from the assumed model.

1.2 Location estimation

Assume $X \sim F_0(x;\xi)$ and $\mathbb{E}_{F_0}(X) = \xi$, our goal is to estimate the location parameter ξ (in 1-dimension).

1.2.1 Preliminaries

If $X_1, ..., X_n$ are an iid sample of $F_0(x;\xi)$, then we can use the sample mean as an unbiased estimator. But in reality we often only observe a contaminated model defined as follows.

Definition 1. Consider the class of distributions with cdf in the set

$$\mathcal{P}_\epsilon(F_0) = \{F, F = (1-\epsilon)F_0 + \epsilon H, H \in M\} \tag{1.1}$$

where M is a set of all possible cdfs. Then $\mathcal{P}_\epsilon(F_0)$ is called the ϵ-neighborhood of F_0 under Huber's contamination model.

Remark 1.
- Elements of $\mathcal{P}_\epsilon(F_0)$ are mixture distributions, with $1-\epsilon$ proportion on F_0 and ϵ proportion on H.
- We could also define the ϵ-neighborhood of F_0 differently, for example,
  - Kolmogorov distance: $\{F : kolm(F, F_0) = \sup_{t\in\mathbb{R}} |F(t) - F_0(t)| \leq \epsilon\}$. It can be shown that $\mathcal{P}_\epsilon(F_0)$ is a subset of this.
  - Levy neighborhood: max side length of a square lying between F_0 and F.

So a very natural question here is if it is reasonable to use sample mean under Huber's contamination model. Suppose $\xi = 0$. If we take $X_i \sim F \in \mathcal{P}_\epsilon(F_0)$, then sample mean has expectation

$$\mathbb{E}_F(X_i) = \epsilon \mathbb{E}_H(X_i), \tag{1.2}$$

which could be arbitrarily large. Therefore the sample mean in general is biased in this setting and it is not a good choice.

1.2.2 Measuring robustness

In this section we'll introduce how to measure the robustness and show sample median is robust in some sense.

Breakdown point

Definition 2 (Breakdown point)**.** Consider a data set $X = (X_1, ..., X_n)$ and an estimator $T_n(X)$. For $m \leq n$, define

$$b(m, X, T_n) = \sup_{X' \in X_m} |T_n(X') - T_n(X)|, \tag{1.3}$$

where $X_m \subset R^n$ is the set of all alternative data sets obtained from X by changing **at most** m data points. Then breakdown point is defined by

$$\varepsilon^*(X, T_n) = \frac{1}{n} \max_{m \geq 0} \{m; b(m, X, T_n) < \infty\}. \tag{1.4}$$

Therefore, the finite sample breakdown point of en estimator is the fraction of data that can be given arbitrary values without making the estimator arbitrarily bad. For example, the breakdown point of sample mean is 0. The breakdown point of median is $\frac{1}{n}(\frac{n}{2} - 1)$ when n is even, and $\frac{1}{n}(\frac{n-1}{2})$ when n is odd.

Remark 2.

- The "at most" in Definition 2 is important. It makes sure changing all less than m data points arbitrarily bad will also not make the estimator arbitrarily bad.
- This is a finite sample definition of breakdown point. There are many variants of breakdown point, such as asymptotic notions.
- The sample median is robust in the sense of breakdown point. Indeed, it has the highest breakdown point among translation-invariant estimators, see exercise.

Minimax bias

In all cases of practical interest, we assume there is a value depending on F, $T_\infty = T_\infty(F)$, such that

$$T_n \to T_\infty(F). \tag{1.5}$$

For example, when T_n is the sample mean, then $T_\infty(F)$ is the population mean. When T_n is the sample median, then $T_\infty(F)$ is the population median(?).

Definition 3. The asymptotic bias of T_n at any $F \in \mathcal{P}_\epsilon(F_0)$ is

$$b(T_n, F, \xi) = T_\infty(F) - \xi, \tag{1.6}$$

and the maximum asymptotic bias is

$$MB(T_n, \varepsilon, \xi) = \max\{|b(T_n, F)|, F \in \mathcal{P}_\epsilon(F_0)\}. \tag{1.7}$$

Remark 3. The maximum asymptotic bias is the worst-case analysis.

Theorem 1.1 ([Huber, 1964]). *The median has the smallest maximum asymptotic bias among all translation-invariant estimates if the underlying distribution is symmetric and unimodal.*

Proof. **Step 1: calculate** $MB(Median(X_1, ..., X_n), \varepsilon, 0)$. Let F_0 has a density f_0 which is symmetric and unimodal, then $\xi = 0$. It's easy to show that the maximum asymptotic bias of the median is

$$b_\epsilon := MB(Median(X_1, X_2, ..., X_n), \varepsilon, 0) = F_0^{-1}(\frac{1}{2(1-\varepsilon)}). \tag{1.8}$$

Step 2: show $\min_{T_n \in \mathcal{T}} MB(T_n, \varepsilon, 0) = b_\epsilon$ **where** $\mathcal{T}$ **is the set of all translation-invariant estimates.** Let T_n be any translation-invariant estimate, it suffices to show that the maximum asymptotic bias of T_n in $P_\epsilon(F_0)$ is not smaller than b_ϵ. Let F_+ be the distribution with density

$$f_+(x) = \begin{cases} (1-\varepsilon)f_0(x) & x \le b_\epsilon \\ (1-\varepsilon)f_0(x - 2b_\epsilon) & \text{otherwise} \end{cases}, \tag{1.9}$$

then f_+ belongs to $\mathcal{P}_\epsilon(F_0)$. In fact,

$$f_+ = (1-\varepsilon)f_0 + \varepsilon g, \tag{1.10}$$

where $g(x) = \frac{1-\varepsilon}{\varepsilon}(f_0(x - 2b_\epsilon) - f_0(x))1_{(x > b_\epsilon)}$. It's not hard to show that g is a density.

Furthermore, we define $F_-(x) = F_+(x + 2b_\epsilon)$, which also belongs to $\mathcal{P}_\epsilon(F_0)$. Since T_n is translation invariant, so we have

$$T_\infty(F_+) - F_\infty(F_-) = 2b_\epsilon. \tag{1.11}$$

Therefore, $|T_\infty(F_+)|$ and $|T_\infty(F_-)|$ can not both be less than b_ϵ, which implies $MB(T_n, \varepsilon, 0) \ge b_\epsilon$.

Hence the proof is complete. □

Variance

Let's first recall some basic theory of maximum likelihood estimator. Suppose X_i's have pdf $f(X_i; \xi)$, under appropriate regularity conditions, MLE $\hat{\xi}_{mle} = \text{argmax} \sum_{i=1}^n \log f(x_i, \xi)$ is asymptotic normal:

$$\sqrt{n}(\hat{\xi}_{mle} - \xi) \to N(0, \frac{1}{I(\xi)}), \tag{1.12}$$

where $I(\xi) = Var(\frac{\partial \log f(x_i, \xi)}{\partial \xi})$ is the Fisher information. Furthermore, $\frac{1}{I(\xi)}$ is the minimum possible variance among all asymptotically normal estimators (Shao, 2003).

1.3 Exercise

Definition 4. An estimator T_n is translation-invariant if

$$T_n(X_1 + a, ..., X_n + a) = T_n(X_1, ..., X_n) + a \tag{1.13}$$

Show that median has the highest possible breakdown point among all translation-invariant estimators.

Proof. Let $m^* = \max_{m\geq 0}\{m; b(m, X, T_n) < \infty\}$, then It suffices to show that for all $m \leq m^*$, we have $n - m \geq m^*$. Let $X'_m = \{X_1 - a, X_2 - a, ..., X_m - a, X_{m+1}, ..., X_n\}$ and $X'_m + a = \{X_1, X_2, ..., X_m, X_{m+1} + a, ..., X_n + a\}$, then

$$T_n(X'_m) = T_n(X'_m + a) - a, \tag{1.14}$$

where $T_n(X')$ is bounded by the definition of breakdown point. Then let $a \to \infty$, we have $T_n(X'_m + a)$ also goes to ∞. Note $X'_m + a$ has $n-m$ data points which are different with X. Therefore, $n-m > m^*$. Hence we prove the statement.

□

1.4 Discussion

If we define the asymptotic bias of an estimator T_n as

$$b(T_n, F) = \lim_{n\to\infty} |\mathbb{E}_F(T_n(X_1, ..., X_n))| \tag{1.15}$$

where $X_i \sim F$, and claim that the maximum asymptotic bias of sample median among $\mathcal{P}_\epsilon(\Phi)$ achieves

$$\min_{T_n\in\mathcal{T}} \max_{F\in\mathcal{P}_\epsilon(\Phi)} b(T_n, F) = b_0 \tag{1.16}$$

where $b_0 = \Phi^{-1}(\frac{1}{2(1-\varepsilon)})$ and $\mathcal{T}$ is the set of translation invariant estimators, then we need to show

$$\max_{F\in\mathcal{P}_\epsilon(\Phi)} b(Med(X_1, ..., X_n), F) = b_0. \tag{1.17}$$

However, is this trivial? First we calculate b_0 by

$$b_0 = \max\{b; (1-\varepsilon)\Phi(b) + \varepsilon H(b) = \frac{1}{2}\}, \tag{1.18}$$

so b_0 is the maximum population median for $F \in \mathcal{P}_\epsilon(\Phi)$.

In general we have when $n \to \infty$,

$$Med(X_1, ..., X_n) \to \mu \text{ in probability}, \tag{1.19}$$

where μ is the population median of F. But is that true

$$\mathbb{E}_F(Med(X_1, ..., X_n)) \to \mu? \tag{1.20}$$

The definition of asymptotic variance given in the lecture might have the same issue.

Bibliography

[Huber, 1964] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.

STAT992: Robustness Theory Fall 2018

Lecture 2: Asymptotic theory of M-estimates

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, September 11*

2.1 M-estimates

Definition 1. Consider a nonnegative symmetric function $\rho(x)$ in $\mathbb{R}$ and assume $\rho(x)$ is nondecreasing in $\mathbb{R}^+$,

$$T_n = \text{argmin}_t \sum_{i=1}^{n} \rho(X_i - t) \tag{2.1}$$

is called M-estimator with the associated loss function ρ.

Example 1.

- Maximum likelihood estimator. For example, if we have a Gaussian family, $f(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right)$, then MLE is equivalent to minimizing

$$\sum_{i=1}^{n} (X_i - t)^2, \tag{2.2}$$

  then $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Suppose $\rho(t) = |t|$, then the M-estimator is

$$\text{argmin} \sum_{i=1}^{n} |X_i - t|. \tag{2.3}$$

  This corresponds to the sample median. Note this is also the MLE when $f(t)$ is Laplacian.

Note if ρ is convex and differentiable, solutions to the optimization problem correspond to solution of the estimating equation:

$$\sum_{i=1}^{n} \psi(X_i - t) = 0, \tag{2.4}$$

where $\psi(x) = \rho'(x)$ is continuous and monotonically increasing by Lemma 1, so in general there exists a solution in (2.4).

Remark 1. If ρ is only differentiable (we do not assume convexity), then solution of (2.4) may not exist. Furthermore, if there is a solution, it may be a local minima. This connects to the non-convexity and redescending M-estimates.

2.2 Consistency

Throughout, we assume $\psi(x)$ is nondecreasing in x. (Then $\sum_{i=1}^{n} \psi(X_i - t)$ is nonincreasing in t.) Let

$$T_n^* = \sup\{t; \sum_{i=1}^{n} \psi(X_i - t) \geq 0\} \tag{2.5}$$

and

$$T_n^{**} = \inf\{t; \sum_{i=1}^{n} \psi(X_i - t) \le 0\}, \tag{2.6}$$

then we define the M-estimator T_n to be T_n^* with probability $\frac{1}{2}$, and T_n^{**} with probability $\frac{1}{2}$ if $T_n^* \neq T_n^{**}$. Note in the discontinuous setting, there may not exist a solution of (2.4), the above definition corresponds to the point at which $\sum_{i=1}^n \psi(X_i - t)$ changes sign.

Theorem 1. *Suppose there exists a unique(?) $t_0 \in \mathbb{R}$, s.t. $\mathbb{E}_F[\psi(X-t)] \ge 0$ for $t < t_0$ and $\mathbb{E}_F[\psi(X-t)] \le 0$ for $t > t_0$, then we have T_n, T_n^*, T_n^{**} converge in probability to t_0.*

Proof. It suffices to show $T_n^* \to t_0$ and $T_n^{**} \to t_0$ in probability, then since $T_n = (1 - Z_n)T_n^* + Z_n T_n^{**}$ where Z_n are iid and $\mathbb{P}(Z_n = 1) = \mathbb{P}(Z_n = 0) = \frac{1}{2}$, by Slutsky's theorem, we have $T_n \to t_0$ in probability. We will show for every $\varepsilon > 0$,

$$\mathbb{P}(|T_n^* - t_0| \ge \varepsilon) \to 0. \tag{2.7}$$

We write

$$\mathbb{P}(|T_n^* - t_0| \ge \varepsilon) = \mathbb{P}(T_n^* - t_0 \ge \varepsilon) + \mathbb{P}(T_n^* - t_0 \le -\varepsilon) := A + B. \tag{2.8}$$

For the term B, we have

$$\mathbb{P}(T_n^* \le t_0 - \varepsilon) = \mathbb{P}(\sum_{i=1}^{n} \psi(X_i - (t_0 - \varepsilon)) \le 0) \tag{2.9}$$

by the definition of T_n^* (sign change point) and the monotonicity of $\sum_i^n \psi(X_i - t)$.

By LLN, for any $\varepsilon > 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - (t_0 - \varepsilon)) \to \mathbb{E}[\psi(X - (t_0 - \varepsilon))] \ge 0 \tag{2.10}$$

in probability, therefore $B \to 0$. Similarly we can prove $A \to 0$ and $T_n^{**} \to t_0$ in probability. □

Example 2. If $\psi(x) = x$, then $\sum_i^n \psi(X_i - t)$ is continuous and deceasing, we have $T_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $t_0 = \mathbb{E}_F(X)$. If $\psi(x) = sgn(x)$, we have

$$\mathbb{E}_F[sgn(X - t)] = \mathbb{P}(X > t) - \mathbb{P}(X < t). \tag{2.11}$$

Therefore, t_0 is the median of F. If t_0 is unique, then for $n = 2m$, the interval $(X_{(m)}, X_{(m+1)})$ shrinks to a single point when $n \to \infty$.

2.3 Asymptotic normality

Theorem 2. *Suppose there exists a $t_0 \in \mathbb{R}$, such that $\mathbb{E}_F[\psi(X - t_0)] = 0$. Assume the function $\lambda(t) = \mathbb{E}_F[\psi(X-t)]$ is differentiable at t_0 and $\lambda'(t_0) < 0$, also suppose $\sigma^2(t) = \mathbb{E}_F[\psi^2(X-t)] - \lambda^2(t)$ is finite, continuous and nonzero at t_0, then we have*

$$\sqrt{n}(T_n - t_0) \to_d N(0, \frac{\sigma^2(t_0)}{(\lambda'(t_0))^2}). \tag{2.12}$$

Corollary 1. *Suppose ρ is strictly convex and symmetric, and the distribution with cdf F is symmetric around 0 (this implies $t_0 = 0$). Then we have*

$$\sqrt{n}(T_n - 0) \to_d N(0, \frac{\mathbb{E}_F[\psi^2(X)]}{(\mathbb{E}_F[\psi'(X)])^2}) \tag{2.13}$$

if $\lambda'(0) = -\mathbb{E}_F[\psi'(X)]$.

Proof of Theorem 2. WLOG, suppose $t_0 = 0$ and define $\sigma_0 = \frac{\sigma(0)}{-\lambda'(0)}$. For any $u \in \mathbb{R}$, we write

$$\mathbb{P}(\sqrt{n}T_n^*/\sigma_0 \le u) = \mathbb{P}(T_n^* \le \frac{u\sigma_0}{\sqrt{n}}) = \mathbb{P}(\sum_{i=1}^n \psi(X_i - \frac{u\sigma_0}{\sqrt{n}}) \le 0). \tag{2.14}$$

Let

$$Y_{ni} = \frac{\psi(X_i - \frac{u\sigma_0}{\sqrt{n}}) - \lambda(\frac{u\sigma_0}{\sqrt{n}})}{\sigma(\frac{u\sigma_0}{\sqrt{n}})}, \tag{2.15}$$

then Y_{ni} are independent for a fixed n, and $\mathbb{E}(Y_{ni}) = 0$, $Var(Y_{ni}) = 1$. So if Lindeberg's condition holds, we have RHS in (2.14) is equal to

$$\mathbb{P}(\sum_{i=1}^n Y_{ni} \le \frac{-n\lambda(\frac{u\sigma_0}{\sqrt{n}})}{\sigma(\frac{u\sigma_0}{\sqrt{n}})}) \to \lim_{n\to\infty} \Phi(\frac{-\sqrt{n}\lambda(\frac{u\sigma_0}{\sqrt{n}})}{\sigma(\frac{u\sigma_0}{\sqrt{n}})}) = \Phi(\lim_{n\to\infty} \frac{-\sqrt{n}\lambda(\frac{u\sigma_0}{\sqrt{n}})}{\sigma(\frac{u\sigma_0}{\sqrt{n}})}) = \Phi(u). \tag{2.16}$$

Therefore it suffices for us to check the Lindeberg's condition: for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_{ni}^2 1_{(|Y_{ni}| > \varepsilon\sqrt{n})}] = 0. \tag{2.17}$$

Note $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_{ni}^2 1_{(|Y_{ni}| > \varepsilon\sqrt{n})}] = \lim_{n\to\infty} \mathbb{E}[Y_{ni}^2 1_{(|Y_{ni}| > \varepsilon\sqrt{n})}]$. By the continuity of λ and σ, it suffices to show

$$\lim_{n\to\infty} \mathbb{E}[\psi^2(X_i - \frac{u\sigma_0}{\sqrt{n}}) 1_{(\psi(X_i - \frac{u\sigma_0}{\sqrt{n}}) > \varepsilon\sqrt{n})}] = 0. \tag{2.18}$$

By assumption, ψ is monotone, and $\mathbb{E}[\psi^2(X_i - t)]$ is finite in a neighborhood of 0, so for n large enough, we have

$$\psi^2(X_i - \frac{u\sigma_0}{\sqrt{n}}) \le \psi^2(X_i + \delta) + \psi^2(X_i - \delta). \tag{2.19}$$

So if we define $Z_i = \psi^2(X_i - \delta) + \psi^2(X_i + \delta)$, then (2.18) is dominated by

$$\mathbb{E}[Z_i 1_{(Z_i > \varepsilon^2 n)}] \to 0 \tag{2.20}$$

as $n \to \infty$ since $\mathbb{E}[Z_i] < \infty$. Therefore we prove the theorem. □

Example 3. For the median, ψ is discontinuous, but if F has a density f, explicit calculation yields

$$\lambda(t) = \mathbb{P}(X > t) - \mathbb{P}(X < \theta) = 1 - 2F(t), \tag{2.21}$$

and hence $\lambda'(t_0) = -2f(t_0)$.

Remark 2. Recall the dominated convergence theorem: if $\lim_{n\to\infty} f_n = f$ and there exists an integrable function g such that $|f_n| \leq g$ for any n, then

$$\lim_{n\to\infty} \int f_n dv = \int \lim_{n\to\infty} f_n dv. \tag{2.22}$$

Therefore, if

$$|\frac{\partial \psi(x,t)}{\partial t}| \leq g(x) \tag{2.23}$$

for all t, and $\mathbb{E}_F[g(X)] < \infty$, we have $\lambda'(0) = -\mathbb{E}_F[\psi'(X)]$. Bounded $\psi'(t)$ satisfies the condition in DCT.

2.4 Exercise

1. Show that $\rho(t) = |t|$ corresponds to the sample median.
2. If we define

$$\psi(t) = \begin{cases} 1 & t < 0 \\ 0 & t = 0 \\ -1 & t > 0 \end{cases} \tag{2.24}$$

then the solution to the estimating equation also corresponds to the sample median.

Proof. When $\rho(t) = |t|$, the subgradient of $\sum_{i=1}^n |X_i - t|$ includes $\sum_{i=1}^n sgn(X_i - t)$, then by subgradient optimality condition, if there exists T_n such that

$$\sum_{i=1}^n sgn(X_i - T_n) = 0, \tag{2.25}$$

then we have $T_n = \text{argmin}_t \sum_{i=1}^n |X_i - t|$.

Furthermore,

$$\sum_{i=1}^n sgn(X_i - t) = |i; X_i > \mu| - |i; X_i < \mu|, \tag{2.26}$$

so if T_n is the sample median, we have $\sum_i^n sgn(X_i - T_n) = 0$. Hence we prove the above problems.

□

2.5 Appendix

Lemma 1. *If a differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex, the derivative f' is monotonically increasing and continuous.*

Proof. https://math.stackexchange.com/questions/1901753/continuity-of-derivative-of-convex-function

□

Lemma 2 (Slutsky's theorem)**.** *Let $X, X_1, ..., Y_1, Y_2, ...,$ be random variable on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_d c$, where c is a fixed real number, then*

1. $X_n + Y_n \to_d X + c$;
2. $Y_n X_n \to_d cX$;
3. $X_n / Y_n \to_d X/c$ *if* $c \neq 0$.

Lemma 3 (Subgradient optimality condition)**.** *For any f (convex or not), $f(x^*) = \min_x f(x)$ if and only if $0 \in \partial f(x^*)$.*

STAT992: Robustness Theory Fall 2018

Lecture 3: Minimax variance

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, September 13*

3.1 Problem statement

Suppose $T_n \to T(F)$ in probability and $\sqrt{n}(T_n - T(F)) \to N(0, A(T,F))$, we want to solve

$$\min_{T_n \in \mathcal{T}} \max_{F \in \mathcal{P}_\varepsilon(\Phi)} A(T,F). \tag{3.1}$$

where $\mathcal{T}$ is some class.

If T_n is the sample mean, then $\psi(x) = x$, which is unbounded, we have

$$A(T,F) = \mathbb{E}_F[(X - t_0)^2] \tag{3.2}$$

where $t_0 = \mathbb{E}_F(X)$. Therefore, if $F = (1-\varepsilon)\Phi + \varepsilon H$, then $t_0 = \varepsilon \mathbb{E}_H(X)$, then

$$A(T,F) = \mathbb{E}_F[(X - \varepsilon\mathbb{E}_H(x))^2] = (1-\varepsilon)\mathbb{E}_\Phi(X^2) + \varepsilon \mathbb{E}_H(X^2) - \varepsilon^2 \mathbb{E}_H^2(X), \tag{3.3}$$

which is unbounded.

For simplicity, we consider the symmetric contaminated distributions, $F \sim \mathcal{P}_\varepsilon(\Phi) = \{F; F = (1-\varepsilon)\Phi + \varepsilon H,\ \text{H is symmetric}\}$, then $t_0 = 0$. Note for a fixed bounded ψ, that is, $\|\psi\|_\infty \le k$ for some constant k, and assume $\psi' \ge 0$, we have

$$\begin{aligned} A(T,F) = \frac{\mathbb{E}_F[\psi^2(X)]}{[\mathbb{E}_F\psi'(X)]^2} &= \frac{(1-\varepsilon)\mathbb{E}_\Phi[\psi^2(X)] + \varepsilon\mathbb{E}_H[\psi^2(X)]}{((1-\varepsilon)\mathbb{E}_\Phi[\psi'(X)] + \varepsilon\mathbb{E}_H[\psi'(X)])^2} \\ &\le \frac{(1-\varepsilon)\mathbb{E}_\Phi[\psi^2(X)] + \varepsilon k^2}{((1-\varepsilon)\mathbb{E}_\Phi[\psi'(X)])^2} := g(\psi, k). \end{aligned} \tag{3.4}$$

3.2 Huber loss

Define

$$\rho_k(t) = \begin{cases} \frac{t^2}{2} & |t| \le k \\ k|t| - \frac{k^2}{2} & |t| > k \end{cases} \tag{3.5}$$

then we call $\rho_k(t)$ Huber loss. For Huber loss, we have

$$\psi_k(t) = \begin{cases} k & t > k \\ t & -k \le t \le k \\ -k & t \le -k \end{cases} \tag{3.6}$$

Note $\psi_k(t)$ is not differentiable at $t = -k$ and $t = k$, but it's still good. Note we have

$$\lambda(t) = \mathbb{E}_F[\psi_k(X - t)] = k(1 - F(k+t)) - kF(t-k) + \int_{t-k}^{t+k} (x-t) f(x) dx, \tag{3.7}$$

where f is the pdf of X. Therefore, we have

$$\lambda'(t) = -kf(k+t) - kf(t-k) + kf(t+k) + kf(t-k) - \int_{t-k}^{t+k} f(x)dx = -\int_{t-k}^{t+k} f(x)dx. \quad (3.8)$$

If we define $\psi_k'(k) = 1$ and $\psi_k'(-k) = 1$, then we have

$$\lambda'(t) = -\mathbb{E}[\psi_k'(X-t)]. \quad (3.9)$$

Remark 1. Note for ψ_k, the upper bound in (3.4) can be achieved by any H that puts all mass outside $(-k, k)$.

Next we use Huber loss function ρ_k to construct some distribution $f_{\rho_k} = C\exp(-\rho_k)$ which is in $\mathcal{P}_\varepsilon(\Phi)$ and the corresponding H puts all mass outside of $(-k, k)$ (k is to be determined). In oder to make

$$f_{\rho_k} = \begin{cases} (1-\varepsilon)\phi(x) + \varepsilon \times 0 & |x| \le k \\ \frac{1-\varepsilon}{\sqrt{2\pi}}\exp(-k|x| + \frac{1}{2}k^2) = (1-\varepsilon)\phi(x) + \varepsilon\big[\frac{1-\varepsilon}{\varepsilon\sqrt{2\pi}}\exp(-k|x| + \frac{1}{2}k^2) - \frac{1-\varepsilon}{\varepsilon}\phi(x)\big] & |x| > k \end{cases} \quad (3.10)$$

a distribution where $\phi(x)$ is the pdf of standard normal, it requires

$$\int_{|x|>k} \frac{1-\varepsilon}{\varepsilon\sqrt{2\pi}}\exp(-k|x| + \frac{1}{2}k^2) - \frac{1-\varepsilon}{\varepsilon}\phi(x)dx = 1, \quad (3.11)$$

which is equivalent to

$$\frac{2\phi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon}. \quad (3.12)$$

Therefore we define the distribution F_0 such that its pdf is $\frac{1-\varepsilon}{\sqrt{2\pi}}\exp(-\rho_k(x))$ with k satisfying (3.12).

3.3 Minimax variance

Now we will show Huber M-estimator is minimax optimal over some nice class. Let

$$\Psi = \{\psi;\ \text{the asymptotic variance of } \psi \text{ is given by } V(\psi, F) = \frac{\mathbb{E}_F[\psi^2(X)]}{(\mathbb{E}_F\psi'(X))^2}\}. \quad (3.13)$$

Clearly, Ψ includes all $\psi_k(t)$. Therefor, for any $\psi \in \Psi$, we have

$$\max_{F\in\mathcal{P}_\varepsilon(\Phi)} V(\psi, F) \ge V(\psi, F_0) \ge V(\psi_k, F_0) \quad (3.14)$$

where k satisfies (3.12), and the last inequality follows from the fact that MLE minimizes the asymptotic variance (see below).

Therefore, we have

$$\min_{\psi\in\Psi}\max_{F\in\mathcal{P}_\varepsilon(\Phi)} V(\psi, F) = V(\psi_k, F_0). \quad (3.15)$$

For completeness, let's verify $V(\psi, F_0) \geq V(\psi_k, F_0)$. For $\psi \in \Psi$, we have

$$\begin{aligned} V(\psi, F_0) &= \frac{\int \psi^2(x) f_0(x) dx}{(\int \psi'(x) f_0(x) dx)^2} = \frac{\int \psi^2(x) f_0(x) dx}{(\int \psi(x) f_0'(x) dx)^2} \\ &= \frac{\int \psi^2(x) f_0(x) dx}{(\int \psi(x) \frac{f_0'(x)}{f_0(x)} f_0(x) dx)^2} \geq \frac{1}{\int (\frac{f_0'(x)}{f_0(x)})^2 f_0(x) dx} \end{aligned} \tag{3.16}$$

where the second equality uses the integration by parts. The last inequality uses the Cauchy-Schwarz inequality. Furthermore, equality holds if $\psi(x) \propto \frac{f_0'}{f_0}$. Hence we prove the claim.

3.4 Exercise

What is $\max_{F \in \mathcal{P}_\varepsilon(\Phi)} A(T, F)$ when T is the median estimator? How does it compare to $\min_{\psi \in \Psi} \max_{F \in \mathcal{P}_\varepsilon(\Phi)} V(\psi, F)$ achieved by Huber loss?

Proof. We assume $F = (1-\varepsilon)\Phi + \varepsilon H$ where H is symmetric, then

$$\max_{F \in \mathcal{P}_\varepsilon(\Phi)} = \frac{1}{(1-\varepsilon)^2 \phi^2(0)}. \tag{3.17}$$

For given ε, we have a specific k by (3.12), then we have

$$V(\psi_k, F_0) = \frac{(1-\varepsilon)\mathbb{E}_\Phi(\psi_k^2) + \varepsilon k^2}{((1-\varepsilon)\mathbb{E}_\Phi \psi_k')^2} = \frac{(1-\varepsilon)[2k^2\Phi(-k) + \int_{-k}^{k} x^2 \phi(x) dx] + \varepsilon k^2}{(1-\varepsilon)^2 (2\Phi(k) - 1)^2} \tag{3.18}$$

□

STAT992: Robustness Theory Fall 2018

Lecture 4: Minimax variance

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, September 18*

4.1 Recap

Theorem 1. *Suppose there exists a cdf $F_0 \in \mathcal{P}_\epsilon(\Phi)$, such that if we define $\psi_0 = \frac{-f_0'}{f_0}$, the following holds:*

$$F_0 \in argmax_{F \in \mathcal{P}_\epsilon(\Phi)} V(\psi_0, F), \tag{4.1}$$

then

$$\min_{\psi \in \Psi} \max_{F \in \mathcal{P}_\epsilon(\Phi)} V(\psi, F) = V(\psi_0, F_0). \tag{4.2}$$

Therefore, ψ_0 is minimax optimal.

Proof. By MLE efficiency, we have $\psi_0 \in \text{argmin}_{\psi \in \Psi} V(\psi, F_0)$. Then for any $\psi \in \Psi$, we have

$$\max_{F \in \mathcal{P}_\epsilon(\Phi)} V(\psi, F) \geq V(\psi, F_0) \geq V(\psi_0, F_0). \tag{4.3}$$

Therefore, we have

$$\min_{\psi \in \Psi} \max_{F \in \mathcal{P}_\epsilon(\Phi)} V(\psi, F) \geq V(\psi_0, F_0). \tag{4.4}$$

Furthermore, by the assumption that $V(\psi_0, F_0) = \max_{F \in \mathcal{P}_\epsilon(\Phi)} V(\psi_0, F)$, we prove the theorem. □

4.2 Minimax variance for $\mathcal{P}_\epsilon(G)$

Let G be arbitrary symmetric distribution and $\mathcal{P}_\epsilon(G) = \{F : F = (1-\varepsilon)G + \varepsilon H, \text{H is symmetric}\}$. We want to solve

$$\min_{\psi \in \Psi} \max_{F \in \mathcal{P}_\epsilon(G)} V(\psi, F). \tag{4.5}$$

we have a more general theorem in the following.

Theorem 2. *Suppose G is a cdf of a symmetric distribution with twice continuously differentiable pdf g, such that $-\log g(x)$ is convex, then we have*

(i) $V(\psi, F)$ has a saddle point. That is

$$\max_{F \in \mathcal{P}_\epsilon(G)} V(\psi_0, F) = V(\psi_0, F_0) = \min_{\psi \in \Psi} V(\psi, F_0). \tag{4.6}$$

Hence, ψ_0 is minimax optimal.

(ii) Furthermore, the explicit expression of ψ_0 *is* $\psi_0 = \frac{-f_0'}{f_0}$, *and*

$$f_0(x) = \begin{cases} (1-\varepsilon)g(x_0)\exp\left(k(x-x_0)\right) & x \le x_0 \\ (1-\varepsilon)g(x) & x_0 < x < x_1 \\ (1-\varepsilon)g(x_1)\exp\left(-k(x-x_1)\right) & x \ge x_1 \end{cases} \tag{4.7}$$

where $x_0 < x_1$ *are endpoints of the interval where* $|g'/g| \le k$ *and* k *is related to* ε *through*

$$\frac{1}{1-\varepsilon} = \int_{x_0}^{x_1} g(x)dx + \frac{g(x_0)+g(x_1)}{k}, \tag{4.8}$$

and F_0 *is the cdf corresponding to* f_0.

Remark 1.

- (4.8) is required to ensure that f_0 integrates to 1. It always has a solution. (How to show?)
- Need to show $F_0 \in \mathcal{P}_\epsilon(G)$, that is $F_0 = (1-\epsilon)G + \epsilon H$ for some symmetric distribution H. Note the pdf of H is

$$h(x) = \begin{cases} \frac{1-\varepsilon}{\varepsilon}[g(x_0)\exp\left(k(x-x_0)\right) - g(x)] & x \le x_0 \\ 0 & x_0 < x < x_1 \\ \frac{1-\varepsilon}{\varepsilon}[g(x_1)\exp\left(-k(x-x_1)\right) - g(x)] & x \ge x_1 \end{cases} \tag{4.9}$$

  since $-\log g$ is convex, so we have $h(x) \ge 0$.
- Since $-\log g(x)$ is convex and g is twice differentiable, then $-\frac{g'}{g}$ is nondecreasing. If x_0 and x_1 are finite, then by continuity, we have

$$\frac{g'(x_0)}{g(x_0)} = k, \frac{g'(x_1)}{g(x_1)} = -k. \tag{4.10}$$

  Therefore, we have f_0 and its derivative are continuous, and

$$\psi_0(x) = \begin{cases} -k & x \le x_0 \\ \frac{-g'(x)}{g(x)} & x_0 < x < x_1 \\ k & x \ge x_1 \end{cases} \tag{4.11}$$

  hence this result may be described as a **truncated MLE**.
- When $G = \Phi$, then $g(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$, so we have $x_0 = -k, x_1 = k$, then (4.8) becomes

$$\frac{2\phi(k)}{k} - 2\Phi(-k) = \frac{\epsilon}{1-\epsilon}. \tag{4.12}$$

Proof of Theorem 2. Only need to show $F_0 = \text{argmax}_{F\in\mathcal{P}_\epsilon(G)} V(\psi_0, F)$. For any $F = (1-\epsilon)G + \epsilon H$, we have

$$\begin{aligned} V(\psi_0, F) &= \frac{(1-\epsilon)\mathbb{E}_G(\psi_0^2) + \epsilon\mathbb{E}_H(\psi_0^2)}{((1-\epsilon)\mathbb{E}_G\psi_0' + \epsilon\mathbb{E}_H\psi_0')^2} \\ &\le \frac{(1-\epsilon)\mathbb{E}_G(\psi_0^2) + \epsilon k^2}{((1-\epsilon)\mathbb{E}_G\psi_0')^2} \end{aligned} \tag{4.13}$$

where the last inequality comes from (4.11). Note when H has pdf $h(x)$ given in (4.9), the upper bound is achieved. Hence, we prove the theorem. □

4.3 Distributions minimizing Fisher information

4.3.1 Equivalence of two problems

We define the Fisher information at the distribution $F : I(F) = \int (\frac{f'}{f})^2 f dx$ where f is the pdf of F. We first state the following important and useful theorem.

Theorem 3. *Under the same assumption with Theorem 2, the following two problems are equivalent:*

1. $\min_{\psi\in\Psi} \max_{F\in\mathcal{P}_\epsilon(G)} V(\psi, F)$
2. $\min I(F)$

Proof. Let $V(\psi_0, F_0) = \min_\Psi \max_{\mathcal{P}_\epsilon(G)} V(\psi, F)$ and $I(F^*) = \min I(F)$. For any $\psi \in \Psi$, we have

$$\max_{F\in\mathcal{P}_\epsilon} V(\psi, F) \geq V(\psi, F^*) \geq V(\psi^*, F^*). \tag{4.14}$$

where $\psi^* = -\frac{(f^*)'}{f^*}$ and f^* is the pdf of F^* and the last inequality follows from the efficiency of MLE. Therefore,

$$\frac{1}{I(F_0)} = V(\psi_0, F_0) = \min_\Psi \max_{F\in\mathcal{P}_\epsilon(G)} V(\psi, F) \geq V(\psi^*, F^*) = \frac{1}{I(F^*)}. \tag{4.15}$$

Furthermore, we have

$$I(F^*) \leq I(F_0), \tag{4.16}$$

so we have

$$I(F_0) = I(F^*). \tag{4.17}$$

If there is a unique solution of $\min I(F)$, then we have $F_0 = F^*$.

□

Remark 2. See proposition 4.5 in Huber's book on uniqueness.

4.3.2 Sufficient and necessary condition

Lemma 1. *I(F) is a convex function of F.*

Proof. Note $I(F) = \sup_\psi \frac{(\int \psi' dF)^2}{\int \psi^2 dF}$ and $\frac{(\int \psi' dF)^2}{\int \psi^2 dF}$ is a convex function of F. Since the supremum of the convex functions is convex, so we prove the lemma. □

Assume $F_0 = \text{argmin}_F I(F)$. Let $F_t = (1-t)F_0 + tF_1$ for any $F_1 \in \mathcal{P}_\epsilon(G)$. Since $I(F_t)$ is a convex function of t, so by the optimality condition, we have

$$[\frac{d}{dt} I(F_t)]_{t=0} = \int \frac{2f_0'}{f_0}(f_1' - f_0') - (\frac{f_0'}{f_0})^2 (f_1 - f_0) dx \geq 0. \tag{4.18}$$

By integration by parts, we have

$$\int \frac{f_0'}{f_0}(f_1' - f_0')dx = -\int (\frac{f_0'}{f_0})'(f_1 - f_0)dx. \tag{4.19}$$

Therefore, we have

$$[\frac{d}{dt}I(F_t)]_{t=0} = \int (2\psi_0' - \psi_0^2)(f_1 - f_0)dx \geq 0 \tag{4.20}$$

for any $F_1 \in \mathcal{P}_\epsilon(G)$.

4.3.3 An example

Recall the Kolmogorov neighborhood of Φ is

$$\mathcal{P}_\epsilon^k(\Phi) = \{F; \sup_t |F(t) - \Phi(t)| \leq \epsilon\}. \tag{4.21}$$

Note $\mathcal{P}_\epsilon(\Phi) \subset \mathcal{P}_\epsilon^k(\Phi)$. We want to solve the problem $\min_{F \in \mathcal{P}_\epsilon^k(\Phi)} I(F)$. Therefore, it suffices to find f_0 satisfying

$$\int (2\psi_0' - \psi_0^2)(f_1 - f_0)dx \geq 0 \tag{4.22}$$

for all $F_1 \in \mathcal{P}_\epsilon^k(\Phi)$.

We use the following ideas: like Huber function, we hope ψ_0 is a constant on some interval, so we have $f_0(x) \propto \exp(-cx)$. See more details from Example 4.3 in Huber's book.

4.4 Exercise

Show that Huber function $\psi_k(t)$ is in the class Ψ.

Proof. See lecture 3. □

4.5 Appendix

Lemma 2 (Optimality condition for constrained convex problem)**.** *For a convex problem,*

$$\min f(x), \ \textit{subject to } x \in C \tag{4.23}$$

and differentiable f, a feasible point x is optimal if and only if

$$\nabla f(x)^T (y - x) \geq 0 \tag{4.24}$$

for all $y \in C$.

STAT992: Robustness Theory Fall 2018

Lecture 5: Hampel's approach to robustness

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, September 20*

5.1 Introduction to Hampel's approach

Huber theory is nice, but calculations depend on symmetry, normality, M-estimators and etc. Hampel looked at notions such as:

1. Qualitative robustness (continuity of $T(F)$).
2. Influence function (effect of infinitesimal perturbation).
3. Breakdown point.

5.2 Influence function

Assume $\{T_n\}$ is a sequence of estimators such that for $X_i \sim F$, we have

$$T_n(X_1, ..., X_n) \to T(F) \tag{5.1}$$

in probability.

Definition 1. The Gateaux derivative of a functional $T : \mathcal{M} \to \mathbb{R}$ at F in the direction H is defined by

$$\lim_{t \to 0} \frac{T(F + t(H - F)) - T(F)}{t} := T'_F(H - F) \tag{5.2}$$

From a mathematical perspective, the Gateaux derivative is a generalization of the concept of a directional derivative to functional analysis. From a statistical perspective, it represents the rate of change in a statistical functional upon a small amount of contamination by another distribution.

Definition 2. The influence function $IF(\cdot; T, F) : \mathbb{R} \to \mathbb{R}$ of a functional T at F is

$$IF(x; T, F) = \lim_{t \to 0} \frac{T((1 - t)F + t\Delta_x) - T(F)}{t} \tag{5.3}$$

where Δ_x is the point mass at x.

The influence function measures the infinitesimal perturbations in the direction of a point mass at x.

5.3 Exercise

What is the worse-case bias for Huber estimator over $\mathcal{P}_\epsilon(\Phi) = \{(1-\epsilon)\Phi + \epsilon H\}$ where H is any distribution? Compare the result with worst-case bias for median $\Phi^{-1}(\frac{1}{2(1-\epsilon)})$.

Part I: For any distribution $F \in \mathcal{P}_\epsilon(\Phi)$, the asymptotic bias of Huber estimator is is the solution of

$$\mathbb{E}_F \psi_k(x-b) = 0. \tag{5.4}$$

Therefore, we have

$$(1-\epsilon)\mathbb{E}_\Phi \psi_k(x-b) + \epsilon \mathbb{E}_H \psi_k(x-b) = 0. \tag{5.5}$$

Now we hope $|b|$ is as large as possible. Since $\|\psi_k\|_\infty \le k$, so we have

$$|\mathbb{E}_H \psi_k(x-b)| \le \mathbb{E}_H |\psi_k(x-b)| \le k. \tag{5.6}$$

Therefore we have

$$-\frac{k\epsilon}{1-\epsilon} \le \mathbb{E}_\Phi \psi_k(x-b) \le \frac{k\epsilon}{1-\epsilon}. \tag{5.7}$$

Since $\frac{d\mathbb{E}_\Phi \psi_k(x-b)}{db} = -\mathbb{E}_\Phi \psi_k'(x-b) < 0$, so $\mathbb{E}_\Phi \psi_k(x-b)$ is decreasing. Furthermore, $\mathbb{E}_\Phi \psi_k(x-b)$ is an odd function. Therefore, the maximum $|b|$ is achieved when

$$\mathbb{E}_\Phi \psi_k(x-b) = \frac{-k\epsilon}{1-\epsilon}. \tag{5.8}$$

or

$$k[1-\Phi(k+b)-\Phi(b-k)] + \int_{b-k}^{b+k} (x-b)\phi(x)dx = \frac{-k\epsilon}{1-\epsilon}. \tag{5.9}$$

where $\phi(x)$ is the pdf of standard normal distribution. Note (5.8) has the solution if only if $\epsilon < 0.5$ (when $b \to \infty$, $\mathbb{E}_\Phi \psi_k(x-b) \to -k$; When $b \to -\infty$, $\mathbb{E}_\Phi \psi_k(x-b) \to k$). For the maximum bias b, we have $\mathbb{E}_H \psi_k(x-b) = k$. Therefore, H can be any distribution which puts mass on $(k+b, \infty)$.

When $\epsilon \ge 0.5$, the maximum bias is ∞. Then we have $\mathbb{E}_H \psi_k(x-b) = \frac{(1-\epsilon)k}{\epsilon} \le k$.

Part II: Since Huber estimator is translation invariant, so by previous theorem, we have the worst-case bias for Huber estimator is larger than $\Phi^{-1}(\frac{1}{2(1-\epsilon)})$.

	k=1	k=2	k=3	Median
$\epsilon = 0.1$	0.1633	0.2333	0.3344	0.1397
$\epsilon = 0.2$	0.3722	0.5295	0.7543	0.3186
$\epsilon = 0.5$	∞	∞	∞	∞
$\epsilon = 0.6$	∞	∞	∞	∞

STAT992: Robustness Theory Fall 2018

Lecture 6: Influence function

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, September 25*

6.1 Related quantities

1. Gross error sensitivity:
$$\gamma^*(T,F) = \sup_x |IF(x;T,F)| \tag{6.1}$$
2. Local-shift sensitivity:
$$\lambda^*(T,F) = \sup_{x\neq y} \frac{|IF(x;T,F) - IF(y;T,F)|}{|x-y|} \tag{6.2}$$
3. Rejection point:
$$\rho^*(T,F) = \inf\{r>0, IF(x;T,F) = 0 \text{ when } |x|>r\} \tag{6.3}$$
Estimators with finite rejection point ignore bad enough outliers.

6.2 Von Mises Calculus

Given $X_1,...,X_n$ from a distribution F, and a parameter of interest $T(F)$, suppose that we estimate $T(F)$ by $T(F_n)$, where F_n is the empirical distribution. Also assume $T_n(X_1,...,X_n) \to^p T(F)$.

Define the derivative $T_F^{(k)}(H)$ of the map $t \to T(F+tH)$ at t=0, where H is a perturbation direction. Then if the derivatives exist we have a Taylor series expansion:
$$T(F+tH) - T(F) = tT_F'(H) + \frac{1}{2}t^2T_F^{(2)}(H) + ... + \frac{1}{m!}t^mT_F^{(m)}(H) + o(t^m). \tag{6.4}$$
Substituting $t = 1/\sqrt{n}$ and $H = G_n$ where $G_n = \sqrt{n}(F_n - F)$, we obtain the von Mises expansion:
$$T(F_n) - T(F) = \frac{1}{\sqrt{n}}T_F'(G_n) + \frac{1}{2n}T_F^{(2)}(G_n) + ... + \frac{1}{m!n^{m/2}}T_F^{(m)}(G_n) + o(n^{-m/2}). \tag{6.5}$$

If T_F' is a linear map, we have
$$T(F_n) - T(F) \approx \frac{1}{\sqrt{n}}T_F'(G_n) = T_F'(F_n - F) = \frac{1}{n}\sum_{i=1}^n T_F \tag{6.6}$$

Therefore, we have
$$\sqrt{n}(T(F_n) - T(F)) = \frac{1}{\sqrt{n}}T_F'(\Delta_{X_i} - F) = \frac{1}{\sqrt{n}}\sum_{i=1}^n IF(X_i;T,F). \tag{6.7}$$

By CLT, we have $\sqrt{n}(T(F_n) - T(F)) \to N(0, A(T,F))$ where $A(T,F) = \int (IF(x;T,F))^2 dF(x)$.

Remark 1. In some cases, we have $T(F_n) = T_n$; in other cases, $T(F_n) \approx T_n$.

6.3 Functional Delta method

Recall the following Delta method:

Theorem 1. *If $\phi : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable at θ, and $\sqrt{n}(T_n - \theta) \to^d T$, then*

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \to^d \phi'_\theta(T) \tag{6.8}$$

Here $\phi'(\theta)$ is the derivative (linear map) satisfying

$$\phi(\theta + h) - \phi(h) = \phi'_\theta(h) + o(\|h\|). \tag{6.9}$$

for $h \to 0$.

See Van der Vaart Chapter 20.

6.4 Exercise

For $0 \leq \alpha \leq \frac{1}{2}$, the α-trimmed mean is defined by

$$T_n(X_1, ..., X_1) = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} X_{(i)}, \tag{6.10}$$

where $m = \lfloor \alpha n \rfloor$. If we define the functional

$$T(F) = \frac{1}{1 - 2\alpha} \int_\alpha^{1-\alpha} F^{-1}(s) ds = \frac{1}{1 - 2\alpha} \mathbb{E}_F(X 1_{\alpha \leq F(X) \leq 1-\alpha}) \tag{6.11}$$

It can be shown that $T_n \to^p T(F)$ if $X_i \sim F$.

- Write the expression for $T(F_n)$, and check that in general, $T(F_n) \neq T_n$, but they are close.
- Compute the influence function of T.

Proof.

Part 1: We have

$$T(F_n) = \frac{1}{1 - 2\alpha} \frac{1}{n} \sum_{i=\lceil na \rceil}^{\lfloor n(1-\alpha) \rfloor} X_{(i)} \tag{6.12}$$

Part 2: See notes.

□

ECE901/CS838/STAT992: Robustness Theory Fall 2018

Lecture 10: Reconciling Huber's and Hampel's approaches to robustness

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, October 11*

In previous lectures, we have introduced two common approaches, Huber's minimax approach and Hampel's influence function approach, to study robustness. In today's lecture, we will show these two approaches are equivalent in the sense that the minimax bias criterion corresponds to the most B-robustness and the minimax variance criterion corresponds to the optimal B-robustness (for the class of monotone ψ functions).

10.1 A brief review

- An estimator minimizing gross-error sensitivity γ^* is called a most B-robust estimator. An estimator that minimizes the asymptotic variance to a bound on γ^* is called the optimal B-robust estimator.
- The optimal B-robust estimator is given by truncation:

$$\tilde{\psi}(y) = [s(y,\theta) - a]_{-b}^{b}.$$

- For location M-estimator such that $\mathbb{E}_F[\psi(X - T(F))] = 0$ and the density of F is twice continuously differentiable, symmetric and log-concave. Consider the class of functions Ψ which satisfies

  1. $C(\psi)$, the set of discontinuity points of ψ, is finite. At each point in $C(\psi)$, there exist finite left and right limits of ψ which are different. Also, $\psi(-x) = -\psi(x)$ if $\{x, -x\} \subset R \backslash C(\psi)$, and $\psi(x) \geq 0$ for $x \notin C(\psi)$ and $x \geq 0$.
  2. $D(\psi)$, points at which ψ is continuous but ψ' is discontinuous or undefined, is finite.
  3. $\int \psi^2(x) dF(x) < \infty$.
  4. $0 < \int \psi'(x) dF(x) < \infty$.

  Then median is the unique most B-robust estimator over Ψ.

  Remark 1. The conditions on ψ doesn't necessarily imply ψ is monotone, but ψ has to be odd.

10.2 Minimax bias and most B-robustness

Recall for symmetric F, we define $\mathcal{P}_\epsilon(F) = \{(1-\epsilon)F + \epsilon H, \text{ where H is any distribution}\}$. Huber considered the minimax bias problem

$$\min_{\psi} \sup_{G \in \mathcal{P}_\epsilon(F)} |T(G) - T(F)| \tag{10.1}$$

where T is the functional corresponding to some odd function ψ. For small ϵ and assume T is linear, we have

$$T((1-\epsilon)F+\epsilon H)-T(F) \approx \int [T((1-\epsilon)F+\epsilon\Delta_x)-T(F)]dH(x) = \epsilon \int IF(x;\psi,F)dH(x).$$

Therefore,

$$\begin{aligned} \sup_{G\in\mathcal{P}_\epsilon(F)} |T(G)-T(F)| &= \sup_H |T((1-\epsilon)F+\epsilon H)-T(F)| \\ \approx \sup_H |\epsilon \int IF(x;\psi,F)dH(x)| &= \epsilon \sup_x |IF(x;\psi,F)| = \epsilon\gamma^*(\psi,F). \end{aligned} \tag{10.2}$$

Hence (10.1) corresponds to

$$\min_\psi \gamma^*(\psi,F). \tag{10.3}$$

As we have shown before, (10.1) and (10.3) yield the same estimator, the median.

Next let us visualize the connections between the above various robustness notions. We make a plot of the maximal asymptotic bias $\sup_{G\in\mathcal{P}_\epsilon(F)} |T(G)-T(F)|$ as a function of the fraction ϵ of contamination. The value of ϵ for which this function becomes infinite is the asymptotic contamination breakdown point ϵ^*. If the function is continuous at $\epsilon = 0$, the estimator is "qualitatively robust" in the sense that the maximum possible asymptotic bias goes to 0 as the fraction of contamination goes to 0. For "qualitatively robust" estimators, the gross-error sensitivity γ^* of the estimator equals the slope of this function at $\epsilon = 0$ by (10.2).

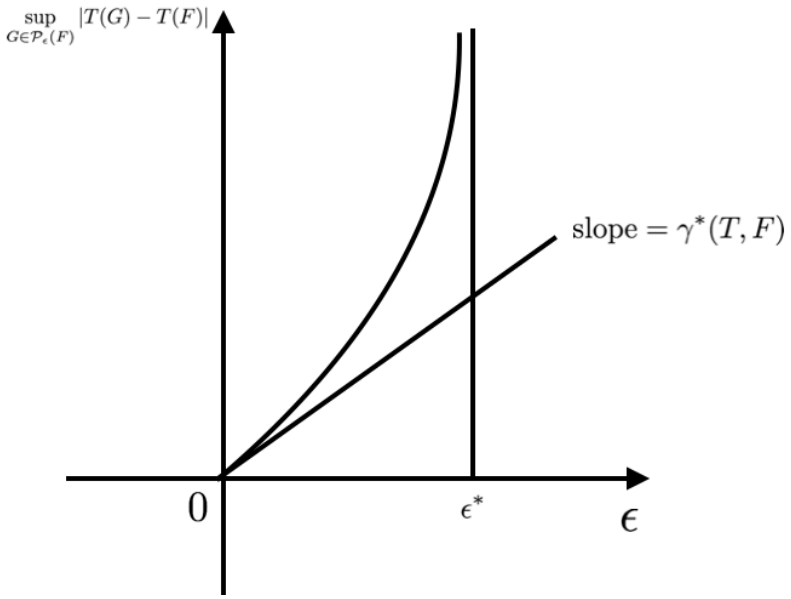

Figure 10.1: Plot of $\sup_{G\in\mathcal{P}_\epsilon(F)} |T(G)-T(F)|$ as a function of ϵ.

Remark 2. Note (10.2) is an approximation. From Figure 10.2, the breakdown point ϵ^* tells us the radius in which the linear approximation is valid. Clearly, for $\epsilon \geq \epsilon^*$, the approximation fails completely.

10.3 Minimax variance and optimal B-robustness

Assume we are considering location M-estimators over class of estimators Ψ and class of distributions F defined in Section 10.1. Define

$$A(\psi) = \int \psi^2(x)dF(x), B(\psi) = \int \psi'(x)dF(x),$$

then the influence function is $IF(x;\psi,F) = \frac{\psi(x)}{B(\psi)}$ and the asymptotic variance is $V(\psi,F) = \int IF^2(x;\psi,F)dF(x) = \frac{A(\psi)}{B^2(\psi)}$.

We define change-of-variance function and change-of-variance sensitivity in the following.

Definition 1. The change-of-variance function $CVF(x;\psi,F)$ of ψ at F is defined as the sum of the regular part

$$\frac{A(\psi)}{B(\psi)^2}(1 + \frac{\psi^2(x)}{A(\psi)} - 2\frac{\psi'(x)}{B(\psi)})1_{\mathbb{R}\backslash\{C(\psi)\cup D(\psi)\}}(x)$$

which is continuous on $\mathbb{R}\backslash\{C(\psi)\cup D(\psi)\}$, and

$$\frac{A(\psi)}{B(\psi)^2}(-\frac{2}{B(\psi)}[\sum_{i=1}^{m}(\psi(c_i+) - \psi(c_i-))\delta_{c_i}(x)]).$$

For continuous ψ, it is equivalent to

$$CVF(x;\psi,F) = \frac{\partial}{\partial t}V(\psi,(1-t)F + t(\frac{1}{2}\Delta_x + \frac{1}{2}\Delta_{-x}))|_{t=0}.$$

Definition 2. The change-of-variance sensitivity $\kappa^*(\psi,F)$ is defined as $+\infty$ if a delta function with positive factor occurs in CVF, and otherwise

$$\kappa^*(\psi,F) = \sup_{x\in\mathbb{R}\backslash\{C(\psi)\cup D(\psi)\}} \frac{CVF(x;\psi,F)}{V(\psi,F)}.$$

If $\kappa^*(\psi,F) < \infty$, then we call the estimator V-robust.

Theorem 1. *1. Suppose $\psi \in \Psi$ is V-robust, then*

$$\gamma^*(\psi,F) \le \sqrt{(\kappa^*(\psi,F)-1)V(\psi,F)}. \tag{10.4}$$

Thus V-robustness implies B-robustness.

2. If ψ is nondecreasing, then the inequality in (10.4) is an equality. Thus, B-robustness and V-robustness are equivalent.

Proof. See Theorem 1 and Theorem 2 in Section 2.5b in [Hampel et al., 2011]. □

Remark 3. In general we do not have equivalence of V-robustness and B-robustness, as is exemplified by the Huber-type skipped mean, which is B-robust but not V-robust. See more details from the exercise.

Recall Huber's minimax variance problem is stated as

$$\min_{\psi} \max_{G \in \mathcal{P}_\epsilon(F)} V(\psi, G). \tag{10.5}$$

We can make a good approximation of $\max_{G \in \mathcal{P}_\epsilon(F)} V(\psi, G)$ by means of the change-of-variance sensitivity κ^* for small ϵ:

$$\begin{aligned}
\sup_{G \in \mathcal{P}_\epsilon(F)} V(\psi, G) &= \exp\left(\sup_H [\ln V(\psi, (1-\epsilon)F + \epsilon H)]\right) \\
&\approx \exp\left(\sup_H \left[\ln V(\psi, F) + \epsilon \int \frac{CVF(x; \psi, F)}{V(\psi, F)} dH(x)\right]\right) \\
&= \exp\left(\ln V(\psi, F) + \epsilon \sup_x \frac{CVF(x; \psi, F)}{V(\psi, F)}\right) \\
&= V(\psi, F) \exp\left(\epsilon \kappa^*(\psi, F)\right),
\end{aligned}$$

where the approximation uses the Taylor expansion of $\ln V$. Therefore, (10.5) is equivalent to

$$\min_{\psi} V(\psi, F) \exp\left(\epsilon \kappa^*(\psi, F)\right). \tag{10.6}$$

Note any solution $\tilde{\psi}$ of (10.6) is also a minimizer of

$$\begin{aligned}
\min_{\psi} \quad & V(\psi, F) \\
\text{subject to} \quad & \kappa^*(\psi, F) \leq \kappa^*(\tilde{\psi}, F).
\end{aligned} \tag{10.7}$$

Over the class of monotone ψ functions, by Part 2 of Theorem 1, the problem (10.7) may equivalently be written as

$$\begin{aligned}
\min_{\psi} \quad & V(\psi, F) \\
\text{subject to} \quad & \gamma^*(\psi, F) \leq \gamma^*(\tilde{\psi}, F).
\end{aligned} \tag{10.8}$$

Hence $\tilde{\psi}$ is also optimal B-robust. In the case $F = \Phi$, we conclude that $\tilde{\psi}$ is the Huber estimator, using the theorem on optimal B-robust estimators from before.

10.4 Exercise

Show that Part 2 of Theorem 1 is not necessarily true for non-monotone ψ. In particular, consider skipped mean estimator which corresponds to $\psi(x) = x * 1_{\{|x| \leq r\}}$.

Proof. For symmetric F, the influence function is

$$IF(x; \psi, F) = \frac{\psi(x)}{\int \psi' dF(x)}. \tag{10.9}$$

Therefore, the gross-error sensitivity is

$$\gamma^* = \frac{r}{F(r) - F(-r)}. \tag{10.10}$$

On the other hand, we have

$$CVF(x;\psi,F)=\frac{A(\psi)}{B(\psi)^2}(1+\frac{\psi^2(x)}{A(\psi)}-2\frac{\psi'(x)}{B(\psi)})1_{\mathbb{R}\setminus\{-r,r\}}(x)+\frac{A(\psi)}{B(\psi)^2}(\frac{2r}{B(\psi)}(\delta_{-r}(x)+\delta_r(x))). \quad (10.11)$$

where $B(\psi)=F(r)-F(-r)>0$. Therefore the change-of-variance sensitivity is $+\infty$. Hence the skipped mean estimation is B-robust but not V-robust. □

Bibliography

[Hampel et al., 2011] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.

ECE901/CS838/STAT992: Robustness Theory Fall 2018

Lecture 11: Robust linear regression

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, October 16*

11.1 Problem setup

Assume the linear regression model $y_i = \sum_{j=1}^{p} X_{ij}\theta_j + u_i, 1 \leq i \leq n$ where $X_i = (X_{i1}, ..., X_{ip}) \in \mathbb{R}^p$ are iid from distribution with CDF $K(X)$ and $u_i \in \mathbb{R}$ are iid from distribution with G_σ, and X_i are independent with u_i. Then the joint distribution of (X_i, y_i) parameterized by (θ, σ) is

$$f_{\theta,\sigma}(X_i, y_i) = f(X_i)f(y_i|X_i) = k(X_i)\frac{1}{\sigma}g(\frac{y_i - X_i^T\theta}{\sigma}) \tag{11.1}$$

Maximum likelihood estimator corresponds to maximizing $\prod_{i=1}^{n} f_{\theta,\sigma}(X_i, y_i)$. In particular, if $g(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2)$, then the MLE is equivalent to minimizing $\sum_{i=1}^{n}(y_i - X_i^T\theta)^2$ assuming σ is known. Hence ordinary least squares is efficient when $G = \Phi$. However, under deviation from normality on distribution of u_i, this is not a robust estimator (see influence function calculation below).

11.2 Regression M-estimators

Instead we look at an M-estimator with loss function ρ:

$$\min_{\theta} \sum_{i=1}^{n} \rho(y_i - X_i^T\theta). \tag{11.2}$$

Then the corresponding estimating equation is

$$\sum_{i=1}^{n} \psi(y_i - X_i^T\theta)X_i = 0. \tag{11.3}$$

and the corresponding functional should satisfy

$$\int \psi(y - X^T T(F))X dF(X, y) = 0. \tag{11.4}$$

Furthermore, we can define the following generalized M-estimators

$$\sum_{i=1}^{n} \eta(X_i, X_i^T\theta - y_i)X_i = 0, \tag{11.5}$$

where η takes the form $\eta(X_i, r_i) = w(X_i)\psi(r_i v(X_i))$, w and v are functions from $\mathbb{R}^p$ to $\mathbb{R}$. Type of weight functions might be

$$w(X_i) = \frac{1}{\|AX_i\|_2}. \tag{11.6}$$

Then the corresponding functional should satisfy

$$\int \psi((y - X^T T(F))v(X))W(X)XdF(X,y) = 0 \tag{11.7}$$

11.2.1 Influence function

Next we calculate the influence function. For simplicity, we consider the functional $T(F)$ which solves

$$0 = \int X\psi(y - X^T\theta)dF(X,y). \tag{11.8}$$

Define $F_t = (1-t)F + t\Delta_{(X_0,y_0)}$, then we have

$$0 = \int \psi(y - X^T T(F_t))XdF_t(X,y) = (1-t)\int \psi(y - X^T T(F_t))XdF(X,y) + t\int \psi(y - X^T T(F_t))Xd\Delta_{(X_0,y_0)} \tag{11.9}$$

Differentiating with respect to t, we have

$$\begin{aligned} 0 = (1-t)\int \psi'(y - X^T T(F_t))(-X^T \frac{d}{dt}T(F_t))XdF(X,y) - \int \psi(y - X^T T(F_t))XdF(X,y) \\ +t\int \psi'(y - X^T T(F_t))(-X^T \frac{d}{dt}T(F_t))Xd\Delta_{(X_0,y_0)} + \int \psi(y - X^T T(F_t))Xd\Delta_{(X_0,y_0)} \end{aligned} \tag{11.10}$$

Plugging in $t = 0$, we have

$$0 = \int \psi'(y - X^T T(F))(-XX^T)IF((X_0,y_0);T,F)dF(X,y) + \psi(y_0 - X_0^T T(F))X_0. \tag{11.11}$$

Therefore, we have

$$IF((X_0,y_0);T,F)) = M^{-1}\psi(y_0 - X_0^T T(F))X_0 \tag{11.12}$$

where $M = \int \psi'(y - X^T T(F))XX^T dF(X,y)$. Furthermore, by independence of u_i and X_i, we have

$$M = (\int \psi'(u)dG(u))(\int XX^T dK(X)) \tag{11.13}$$

If we write $u_0 = y_0 - X_0^T T(F)$, then we can think of influence function as having two factors:

$$IF((X_0,y_0);T,F) = \frac{\psi(u_0)}{\mathbb{E}_{u\sim G}[\psi'(u)]} \times [(\mathbb{E}_{X\sim K}[XX^T])^{-1}X_0] \tag{11.14}$$

where the first term is called influence of residual and the second term is called influence of position. Therefore influence function is bounded if ψ is bounded. Hence, from above, we see OLS is not robust to deviations in y.

Similarly, for the generalized M-estimators, the influence function takes the form

$$IF((X,y;T,F) = W(X)\psi((y - X^T T(F))V(X))M^{-1}X \tag{11.15}$$

where M is an appropriately redefined matrix as before.

11.2.2 Asymptotic variance

Maronna and Yohai (1981) derived consistency and asymptotic normality of regression M-estimators under appropriate conditions. The asymptotic covariance matrix is

$$\begin{aligned} V(T,F) &= \int IF(X,y;T,F)IF(X,y;T,F)^T dF(X,y) \\ &= M^{-1}(\int \psi^2(y - X^T T(F))XX^T dF(X,y))M^{-1} \\ &= M^{-1}(\int \psi^2(u)dG(u))(\int XX^T dK(X))M^{-1} \\ &= \frac{\int \psi^2(u)dG(u)}{(\int \psi'(u)dG(u))^2}(\int XX^T dK(X))^{-1} \end{aligned} \tag{11.16}$$

Therefore, minimizing $V(T,F)$ amounts to minimizing the first term over ψ.

11.3 Optimality

We define the gross-error sensitivity as

$$\gamma^*(T,F) = \sup_{(X,y)} \|IF(X,y;T,F)\|_2, \tag{11.17}$$

Then from the influence function of generalized M-estimators, if ψ is bounded and $\|W(X)X\|_2$ is bounded, we have $\gamma^* < \infty$.

The change-of-variance sensitivity is defined as

$$\kappa^*(T,F) = \sup_{(X,y)} \{\frac{trace(CVF(X,y;T,F))}{trace(V(T,F))}\}. \tag{11.18}$$

Theorem 1. *We have $\gamma^* \geq p\sqrt{\frac{\pi}{2}}\frac{1}{\mathbb{E}\|X\|_2}$. If $\mathbb{E}(\frac{XX^T}{\|X\|_2})$ is a scalar matrix, then $\eta(x,r) = sign(r)/\|x\|_2$ reaches the lower bound, so we call this estimator most B-robust.* [1]

Proof. See page 318 of Hampel's textbook. □

Now we consider the optimal B-robustness. Assuming radical symmetry of K, optimal B-robust estimator is the Hampel-Krasker estimator, which has ψ is the Huber function, $V(X) = \|AX\|_2$ for some appropriate A and $W(X) = \frac{1}{\|AX\|_2}$.

11.4 Exercise

1. Show that when K is radially symmetric (i.e., a function of $\|X\|_2$), the lower bound is achieved by the median estimator which solves

$$\sum_{i=1}^{n} \frac{sgn(y_i - X_i^T\theta)}{\|X_i\|_2} X_i = 0. \tag{11.19}$$

[1] The scalar matrix is a real multiple of the identity matrix.

2. Is this estimator equivalent to

$$\min_{\theta} \sum_{i=1}^{n} \frac{|y_i - X_i^T \theta|}{\|X_i\|_2}? \tag{11.20}$$

ECE901/CS838/STAT992: Robustness Theory Fall 2018

Lecture 12: One-step estimator

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, October 18*

12.1 One-step estimator

Assume we have the linear model $y = X\theta + u$. Then the M-estimator is defined as

$$\min_{\theta} \sum_{i=1}^{n} \rho(y_i - X_i^T\theta). \tag{12.1}$$

One-step estimator starts with initial point $\hat{\theta}_0$ such that

$$\|\hat{\theta}_0 - \theta\|_2 = \mathcal{O}_p(\frac{1}{\sqrt{n}}), \tag{12.2}$$

then take one-step by

$$\hat{\theta}_1 = \hat{\theta}_0 + (X^TX)^{-1}X^T\hat{R}, \tag{12.3}$$

where

$$\hat{R} = \frac{n}{\sum_{i=1}^{n}\psi'(y_i - X_i^T\hat{\theta}_0)}(\psi(y_1 - X_1^T\hat{\theta}_0), \psi(y_2 - X_2^T\hat{\theta}_0), ..., \psi(y_n - X_n^T\hat{\theta}_0))^T. \tag{12.4}$$

12.2 Newton's method

Recall that Taylor expansion of f around $\hat{\theta}_0$ is

$$f(\theta) \approx f(\hat{\theta}_0) + \nabla f(\hat{\theta}_0)^T(\theta - \hat{\theta}_0) + \frac{1}{2}(\theta - \hat{\theta}_0)^T\nabla^2 f(\hat{\theta}_0)(\theta - \hat{\theta}_0). \tag{12.5}$$

Then minimizing RHS over Θ, we have

$$0 = \nabla f(\hat{\theta}_0) + \nabla^2 f(\hat{\theta}_0)(\theta - \hat{\theta}_0). \tag{12.6}$$

Therefore, one step of Newton's method for minimizing $f(\theta)$ is

$$\hat{\theta}_1 = \hat{\theta}_0 - (\nabla^2 f(\hat{\theta}_0))^{-1}\nabla f(\hat{\theta}_0). \tag{12.7}$$

In out context, $f(\theta) = \sum_{i=1}^{n}\rho(y_i - X_i^T\theta)$. Then

$$\nabla f(\theta) = \sum_{i=1}^{n} -\psi(y_i - X_i^T\theta)X_i = (-\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i - X_i^T\theta))X^T\hat{R}, \tag{12.8}$$

and

$$\nabla^2 f(\theta) = \sum_{i=1}^{n}\psi'(y_i - X_i^T\theta)X_iX_i^T \approx n\mathbb{E}[\psi'(u_i)] \times \mathbb{E}[X_iX_i^T] \approx n(\frac{1}{n}\sum_{i=1}^{n}\psi'(y_i - X_i^T\hat{\theta}_0))\frac{X^TX}{n}. \tag{12.9}$$

12.3 Main results

12.3.1 Asymptotical normality

$\hat{\theta}_1$ behaves asymptotically like global minimizer of $\sum_{i=1}^n \rho(y_i - X_i^T\theta)$.

Theorem 1 ([Bickel, 1975]). *Suppose $\|\hat{\theta}_0 - \theta\|_2 = \mathcal{O}_p(\frac{1}{\sqrt{n}})$. Also suppose $\psi = \rho'$ is twice differentiable with $\|\psi''\|_\infty \leq C_2$ and $\mathbb{E}[\psi(u_i)] = 0, \mathbb{E}[\psi^2(u_i)] < \infty, \mathbb{E}[(\psi'(u_i))^2] < \infty, \mathbb{E}[u_i^2\psi'(u_i)^2] < \infty$. Also suppose $\frac{1}{n}\sum_{i=1}^n X_i X_i^T \to \mathbb{E}(X_i X_i^T)$ in probability where $\mathbb{E}(X_i X_i^T)$ is a positive definite matrix, and $\max_{1\leq i\leq n}\|X_i\|_\infty = o_p(\sqrt{n})$. Then*

$$\sqrt{n}(\hat{\theta}_1 - \theta) \to \mathcal{N}(0, \frac{\int \psi^2(u)dG(u)}{(\int \psi'(u)dG(u))^2}[\mathbb{E}(X_i X_i^T)]^{-1}) \tag{12.10}$$

in distribution.

Remark 1. Conditions on ψ, X can be relaxed.

Proof. We denote $\hat{H} = (\frac{XX^T}{n})(\frac{1}{n}\sum_{i=1}^n \psi'(\hat{u}_i))$ and $\hat{u}_i = y_i - X_i^T\hat{\theta}_0$. Then we have

$$\hat{\theta}_1 = \hat{\theta}_0 + \hat{H}^{-1}(\frac{1}{n}\sum_{i=1}^n \psi(\hat{u}_i)X_i). \tag{12.11}$$

Therefore,

$$\sqrt{n}(\hat{\theta}_1 - \theta) = \sqrt{n}(\hat{\theta}_0 - \theta) + \hat{H}^{-1}(\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(\hat{u}_i)X_i). \tag{12.12}$$

By Taylor expansion, we can write

$$\begin{aligned}\sum_{i=1}^n \psi(\hat{u}_i)X_i &= \sum_{i=1}^n (\psi(u_i) - \psi'(u_i)X_i^T(\hat{\theta}_0 - \theta) + \frac{\psi''(t_i)}{2}(X_i^T(\hat{\theta}_0 - \theta))^2)X_i \\ &= \sum_{i=1}^n (\psi(u_i) - \psi'(u_i)X_i^T(\hat{\theta}_0 - \theta))X_i + o_p(\sqrt{n})\end{aligned} \tag{12.13}$$

since

$$\|\sum_{i=1}^n \frac{\psi''(t_i)}{2}(X_i^T(\hat{\theta}_0 - \theta))^2)X_i\|_\infty \leq \frac{C_2}{2}\max_{1\leq i\leq n}\|X_i\|_\infty \sum_{i=1}^n (X_i^T(\hat{\theta}_0 - \theta))^2 = o_p(\sqrt{n})nO_p(1)(O_p(\frac{1}{\sqrt{n}}))^2 = o_p(\sqrt{n}). \tag{12.14}$$

Therefore we have

$$\sqrt{n}(\hat{\theta}_1 - \theta) = \hat{H}^{-1}(\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(u_i)X_i) + \sqrt{n}\hat{H}^{-1}(\hat{H} - \frac{1}{n}\sum_{i=1}^n \psi'(u_i)X_i X_i^T)(\hat{\theta}_0 - \theta) + \hat{H}^{-1}o_p(1) \tag{12.15}$$

Note by CLT, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(u_i)X_i \to^d \mathcal{N}(0, \mathbb{E}[\psi^2(u_i)]\mathbb{E}(X_i X_i^T)), \tag{12.16}$$

and by LLN, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi'(u_i)X_iX_i^T \to^p \mathbb{E}[\psi'(u_i)]\mathbb{E}[X_iX_i^T]. \tag{12.17}$$

Furthermore, by lemma below, we have

$$\hat{H} \to^p \mathbb{E}[\psi'(u_i)]\mathbb{E}[X_iX_i^T]. \tag{12.18}$$

Hence using Slutsky's theorem, the result should follow.

□

Lemma 1.

$$\hat{H} \to^p \mathbb{E}[\psi'(u_i)]\mathbb{E}[X_iX_i^T]. \tag{12.19}$$

Proof. We write $\frac{1}{n}\sum_{i=1}^{n}\psi'(\hat{u}_i) = \frac{1}{n}\sum_{i=1}^{n}[\psi'(u_i) + \psi''(t_i)(\hat{u}_i - u_i)]$ for some t_i between u_i and $\hat{u}_i$ by mean value theorem. Therefore,

$$\begin{aligned} &|\frac{1}{n}\sum_{i=1}^{n}\psi'(\hat{u}_i) - \frac{1}{n}\sum_{i=1}^{n}\psi'(u_i)| \le |\frac{1}{n}\sum_{i=1}^{n}\psi''(t_i)X_i^T(\hat{\theta}_0 - \theta)| \\ &\le C_2\|\frac{1}{n}\sum_{i=1}^{n}X_i\|_\infty\|\hat{\theta}_0 - \theta\|_1 = C_2 o_p(\sqrt{n})O_p(\frac{1}{\sqrt{n}}) = o_p(1). \end{aligned} \tag{12.20}$$

Furthermore, by LLN, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi'(u_i) \to^p \mathbb{E}(\psi'(u_i)) \tag{12.21}$$

Hence we prove the lemma. □

Remark 2.

1. For initial estimator $\hat{\theta}_0$, we can use OLS.

$$\hat{\theta}_0 = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\theta + u) = \theta + (X^TX)^{-1}X^Tu \tag{12.22}$$

   By CLT, we have $\sqrt{n}(\hat{\theta}_0 - \theta) = O_p(1)$.

2. Practically, we might want to iterate Newton's method multiple steps.

12.3.2 Influence function

Theorem 2. *Influence function of one-step estimator is the same as the influence function of the global minimizer.*

Proof. See exercise. □

Definition 1. The one-step functional $T_1(F)$ is defined as

$$T_1(F) = T_0(F) + (\int \psi'(y_i - X_i^TT_0(F))X_iX_i^TdF(X_i, y_i))^{-1}(\int \psi(y_i - X_i^TT_0(F))X_idF(X_i, y_i)) \tag{12.23}$$

12.4 Exercise

Show that

$$IF(X_0, y_0; T_1, F) = [(\int \psi'(u)dG(u))(\int XX^T dK(X))]^{-1}\psi(y_0 - X_0^T T_1(F))X_0 \tag{12.24}$$

assuming density of G is symmetric, ψ is odd and $T_0(F_\theta) = \theta$.

Proof. Use definition of $T_1(F)$. □

Bibliography

[Bickel, 1975] Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434.

ECE901/CS838/STAT992: Robustness Theory Fall 2018

Lecture 13: Linear regression with unknown scale

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, October 23*

13.1 Joint estimation

Assume the linear regression $y_i = X_i^T\theta^* + u_i$ where $X_i \sim K$, $u_i \sim G_{\sigma^*}$ and σ^* is unknown. Then the MLE is

$$\max_{\theta,\sigma} \prod_{i=1}^n k(X_i)\frac{1}{\sigma}g(\frac{y_i - X_i^T\theta}{\sigma}). \tag{13.1}$$

This is equivalent to

$$\min_{\theta,\sigma} \sum_{i=1}^n \{\rho(\frac{y_i - X_i^T\theta}{\sigma}) + \log\sigma\} \tag{13.2}$$

for $\rho = -\log g$. Note if $\rho(\frac{y_i - X_i^T\theta}{\sigma}) = f_1(\sigma)f_2(y_i - X_i^T\theta)$, we can ignore σ when estimating θ^*. Otherwise, we need to care about σ.

Taking derivative of (13.2) with respect to θ and σ, we have the estimating equation:

$$\sum_{i=1}^n \rho'(\frac{y_i - X_i^T\theta}{\sigma})X_i = 0. \tag{13.3}$$

$$\sum_{i=1}^n [\frac{-(y_i - X_i^T\theta)}{\sigma^2}\rho'(\frac{y_i - X_i^T\theta}{\sigma}) + \frac{1}{\sigma}] = 0 \tag{13.4}$$

We can write these equations as

$$\sum_{i=1}^n \psi(\frac{y_i - X_i^T\theta}{\sigma})X_i = 0. \tag{13.5}$$

$$\sum_{i=1}^n \chi(\frac{y_i - X_i^T\theta}{\sigma}) = 0 \tag{13.6}$$

where $\psi = \rho', \chi(t) = t\rho'(t) - 1$.

However, the above system of equations might be complicated to solve and loss function (13.2) is not jointly convex in (θ, σ).

13.2 Huber's method

Instead, Huber proposed to optimize

$$\min_{\theta,\sigma} \sum_{i=1}^n (\rho(\frac{y_i - X_i^T\theta}{\sigma}) + a)\sigma \tag{13.7}$$

where a is a constant chosen later.

Lemma 1. *This objective function is jointly convex in* (θ, σ)*, if* ρ *is convex.*

Proof. For simplicity, denote $f(\theta,\sigma) = \sum_{i=1}^{n}(\rho(\frac{y_i - X_i^T\theta}{\sigma}) + a)\sigma$, then we have

$$\nabla_\theta f = -\sum_{i=1}^{n} \rho'(\frac{y_i - X_i^T\theta}{\sigma})X_i \tag{13.8}$$

$$\nabla_\sigma f = \sum_{i=1}^{n}[\rho'(\frac{y_i - X_i^T\theta}{\sigma})[-\frac{y_i - X_i^T\theta}{\sigma}] + \rho(\frac{y_i - X_i^T\theta}{\sigma}) + a] \tag{13.9}$$

$$\nabla^2_{\theta\theta} f = \sum_{i=1}^{n} \rho''(\frac{y_i - X_i^T\theta}{\sigma})\frac{X_i X_i^T}{\sigma} \tag{13.10}$$

$$\nabla^2_{\theta\sigma} f = \sum_{i=1}^{n} \rho''(\frac{y_i - X_i^T\theta}{\sigma})(\frac{(y_i - X_i^T\theta)X_i}{\sigma^2}) \tag{13.11}$$

$$\nabla^2_{\sigma\sigma} f = \sum_{i=1}^{n} \rho''(\frac{y_i - X_i^T\theta}{\sigma})[\frac{(y_i - X_i^T\theta)^2}{\sigma^3}] \tag{13.12}$$

Therefore,

$$\nabla^2 f = \sum_{i=1}^{n} u_i u_i^T \geq 0 \tag{13.13}$$

where $u_i^T = \sqrt{\rho''(\frac{y_i - X_i^T\theta}{\sigma})}(\frac{X_i^T}{\sqrt{\sigma}}, \frac{y_i - X_i^T\theta}{\sigma^{3/2}})$ (here we use the convexity of ρ). Hence, $f(\theta,\sigma)$ is jointly convex.

□

Remark 1. One drawback of this method is if ρ is not convex, the loss function is no longer jointly convex.

Lemma 2. *The objective function is consistent for* (θ^*, σ^*) *for a proper choice of* $a \in \mathbb{R}$*.*

Proof. Consider the estimating equation (13.8), we have

$$\mathbb{E}[\rho'(\frac{y_i - X_i^T\theta^*}{\sigma^*})X_i] = \mathbb{E}[\rho'(\frac{u_i}{\sigma^*})X_i] = 0. \tag{13.14}$$

Similarly, consider (13.9), we have

$$\mathbb{E}[\rho'(\frac{u_i}{\sigma^*})(\frac{-u_i}{\sigma^*}) + \rho(\frac{u_i}{\sigma^*}) + a] = 0 \tag{13.15}$$

where a can be chosen to solve the equation and note it doesn't depend on σ^*. □

13.3 Breakdown point of above Huber's joint estimator

Lemma 3. *Consider (13.7), the breakdown point is 0 if ρ is convex and ρ' is bounded.*

Proof. We argue by contradiction. If the breakdown point is not 0, then by arbitrarily changing one point, the estimating equation

$$\sum_{i=1}^{n} \psi(\frac{y_i - X_i^T\theta}{\sigma})X_i = 0 \tag{13.16}$$

still has a solution (θ, σ) which satisfy $\|\theta\|_2 \leq B_\theta, |\sigma| \leq B_\sigma$. Since

$$\psi(\frac{y_1 - X_1^T\theta}{\sigma}) + \sum_{i=2}^{n} \psi(\frac{y_i - X_i^T\theta}{\sigma})X_i = 0, \tag{13.17}$$

We proceed by changing (X_1, y_1) such that $\|X_1\|_2 \to \infty$ and $\frac{y_1}{\|X_1\|_2} \to \infty$, then $\sum_{i=2}^{n} \psi(\frac{y_i - X_i^T\theta}{\sigma})X_i$ will always remain bounded, and

$$\frac{y_1 - X_i^T\theta}{\sigma} \geq \frac{y_1 - B_\theta\|X_1\|_2}{B_\sigma} = \frac{\|X_1\|_2}{B_\sigma}(\frac{y_1}{\|X_1\|_2} - B_\theta) \to \infty. \tag{13.18}$$

Therefore,

$$\psi(\frac{y_1 - X_1^T\theta}{\sigma}) \to \sup_t \psi(t) > 0 \tag{13.19}$$

since ψ is increasing. Thus, we have

$$\psi(\frac{y_1 - X_1^T\theta}{\sigma})X_1 \to \infty. \tag{13.20}$$

This is a contradiction to the estimating equation. □

13.4 Other methods for regression estimation with unknown σ

- MM-estimation (Yohai, 1987):
  1. Compute an initial estimate $\hat{\theta}_0$ that is consistent (for example, OLS).
  2. Compute a robust scale estimate $\hat{\sigma}$ based on residuals $\{y_i - X_i^T\theta_0\}_{i=1}^n$ (for example, M-estimate of scale).
  3. Minimize M-estimation $\sum_{i=1}^{n} \rho(\frac{y_i - X_i^T\theta}{\hat{\sigma}})$.
- Least trimmed squares (LTS) (Rousseeuw, 1984)

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n-\lfloor n\alpha \rfloor} (r(\theta)_{(i)})^2 \tag{13.21}$$

  where $r(\theta)_{(i)}$ are order statistics of residuals $y_i - X_i^T\theta$.

13.5 Exercise

Consider the joint estimator

$$\min_{\theta,\sigma} \sum_{i=1}^{n} (\rho(\frac{y_i - X_i^T\theta}{\sigma}) + a)\sigma \tag{13.22}$$

1. Suppose ρ is the Huber loss with parameter k, show that the optimizer $(\hat{\theta}, \hat{\sigma})$ to (13.22) are the same minimizing for

$$\min_{\theta,\sigma,r} \{\frac{1}{2\sigma}\|y - X\theta - r\|_2^2 + a\sigma + k\|r\|_1\} \tag{13.23}$$

2. Show that for any fixed σ, minimizing (13.22) yields the same solution as jointly minimizing (13.23) with respect to (θ, r). Therefore, Huber M-estimation is equivalent to penalized least squares regression.

Proof of Part 1. Note (13.22) is convex, so $(\hat{\theta}, \hat{\sigma})$ is the minimizer of (13.22) if and only if

$$\sum_{i=1}^{n} \psi(\frac{y_i - X_i^T\hat{\theta}}{\hat{\sigma}}) X_i = X^T \begin{pmatrix} \psi(\frac{y_1 - X_1^T\hat{\theta}}{\hat{\sigma}}) \\ \cdot \\ \cdot \\ \cdot \\ \psi(\frac{y_n - X_n^T\hat{\theta}}{\hat{\sigma}}) \end{pmatrix} = 0 \tag{13.24}$$

and

$$\sum_{i=1}^{n} \chi(\frac{y_i - X_i^T\hat{\theta}}{\hat{\sigma}}) = a \tag{13.25}$$

where

$$\chi(t) = t\psi(t) - \rho(t) = \begin{cases} \frac{t^2}{2} & |t| \le k \\ \frac{k^2}{2} & |t| > k \end{cases} \tag{13.26}$$

On the other hand, note (13.23) is also convex. Therefore, $(\tilde{\theta}, \tilde{\sigma}, \tilde{r})$ is the minimizer of (13.23) if and only if

$$X^T(\frac{y - X\tilde{\theta} - \tilde{r}}{\tilde{\sigma}}) = 0 \tag{13.27}$$

$$\frac{1}{2}\|\frac{y - \tilde{r} - X\tilde{\theta}}{\tilde{\sigma}}\|^2 = a \tag{13.28}$$

and

$$0 \in -(y - X\tilde{\theta} - \tilde{r}) + k\tilde{\sigma}\partial\|\tilde{r}\|_1 \tag{13.29}$$

From (13.29), we have $\tilde{r}$ is the soft thresholding. That is,

$$\tilde{r}_i = \begin{cases} 0 & |y_i - X_i^T\tilde{\theta}| \le k\tilde{\sigma} \\ y_i - X_i^T\tilde{\theta} - k\tilde{\sigma}\text{sign}(y_i - X_i^T\tilde{\theta}) & |y_i - X_i^T\tilde{\theta}| > k\tilde{\sigma} \end{cases} \tag{13.30}$$

Plug in $\tilde{r}$ in (13.27) and (13.28), then (13.27) is the same with (13.24) and (13.28) is the same with (13.25). Hence we prove the part 1.

□

Proof of Part 2. For fixed σ, (13.22) reduces to

$$\min_{\theta} \sum_{i=1}^{n} \rho(\frac{y_i - X_i^T\theta}{\sigma}) \tag{13.31}$$

and (13.23) reduces to

$$\min_{\theta,r} \frac{1}{2\sigma}\|y - X\theta - r\|_2^2 + k\|r\|_1. \tag{13.32}$$

Similarly with part 1, $\hat{\theta}$ is the minimizer of (13.31) if and only if

$$\sum_{i=1}^{n} \psi(\frac{y_i - X_i^T\hat{\theta}}{\sigma})X_i = 0 \tag{13.33}$$

and $(\tilde{\theta}, \tilde{r})$ is the minimizer of (13.32) if and only if

$$X^T(\frac{y - X\tilde{\theta} - \tilde{r}}{\sigma}) = 0 \tag{13.34}$$

and

$$\tilde{r}_i = \begin{cases} 0 & |y_i - X_i^T\tilde{\theta}| \le k\sigma \\ y_i - X_i^T\tilde{\theta} - k\sigma \text{sign}(y_i - X_i^T\tilde{\theta}) & |y_i - X_i^T\tilde{\theta}| > k\sigma \end{cases} \tag{13.35}$$

Therefore we prove the part 2. □

13.6 Appendix: optimality conditions for convex constrained optimization

Theorem 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function, and let $C \subset \mathbb{R}^n$ be a nonempty closed convex set. Consider the problem*

$$\begin{aligned} &\underset{x}{\text{minimize}} && f(x) \\ &\text{subject to} && x \in C \end{aligned}$$

A vector x^* is optimal for this problem if and only if $x^* \in C$ and

$$\nabla f(x^*)^T(z - x^*) \ge 0 \tag{13.36}$$

for all $z \in C$.

ECE901/CS838/STAT992: Robustness Theory Fall 2018

Lecture 14: Robust hypothesis testing

Lecturer: Po-Ling Loh *Scribe: Duzhe Wang, October 25*

We are interested in if the data is contaminated, then what will happen to the hypothesis test based on the test statistic T_n.

14.1 One-sample hypothesis tests

We are interested in testing

$$H_o : \theta = \theta_0 \tag{14.1}$$

and

$$\theta > \theta_0 \text{ (or } \theta < \theta_0, \theta \neq \theta_0). \tag{14.2}$$

based on a test statistic $T(X_1, ..., X_n)$. In presence of iid data where $X_i \sim F_\theta$, we might have the test statistic $T_n(X_1, ..., X_n) \to T(F_\theta)$ in probability. However, in general, we won't have $T(F_\theta) = \theta$. For example, for $F_\sigma : N(0, \sigma^2)$, if we want to test $H_o : \sigma = 1$ and $H_A : \sigma \neq 1$ using $T_n = \frac{1}{n}\sum_{i=1}^n X_i^2$, then $T_n \to \sigma^2$ rather than σ. But many of our derivations depend on the consistency property of the functional. We fix this by defining a new functional U.

We assume the map $\xi : \Theta \to \mathbb{R}$ such that $\xi(\theta) = T(F_\theta)$ is strictly monotone with a nonvanishing derivative, and define $U(F) = \xi^{-1}(T(F))$, then we have $U(F_\theta) = \theta$. Therefore we define the test influence function of T at F as

$$IF_{test}(x; T, F) = IF(x; U, F). \tag{14.3}$$

Note that if $F_t = (1-t)F_\theta + t\Delta_x$, then

$$\begin{aligned} IF_{test}(x; T, F_\theta) = IF(x; U, F_\theta) = \frac{d}{dt}U(F_t)|_{t=0} &= \frac{d}{dt}(\xi^{-1}(T(F_t))) = (\xi^{-1})'(T(F_t))\frac{d}{dt}(T(F_t))|_{t=0} \\ &= \frac{1}{\xi'(\xi^{-1}(T(F_\theta)))}IF(x; T, F_\theta) = \frac{1}{\xi'(\theta)}IF(x; T, F_\theta). \end{aligned} \tag{14.4}$$

Therefore IF_{test} is proportional to the standard IF of T.

Furthermore, if we replace test statistic $T(F)$ by another statistic $\tilde{T}(F) = \eta(T(F))$ where η is monotonic, then we have $\tilde{U}(F) = U(F)$ since

$$\tilde{\xi}(\theta) = \tilde{T}(F_\theta) = \eta(T(F_\theta)) = \eta(\xi(\theta)), \tag{14.5}$$

so $\tilde{\xi} = \eta \circ \xi$ and $\tilde{\xi}^{-1} = \xi^{-1} \circ \eta^{-1}$. Thus

$$\tilde{U}(F) = \tilde{\xi}^{-1}(\tilde{T}(F)) = \xi^{-1} \circ \eta^{-1}(\eta(T(F))) = U(F). \tag{14.6}$$

14.2 Two-sample hypothesis tests

We have two samples where $X_1, ..., X_m \sim G$ and $Y_1, ..., Y_n \sim F$ such that $G(x) = F(x - \theta)$. We are interested in testing

$$H_0 : \theta = \theta_0 \tag{14.7}$$

based on the test statistic $T_{m,n}(X_1, ..., X_m; Y_1, ..., Y_n)$. We assume $T_{m,n}(X_1, ..., X_m; Y_1, ..., Y_n) \to T(G, F; \theta)$ in probability. Similarly with the one-sample tests, we assume the map $\xi : \Theta \to \mathbb{R}$ such that $\xi(\theta) = T(G, F; \theta)$ is strictly monotone with a nonvanishing derivative, and define

$$U(G, F) = \xi^{-1}(T(G, F)). \tag{14.8}$$

Then we define the following influence functions

$$IF_{test,1}(x; T, G, F) = \lim_{t \to 0} \frac{U(G_{t,x}, F) - U(G, F)}{t} \tag{14.9}$$

$$IF_{test,2}(y; T, G, F) = \lim_{t \to 0} \frac{U(G, F_{t,y}) - U(G, F)}{t} \tag{14.10}$$

$$IF_{test,3}(x, y; T, G, F) = \lim_{t \to 0} \frac{U(G_{t,x}, F_{t,y}) - U(G, F)}{t} \tag{14.11}$$

where $G_{t,x} = (1 - t)G + t\Delta_x$ and $F_{t,y} = (1 - t)F + t\Delta_y$.

14.3 Level and power of the test

We hope the test has robustness of validity, i.e., stability of level of the test under contamination and robustness of efficiency, i.e., stability of power of the test under contamination.

For a fixed θ_0, define the alternative hypothesis parameter to be $\theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}$ where $\Delta > 0$ is a constant. Also define $U_n = \xi_n^{-1}(T_n)$.

Definition 1. The asymptotic level of the hypothesis test is defined by

$$\alpha(U, F) = \lim_{n \to \infty} \mathbb{P}_{\theta_0}(U_n \geq k_n(\alpha)) \tag{14.12}$$

where $k_n(\alpha)$ is the critical threshold.

Definition 2. The asymptotic power of a test is defined by

$$\beta(U, F) = \lim_{n \to \infty} \mathbb{P}_{\theta_n}(U_n \geq k_n(\alpha)) \tag{14.13}$$

where $k_n(\alpha)$ is the critical threshold.

We now introduce perturbations to the distribution F. Define

$$F^P_{n,t,x} = (1 - t_n)F_{\theta_n} + t_n\Delta_x \tag{14.14}$$

and

$$F^L_{n,t,x} = (1 - t_n)F_{\theta_0} + t_n\Delta_x \tag{14.15}$$

where $t_n = \frac{t}{\sqrt{n}}$. Therefore we define the level influence function by

$$LIF(x; U, F) = \lim_{n\to\infty} \frac{d}{dt} L_{n,t,x}|_{t=0} \tag{14.16}$$

where $L_{n,t,x} = F^L_{n,t,x}(U_n \geq k_n(\alpha))$. we also define the power influence function by

$$PIF(x; U, F) = \lim_{n\to\infty} \frac{d}{dt} P_{n,t,x}|_{t=0} \tag{14.17}$$

where $P_{n,t,x} = F^P_{n,t,x}(U_n \geq k_n(\alpha))$.

Theorem 1.

$$LIF(x; U, F) = \sqrt{E(T, F)}\psi(\lambda_{1-\alpha}) IF_{test}(x; T, F) \tag{14.18}$$

$$PIF(x; U, F, \Delta) = \sqrt{E(T, F)}\psi(\lambda_{1-\alpha} - \Delta\sqrt{E(T, F)}) IF_{test}(x; T, F) \tag{14.19}$$

where ψ is the density of standard normal. $\lambda_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal. $E(T, F) = (\int IF^2(y; T, F_{\theta_0}) dF_{\theta_0}(y))^{-1}$ is the asymptotic efficiency of the test.

Proof. See notes. □

As we can see from the theorem, we can study $IF_{test}(x; T, F)$ to control the behavior of PIF and LIF. Robustness of validity corresponds to upper bounding LIF and robustness of efficiency corresponds to lower bounding PIF. Optimal statistics give rise to censored likelihood ratio tests, truncated test statistics and etc.

14.4 Exercise

1. Compute the test influence function of the Z-test in the location model $N(\theta, 1)$ at $\theta = 0$.
2. Compute the IF of the χ^2-test in the scale model $N(0, \sigma^2)$ at $\sigma = 1$.
3. Which test is more easily affected by outliers?

Proof. 1. Let $T_n = \frac{1}{n}\sum_{i=1}^n X_i$, then the functional $T = \mathbb{E}(X_i)$. By definition of test influence function, we have

$$IF_{test}(x; T, F_\theta) = \lim_{t\to 0} \frac{T((1-t)F_\theta + t\Delta_x) - T(F_\theta)}{t} = x - \theta. \tag{14.20}$$

when $\theta = 0$, we have the test influence function is x.

2. Let $T_n = \frac{1}{n}\sum_{i=1}^n X_i^2$, then $T_n \to T = \mathbb{E}(X_i^2) = \sigma^2$ when $X_i \sim N(0, \sigma^2)$. Therefore, we have $\xi(\sigma) = \sigma^2$. Then

$$IF_{test}(x; T, F_\sigma) = \frac{1}{2\sigma} IF(x; T, F_\sigma) = \frac{x^2 - \sigma^2}{2\sigma} \tag{14.21}$$

so when $\sigma = 1$, the test influence function is $\frac{x^2-1}{2}$.

3. when $x \in (1-\sqrt{2}, 1+\sqrt{2})$, the IF for χ^2-test is smaller. When $x < 1-\sqrt{2}$ or $x > 1+\sqrt{2}$, the IF for Z-test is smaller.

□