Short course on robust statistics
September 2025
Frejus, France

1 Overview

Topics:

- Huber's theory: Minimax bias, minimax variance
- Hampel's theory: Influence functions
- Extensions: Linear regression, hypothesis testing
- Modern perspectives: Adversarial contamination, heavy-tailed data

Books:

- Huber & Ronchetti, "Robust Statistics"
- Hampel, Ronchetti, Rousseeuw & Stahel, "Robust Statistics: The Approach Based on Influence Functions"

2 Intro to robustness

- Deals with deviations from ideal models and their dangers for corresponding inference procedures
- Goal is to develop procedures that are still reliable and reasonably efficient under small deviations from the model (e.g., an ϵ-neighborhood of the assumed model)

Outlier rejection?

- Might consider a two-step procedure which first "cleans" data, then applies classical estimation procedure
- However, outliers may be difficult to recognize without an initial (somewhat) robust estimator

- Multiple outliers may "mask" each other so that none are rejected
- False rejections/false retentions may cause cleaned data to deviate from normal assumptions, too

Robustness desiderata

- **Efficiency:** Should have nearly(?) optimal efficiency under uncontaminated distribution
- **Stability:** Small deviations from uncontaminated distribution should only alter performance slightly
- **Breakdown:** Larger deviations from model should not be catastrophic

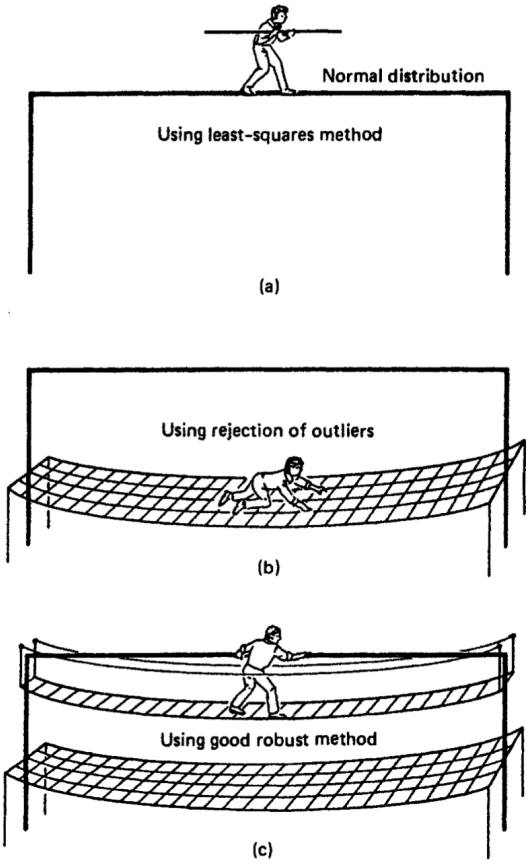

Figure 2. Various ways of analyzing data.

Hampel et al., *Robust Statistics*

3 Huber's perspective

Reference: Huber, "Robust estimation of a location parameter," 1964

We will be interested in estimating the *location parameter* of a distribution in one dimension. Assume $x_1, \dots, x_n \overset{i.i.d.}{\sim} F(t-\xi) = F_\xi(t)$, where $F(t)$ is a cdf defined over a probability space.

Goal is to estimate ξ. If probability distribution corresponding to F is symmetric around 0 ($F(-x) + F(x) = 1$ in the continuous case), then $\mathbb{E}_{F_0}[x_i] = 0$ and $\mathbb{E}_{F_\xi}[x_i] = \xi$, so we could use the mean $\frac{1}{n}\sum_{i=1}^n x_i$.

However, what if the model is contaminated?

Definition. *Consider the class of distributions with cdfs in the set*

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

where $\mathcal{M}$ is the set of all possible cdfs. This is known as (Huber's) ϵ-contamination model.

Note that for $F \in \mathcal{P}_\epsilon(F_0)$, we have

$$\begin{aligned}\sup_t |F(t) - F_0(t)| &= |(1-\epsilon)F_0(t) + \epsilon H(t) - F_0(t)| \\ &= \epsilon \cdot \sup_t |H(t) - F_0(t)| \le \epsilon,\end{aligned}$$

so F also lies in the ϵ-neighborhood of F_0 with respect to the Kolmogorov distance.

If $\mathbb{E}_{F_0}[x_i] = 0$, we have

$$\mathbb{E}_F[x_i] = (1-\epsilon)0 + \epsilon \mathbb{E}_H[x_i].$$

Hence, if the mean of H is not zero, the sample mean will not be consistent (and could be arbitrarily biased).

What about using the median? Nice property of medians is that changing a single point cannot perturb the estimator too much.

3.1 Breakdown point

Definition. *Consider a data set $X = \{x_1, \dots, x_n\}$ and an estimator $T_n(X)$. For $m \le n$, let*

$$b(m; X, T_n) = \sup_{X' \in \mathcal{X}_m} |T_n(X') - T_n(X)|,$$

where $\mathcal{X}_m \subseteq \mathbb{R}^n$ is the set of all data sets differing from X by at most m points. Then

$$\epsilon^*(X, T_n) := \frac{1}{n} \cdot \max_{m \geq 0} \{m : b(m; X, T_n) < \infty\}$$

is the breakdown point *of T_n at X.*

Example. *The breakdown point of the mean is 0. The breakdown point of the median is $\frac{1}{n} \cdot \lfloor \frac{n-1}{2} \rfloor$.*

In fact, the median achieves the highest possible breakdown point:

Exercise. *Show that the median achieves the highest possible breakdown point among all translation-invariant estimators.*

Definition. *An estimator is* translation-invariant *if*

$$T_n(x_1 + a, \ldots, x_n + a) = T_n(x_1, \ldots, x_n) + a,$$

for all $\{x_1, \ldots, x_n\}$ and $a \in \mathbb{R}$.

One can also define asymptotic notions of the breakdown point, for which the maximum breakdown point becomes 50%.

But is this seems like a very rough notion. Also, breakdown point has nothing to do with a distribution.

3.2 Bias

Going back to the ϵ-contamination model, suppose F_0 is symmetric, so (at the population level) mean and median are equal. (Define median as $\inf\left\{m : F_0(m) \geq \frac{1}{2}\right\}$.) Suppose $F_0 = \Phi$ (standard normal) for concreteness, and let $\mathcal{P}_\epsilon := \mathcal{P}_\epsilon(\Phi)$. What is a bound on the (asymptotic) bias of the median estimator?

Clearly, worst case is when H concentrates all mass on one side of origin. Median of $F \in \mathcal{P}_\epsilon$ is then the solution to

$$(1 - \epsilon)\Phi(b) = \frac{1}{2},$$

so maximum bias is $b_0 = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$.

Could we do better? Suppose $\{T_n\}$ is a sequence of estimators for a parameter $T(F_0)$. Define the asymptotic bias of a family of estimators $T = \{T_n\}$:

$$b(T, F) = b(\{T_n\}, F) = \left| \lim_{n\to\infty} \mathbb{E}_F(T_n) - T(F_0) \right|.$$

Then study the minimax problem:

$$\min_{\{T_n\}\subseteq\mathcal{T}} \max_{F\in\mathcal{P}_\epsilon} b(\{T_n\}, F),$$

where we restrict T_n to be in the class $\mathcal{T}$ of translation-invariant estimators.

Under appropriate distributional assumptions, we can show that when $\{T_n\}$ corresponds to the sample median, we have $T_n \xrightarrow{P} T(F)$ and $\lim_{n\to\infty} \mathbb{E}_F(T_n) = T(F)$ when $x_i \sim F$. It then follows easily that b_0 is an upper bound to the optimality problem (achieved using the median estimator).

To prove a lower bound, consider the distribution $F_+ \in P_\epsilon$ constructed as follows (shifted and centered around b_0):

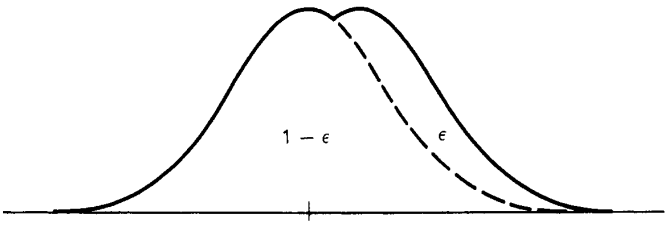

Exhibit 4.1 The distribution F_+ least favorable with respect to bias.

Also consider the version $F_- \in P_\epsilon$ centered around $-b_0$. Claim is that for any $\{T_n\} \subseteq \mathcal{T}$, we have

$$\max\{b(\{T_n\}, F_-), b(\{T_n\}, F_+)\} \geq b_0.$$

Suppose not. Then $b(\{T_n\}, F_-) < b_0$, implying

$$-b_0 < \lim_{n\to\infty} \mathbb{E}_{F_-}(T_n) < b_0.$$

But $\mathbb{E}_{F_+}(T_n) = \mathbb{E}_{F_-}(T_n) + 2b_0$, so

$$b_0 < \lim_{n\to\infty} \mathbb{E}_{F_+}(T_n) < 3b_0,$$

implying $b(\{T_n\}, F_+) > b_0$, a contradiction. We can see that this proof only requires symmetry and unimodality of F_0 for the construction to succeed. Conclusion is that

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

and the sample median is minimax optimal from the point of view of bias.

3.3 Variance

Why do we use the sample mean as a location estimator anyway? Has to do with efficiency—minimum asymptotic variance, assuming a normal distribution.

Theorem 1. *Suppose x_i's have density $f(x;\xi)$. Under appropriate regularity conditions, the maximum likelihood estimator*

$$\widehat{\xi}_{MLE} \in \arg\min_{\xi} \sum_{i=1}^{n} -\log f(x_i;\xi)$$

is asymptotically normal:

$$\sqrt{n}(\widehat{\xi} - \xi) \xrightarrow{d} N\left(0, \frac{1}{I(\xi)}\right),$$

where

$$I(\xi) = \mathbb{E}_\xi\left[\left(\frac{\partial \log f(x_i;\xi)}{\partial \xi}\right)^2\right]$$

is the Fisher information. Furthermore, the ratio $\frac{1}{I(\xi)}$ is the minimum possible variance among all asymptotically unbiased estimators of ξ.

However, the situation may become more complicated when samples are from an ϵ-ball around some distribution. Another minimax problem: Suppose

$$\sqrt{n}(T_n - T(F)) \xrightarrow{d} N(0, A(T,F)),$$

so $A(T,F)$ is the asymptotic variance of the rescaled, recentered sequence of estimators. Consider the minimax problem

$$\min_{\{T_n\}} \max_{F\in\mathcal{P}_\epsilon} A(\{T_n\}, F). \tag{1}$$

Motivated by nice results in MLE theory, we will restrict our attention to a particular class of estimators known as M-estimators.

Definition. *Consider a (symmetric) function ρ. A minimizer $T_n = T_n(x_1, \ldots, x_n)$ of $\sum_{i=1}^n \rho(x_i - T_n)$ is an M-estimator with associated loss function ρ.*

3.4 Asymptotic theory

We will show that under fairly general conditions, the M-estimator T_n consistently estimates the population-level parameter and is also asymptotically normal. Then we can derive formulas for the asymptotic variance and attempt to solve the minimax problem (1). (Note that for the mean estimator,

$$\max_{F \in \mathcal{P}_\epsilon} A(T, F) = \infty,$$

since we could just consider a mixing distribution H corresponding to $N(0, \sigma^2)$, for arbitrarily large σ.)

Throughout, suppose $\psi = \rho'$ is a nondecreasing function. For asymptotic normality, we will assume several conditions:

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

$$\sqrt{n}(T_n - t_0) \xrightarrow{d} N\left(0, \frac{\sigma^2(t_0)}{(\lambda'(t_0))^2}\right).$$

The proof uses the Lindeberg CLT.

3.5 Symmetric losses

Corollary 1. *Suppose ρ is a symmetric, convex function and the x_i's have a symmetric distribution. Suppose the derivative*

$$\lambda'(t) = \frac{\partial \mathbb{E}_F[\psi(x_i - t)]}{\partial t} = -\mathbb{E}_F[\psi'(x_i - t)] \tag{2}$$

exists and $\sigma^2(t) = \mathbb{E}_F[\psi^2(x_i - t)]$ is continuous in a neighborhood around 0. Also suppose $\mathbb{E}_F[\psi^2(x_i)] < \infty$ and $\mathbb{E}[\psi'(x_i)] > 0$. Then

$$T_n \in \arg\min_\xi \left\{ \sum_{i=1}^n \rho(x_i - \xi) \right\}$$

satisfies

$$\sqrt{n}T_n \xrightarrow{d} N\left(0, \frac{\mathbb{E}_F[\psi^2(x_i)]}{\mathbb{E}_F[\psi'(x_i)]^2}\right).$$

Note that in Corollary 1, the assumption that ρ is symmetric and F has a symmetric distribution implies that $\mathbb{E}_F[\psi(x_i)] = 0$, since ψ is an odd function. Thus, we can plug $t_0 = 0$ into Theorem 2.

In particular, we can apply the preceding results to derive asymptotic normality of the sample mean ($\psi(t) = t$) and sample median ($\psi(t) = \text{sign}(t)$). Due to non-differentiability, we have to use Theorem 2 in the case of the median.

3.6 Contaminated distributions

Now we will consider ϵ-neighborhoods. Suppose $\rho(t) = \frac{t^2}{2}$ and $F = (1-\epsilon)\Phi + \epsilon H$, where H is the cdf of a symmetric distribution satisfying conditions of Corollary 1 (from here on, assume we are restricting our attention to such nice distributions). Then

$$A(T, F) = \frac{\mathbb{E}_F[x_i^2]}{\mathbb{E}_F[1]^2} = (1-\epsilon) + \epsilon\mathbb{E}_H[x_i^2].$$

Clearly, this can be arbitrarily large.

On the other hand, suppose we have a function ψ such that $\psi' \geq 0$ and $\|\psi\|_\infty < k$ for some constant k. Then

$$\begin{aligned}\frac{\mathbb{E}_F[\psi^2(x_i)]}{\mathbb{E}_F[\psi'(x_i)]^2} &= \frac{(1-\epsilon)\mathbb{E}_\Phi[\psi^2(x_i)] + \epsilon\mathbb{E}_H[\psi^2(x_i)]}{\Big((1-\epsilon)\mathbb{E}_\Phi[\psi'(x_i)] + \epsilon\mathbb{E}_H[\psi'(x_i)]\Big)^2} \\ &\leq \frac{(1-\epsilon)\mathbb{E}_\Phi[\psi^2(x_i)] + \epsilon k^2}{(1-\epsilon)^2\mathbb{E}_\Phi[\psi'(x_i)]^2}, \end{aligned} \tag{3}$$

which is bounded as H ranges over different cdfs. One example of such a function is ψ corresponding to the Huber loss:

$$\rho(t) = \begin{cases} \frac{t^2}{2}, & \text{if } |t| \leq k, \\ k|t| - \frac{k^2}{2}, & \text{if } |t| > k. \end{cases}$$

Then $\psi(t) = \min\{k, \max(-k, t)\}$. For such a ψ, the upper bound in inequality (3) is achieved when H puts all mass outside the interval $[-k, k]$, since $\psi'(x) = 0$ and $\psi^2(x) = k$ at such points.

We could in theory try to minimize the upper bound (3) with respect to k. In fact, the optimal value of k will correspond to the solution to

$$\frac{2\varphi(k)}{k} - 2\Phi(-k) = \frac{\epsilon}{1-\epsilon}, \tag{4}$$

and the asymptotic variance bound is achieved when $(1-\epsilon)\Phi + \epsilon H$ has pdf

$$\frac{1-\epsilon}{\sqrt{2\pi}} \exp(-\rho(x)).$$

(The equation (4) is actually a little hard to show directly from taking derivatives of the right-hand expression in inequality (3), but fortunately we will derive the optimal value of k in another manner later on.)

3.7 Optimality of Huber loss

We now prove that the Huber estimator is actually minimax over *all* possible ψ. The following result gives a constructive method for determining a saddlepoint solution to the minimax problem.

In fact, a much more general result holds for the minimax variance problem:

Theorem 3. *Suppose G is the cdf of a symmetric distribution with twice continuously differentiable pdf g, such that $-\log g$ is convex (i.e., g is log-concave).*

(i) Then $V(\psi, F)$ has a saddlepoint: there exists $F_0 \in \mathcal{P}_\epsilon(G)$ and $\psi_0 \in \Psi$ such that

$$\max_{F \in \mathcal{P}_\epsilon(G)} V(\psi_0, F) = V(\psi_0, F_0) = \min_{\psi \in \Psi} V(\psi, F_0). \tag{5}$$

Hence,

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

and ψ_0 is minimax optimal.

(ii) Furthermore, we have the explicit expressions

$$\psi_0 = -\frac{f_0'}{f_0},$$

and

$$f_0(x) = \begin{cases} (1-\epsilon)g(x_0)e^{k(x-x_0)}, & \text{if } x \le x_0, \\ (1-\epsilon)g(x), & \text{if } x_0 < x < x_1, \\ (1-\epsilon)g(x_1)e^{-k(x-x_1)}, & \text{if } x \ge x_1, \end{cases} \tag{6}$$

where $x_0 < x_1$ are the endpoints of the interval where $\frac{|g'|}{g} \le k$ (either or both endpoints may be infinity), and k is related to ϵ by

$$\frac{1}{1-\epsilon} = \int_{x_0}^{x_1} g(x)dx + \frac{g(x_0)+g(x_1)}{k}. \tag{7}$$

The proof essentially proceeds by giving a constructive solution to the saddlepoint problem.

We have seen the special case when $g(x) = \varphi(x)$. Then

$$\frac{g'(x)}{g(x)} = \frac{\frac{1}{\sqrt{2\pi}} \cdot -x \exp\left(-\frac{x^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} = -x,$$

implying that the $x_0 = -k$ and $x_1 = k$. Then $f_0(x) = (1-\epsilon)\varphi(x)$ for $|x| \le k$. We can check that the form of f_0 agrees with the density provided earlier on $(-\infty, -k)$ and (k, ∞), as well:

$$\begin{aligned} f_0(x) &= (1-\epsilon)\varphi(k)\exp(-k(x-k)) = (1-\epsilon)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{k^2}{2}\right)\exp(-kx+k^2) \\ &= (1-\epsilon)\frac{1}{\sqrt{2\pi}}\exp\left(-kx+\frac{k^2}{2}\right) \end{aligned}$$

for $x > k$, and $\rho_0(x) = kx - \frac{k^2}{2}$ for $x > k$.

Furthermore, the equation relating k to ϵ reduces to

$$\frac{1}{1-\epsilon} = \int_{-k}^{k} \varphi(x)dx + \frac{2\varphi(k)}{k} = (1 - 2\Phi(k)) + \frac{2\varphi(k)}{k},$$

which is again equation (4), which identifies the optimal parameter for Huber's M-estimator.

Exercise. *What is the minimax optimal solution when G is the cdf of a $\mathcal{N}(0, \sigma^2)$ distribution?*

Solution: We need to substitute $g(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)$. Then

$$\frac{|g'(x)|}{g(x)} = \frac{|x|}{\sigma^2},$$

so $x_0 = -k\sigma^2$ and $x_1 = k\sigma^2$. Then

$$\begin{aligned}(1-\epsilon)g(x_0)\exp(k(x-x_0)) &= (1-\epsilon)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{k^2\sigma^2}{2}\right)\exp(k(x+k\sigma^2)) \\ &= (1-\epsilon)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(kx+\frac{k^2\sigma^2}{2}\right),\end{aligned}$$

so the form of the density is

$$f_0(x) = \begin{cases}(1-\epsilon)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(kx+\frac{k^2\sigma^2}{2}\right), & \text{if } x \le -k\sigma^2, \\ (1-\epsilon)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right), & \text{if } |x| < k\sigma^2, \\ (1-\epsilon)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-kx+\frac{k^2\sigma^2}{2}\right), & \text{if } x \ge k\sigma^2.\end{cases}$$

This means

$$\psi_0(x) = \frac{-f_0'(x)}{f_0(x)} = \begin{cases}-k, & \text{if } x \le -k\sigma^2, \\ \frac{x}{\sigma^2}, & \text{if } |x| < k\sigma^2, \\ k, & \text{if } x \ge k\sigma^2.\end{cases}$$

Furthermore, the value of k corresponds to the solution of the equation

$$\frac{1}{1-\epsilon} = \int_{-k\sigma^2}^{k\sigma^2} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)dx + \frac{2\cdot\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)}{k}.$$

We can easily see that when $\sigma = 1$, we obtain the same expressions for the Huber estimator derived earlier.

Remark. *How much further can we push this theory? Let us consider the minimax variance problem when* $\mathcal{P}_\epsilon^K(\Phi) = \{F : \sup_t |F(t)-\Phi(t)| < \epsilon\}$. *Recall that the* ϵ*-neighborhood we have been considering satisfies* $\mathcal{P}_\epsilon(\Phi) \subseteq \mathcal{P}_\epsilon^K(\Phi)$. *A rather sophisticated and ingenious construction due to Huber leads to a density of the form*

$$f_0(x) = f_0(-x) = \begin{cases}C_0\cos^2\left(\frac{\omega x}{2}\right), & \text{if } 0 \le x < x_0, \\ \varphi(x), & \text{if } x_0 \le x \le x_1, \\ C_1\exp(-\lambda(x-x_1)), & \text{if } x > x_1,\end{cases}$$

with corresponding ψ function

$$\psi(x) = \begin{cases} \omega \tan\left(\frac{\omega x}{2}\right), & \text{if } 0 \le x < x_0, \\ x, & \text{if } x_0 \le x \le x_1, \\ \lambda, & \text{if } x > x_1. \end{cases}$$

(The shape of ρ is that it takes the form $\log(\cos^2(x)) = x^2 + \frac{x^4}{6} + O(x^5)$ in an interval around 0, then becomes quadratic, then becomes linear.)

4 Hampel's perspective

What have we learned so far? Huber loss is minimax optimal in an ϵ-neighborhood of Φ. However, derivation relied heavily on nice form of normal density and symmetric contamination assumption. How can we "robustify" other estimation procedures?

We now discuss a second camp of robustness theory, developed by Hampel (1968) in his PhD thesis: "Contributions to the theory of robust estimation." Basic concepts are qualitative robustness (continuity of limiting functional), influence function (effect of infinitesimal perturbations), and breakdown point (distance to nearest singularity/asymptote).

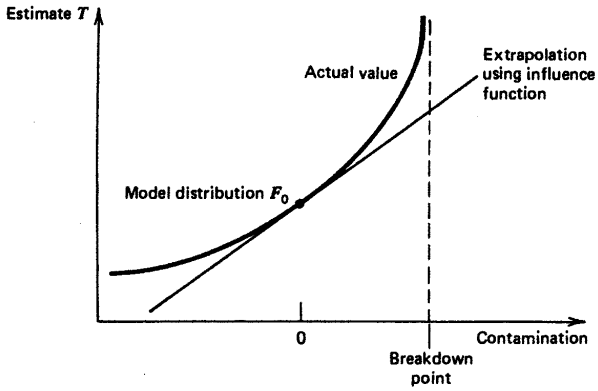

Figure 2. Extrapolation of a functional (estimator), using the infinitesimal approach. (Symbolic, using the analogue of an ordinary one-dimensional function.)

4.1 Influence functions

We will again assume we have a sequence of estimators $T_n(x_1, \ldots, x_n) \xrightarrow{P} T(F)$ when $x_i \sim F$. In order to be robust, we want some stability: T should

be continuous and the derivative should be well-behaved.

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

$$IF(x; T, F) := \lim_{t \to 0} \frac{T\left((1-t)F + t\Delta_x\right) - T(F)}{t}.$$

Influence function provides rate of change of $T(F)$ in direction of point mass Δ_x. In particular, we will be interested in bounding quantities such as *gross-error sensitivity*

$$\gamma^*(T, F) := \sup_x |IF(x; T, F)|$$

(analog of bounded derivative). Other quantities of interest include the *rejection point*

$$\rho^*(T, F) := \inf \{r > 0 : IF(x; T, F) = 0 \text{ when } |x| > r\}.$$

Furthermore, the influence function is related to the asymptotic variance of T_n: It can be shown under appropriate regularity conditions that when $x_i \overset{i.i.d.}{\sim} F$, we have

$$\sqrt{n}(T_n - T(F)) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(x_i; T, F) \xrightarrow{d} N(0, A(T, F)),$$

where $A(T, F) = \int IF(x; T, F)^2 dF(x)$.

Example.

- *(Mean) We have*

$$IF(x; T, F) = \lim_{t \to 0} \frac{((1-t)\mathbb{E}_F[x_i] + tx) - \mathbb{E}_F[x_i]}{t} = x - \mathbb{E}_F[x_i],$$

so when $\mathbb{E}_F[x_i] = 0$ *(e.g.,* F *corresponds to a symmetric distribution),* $IF(x; T, F) = x$*. Recall that under the same condition* $\mathbb{E}_F[x_i] = 0$ *(i.e.,* $t_0 = 0$ *in the earlier notation), we proved that*

$$\sqrt{n}(T_n - T(F)) \xrightarrow{d} N(0, \mathbb{E}_F[x_i^2]),$$

since $\psi(t) = t$ *and* $\frac{\mathbb{E}[\psi^2(x_i)]}{\mathbb{E}[\psi'(x_i)]^2} = \mathbb{E}[x_i^2]$.

- *(Median) We have*

$$IF(x;T,F) = \lim_{t\to 0} \frac{F_t^{-1}(1/2) - F^{-1}(1/2)}{t},$$

where $F_t := (1-t)F + t\Delta_x$*. It is easier to deal with an implicit equation. Note that*

$$F_t\left(F_t^{-1}\left(\frac{1}{2}\right)\right) = \frac{1}{2},$$

and the LHS can also be written as $(1-t)F\left(F_t^{-1}\left(\frac{1}{2}\right)\right) + t\Delta_x\left(F_t^{-1}\left(\frac{1}{2}\right)\right)$*. Differentiating with respect to* t*, we have*

$$-F\left(F_t^{-1}\left(\frac{1}{2}\right)\right) + (1-t)F'\left(F_t^{-1}\left(\frac{1}{2}\right)\right)\frac{d}{dt}F_t^{-1}\left(\frac{1}{2}\right) + \Delta_x\left(F_t^{-1}\left(\frac{1}{2}\right)\right) + t\Delta_x'\left(F_t^{-1}\left(\frac{1}{2}\right)\right)\frac{d}{dt}F_t^{-1}\left(\frac{1}{2}\right) = 0.$$

Evaluating at $t = 0$ *(and using the fact that* $\Delta_x'(y) = 0$ *for* $x \neq y$*), we have*

$$-\frac{1}{2} + F'\left(F^{-1}\left(\frac{1}{2}\right)\right) IF(x;T,F) + \Delta_x\left(F^{-1}\left(\frac{1}{2}\right)\right) = 0,$$

so

$$IF(x;T,F) = \frac{1/2 - \Delta_x(F^{-1}(1/2))}{F'(F^{-1}(1/2))} = \frac{1/2 - 1\{F^{-1}(1/2) > x\}}{F'(F^{-1}(1/2))}.$$

It is not hard to see that this is equal to $\frac{\text{sign}\{x - F^{-1}(1/2)\}}{2F'(F^{-1}(1/2))}$*. Recall that in our earlier results on asymptotic normality, we had*

$$\sqrt{n}(T_n - T) \xrightarrow{d} N\left(0, \frac{1}{4(F'(0))^2}\right)$$

for symmetric distributions (and in fact, we have the asymptotic variance $\frac{1}{4(F'(F^{-1}(1/2))^2}$ *for general distributions when we take* $t_0 = F^{-1}(1/2)$*). Thus, we again have the formula* $A(T,F) = \int IF(x;T,F)^2 dF(x)$ *in the case of the median. Note that the influence function looks a lot like* $\psi\ldots$

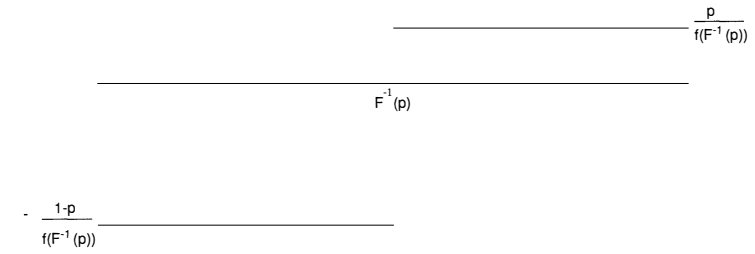

Figure 20.1. Influence function of the pth quantile.

(Note that the derivation above is not exactly rigorous, but these sorts of back-of-the-envelope derivations are often correct, and can be made rigorous under enough regularity assumptions.)

- *(General M-estimators) Recall that $T(F)$ is defined by*

$$\mathbb{E}_F[\psi(x_i - T(F))] = 0.$$

Using an implicit approach, we can show that

$$IF(x; T, F) = \frac{\psi(x - T(F))}{\mathbb{E}_F[\psi'(x_i - T(F))]}.$$

In particular, recall the formula for the asymptotic variance of "nice" M-estimators:

$$A(T, F) = \frac{\mathbb{E}_F[\psi(x_i)^2]}{(\mathbb{E}_F[\psi'(x_i)])^2},$$

when F is the cdf of a symmetric random variable, which is exactly $\int IF(x, T, F)^2 dF(x)$. Furthermore, these computations show that in order to obtain a bounded-influence estimator, we can use an M-estimator with ψ bounded.

4.2 Optimal B-robust estimators

The formula stated above for the influence function of an M-estimator is actually somewhat more general. Consider a family of distributions parametrized by θ, and suppose the functional $T(F_\theta)$ is defined implicitly by

$$\int \psi(y, T(F_\theta)) dF_\theta(y) = 0$$

(the special case of M-estimators is a family of distributions with location parameter θ, and $\psi(y,\theta) = \psi(y-\theta)$—we will also consider scale families later). One can show that

$$IF(x;T,F_\theta) = \frac{\psi(x,T(F_\theta))}{\int \psi(y,\theta)s(y,\theta)dF_\theta(y)},$$

where

$$s(y,\theta) := \frac{\partial}{\partial\theta}(\log f_\theta(y)) = \frac{\frac{\partial}{\partial\theta}f_\theta(y)}{f_\theta(y)}$$

is the score function.

4.2.1 Main result

The following result concerns optimality of estimators. We are interested in the minimal asymptotic variance $\int IF(x;T,F)^2 dF(x)$, subject to an upper bound on the gross error sensitivity $\gamma^*(T,F) = \sup_x |IF(x;T,F)|$.

Theorem 4. *Suppose $F = F_\theta$ (for a fixed θ) and*

$$I(F) = \int s(x,\theta)^2 dF(x) > 0$$

(this is the Fisher information). Let $b > 0$ be a constant. Then there exists a real number a such that

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

(truncated function, which becomes constant outside of $[-b,b]$) satisfies $\int \widetilde{\psi}(y)dF(y) = 0$ and $d := \int \widetilde{\psi}(y)s(y,\theta)dF(y) > 0$. Furthermore, $\widetilde{\psi}$ minimizes

$$\int IF(y;T,F)^2 dF(y)$$

among all mappings ψ satisfying

(i) $\int \psi(y)dF(y) = 0$,

(ii) $\int \psi(y)s(y,\theta)dF(y) \neq 0$,

(iii) and

$$\gamma^*(T,F) \leq c := \frac{b}{d}.$$

Any other solution to this optimization problem coincides with a nonzero multiple of $\widetilde{\psi}$, almost everywhere with respect to F.

Remark. *The condition $\int \psi(y)dF(y) = 0$ is known as "Fisher consistency": for location M-estimators, we have $\psi_\theta(y) = \psi(y - \theta)$, so this is the familiar condition $\mathbb{E}_{F_\theta}[\psi(x_i - \theta)] = 0$.*

We call estimators that minimize the asymptotic variance subject to a bound on gross error sensitivity *optimal B-robust estimators*. (The B stands for "bias," whereas there are also V-robust estimators.) We call any estimator such that $\gamma^*(T, F) < \infty$ *B-robust*.

4.2.2 Location M-estimators

Consider the family of M-estimators for location. Then

$$s(y, \theta) = \frac{\frac{\partial}{\partial\theta} f_\theta(y)}{f_\theta(y)} = \frac{\frac{\partial}{\partial\theta} f(y - \theta)}{f(y - \theta)} = \frac{-f'(y - \theta)}{f(y - \theta)}.$$

By Theorem 4, the optimal B-robust estimator at $\theta = 0$ is given by

$$\widetilde{\psi}(y) = \left[\frac{-f'(y)}{f(y)} - a\right]_{-b}^{b}.$$

Hence, the optimal (finite-sample) B-robust estimator is the solution to

$$\sum_{i=1}^{n} \left[\frac{-f'(x_i - \theta)}{f(x_i - \theta)} - a\right]_{-b}^{b} = 0$$

(finite-sample version).

In particular, if the density f is symmetric around 0, then $s(y, 0) = \frac{-f'(y)}{f(y)}$ is an odd function. It follows that $\int [s(y, 0)]_{-b}^{b} dF(y) = 0$, so $a = 0$, implying that $\widetilde{\psi}(y) = \left[\frac{-f'(y)}{f(y)}\right]_{-b}^{b}$. If $F = \Phi$, this reduces to the Huber estimator with parameter b: $\widetilde{\psi}(y) = [y]_{-b}^{b}$, which is the familiar Huber ψ function. Hence, the Huber estimator also emerges as the optimal M-estimator for the location of a normal family, this time with respect to B-robustness. (Different Huber parameters correspond to different bounds on the GES.)

4.2.3 Scale M-estimators

We can also consider a family of distributions parametrized by scale:

$$f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right).$$

(For instance, consider the $N(0, \theta^2)$ family, where θ is unknown. Then $f_\theta(x) = \frac{1}{\sqrt{2\pi}\theta} \exp\left(\frac{-x^2}{2\theta^2}\right)$.)

The M-estimator corresponds to solving $\int \psi(x, \theta) dF_\theta(x) = 0$, where ψ is a function such that $\psi(x, \theta) = \psi\left(\frac{x}{\theta}\right)$. (We can easily check that if f_θ takes the form above, the MLE will correspond to such an M-estimator for an appropriately defined choice of $\psi(u) = -\frac{uf'(u)}{f(u)} - 1$.)

What is an optimal B-robust estimator? We have

$$s(y, \theta) = \frac{\frac{\partial}{\partial\theta} f_\theta(y)}{f_\theta(y)} = \frac{\frac{1}{\theta} f'\left(\frac{y}{\theta}\right)\left(\frac{-y}{\theta^2}\right) - \frac{1}{\theta^2} f\left(\frac{y}{\theta}\right)}{\frac{1}{\theta} f\left(\frac{y}{\theta}\right)}.$$

If we substitute $\theta = 1$, we obtain $s(y, 1) = \frac{-yf'(y)}{f(y)} - 1$. Hence, the optimal B-robust estimator, according to the theorem, is

$$\widetilde{\psi}_1(y) = \left[\frac{-yf'(y)}{f(y)} - 1 - a\right]_{-b}^{b}.$$

When $F = \Phi$, this becomes

$$\widetilde{\psi}_1(y) = [y^2 - 1 - a]_{-b}^{b},$$

for an appropriate value of a, which generally depends on b.

The (finite-sample) optimal B-robust M-estimator then corresponds to

$$\sum_{i=1}^{n} \left[\left(\frac{x_i^2}{\theta^2}\right) - 1 - a\right]_{-b}^{b} = 0$$

(truncation of MLE expression, above or below, depending on the value of b).

5 Robust linear regression

Analysis of multidimensional estimators becomes a bit more complicated. However, the results from the univariate case translate conveniently (in some cases) into the context of linear regression.

Model:

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + u_i, \qquad \forall 1 \le i \le n,$$

or in matrix notation, $y = X\theta + u$. Assume $x_i \overset{i.i.d.}{\sim} K$ and $u_i \overset{i.i.d.}{\sim} G_\sigma$, where u_i's are independent of x_i's and σ is scale parameter of error distribution. Then joint distribution is

$$f_{\theta,\sigma}(x,y) = f(x)f(y|x) = k(x) \cdot \frac{1}{\sigma} g\left(\frac{y - x^T\theta}{\sigma}\right).$$

MLE would correspond to maximizing

$$\prod_{i=1}^{n} k(x_i) \cdot \frac{1}{\sigma} g\left(\frac{y_i - x_i^T\theta}{\sigma}\right),$$

or equivalently, maximizing

$$\sum_{i=1}^{n} \log\left\{\frac{1}{\sigma} g\left(\frac{y_i - x_i^T\theta}{\sigma}\right)\right\}. \tag{8}$$

In the case when G_σ is cdf of $N(0,\sigma^2)$, we have $g(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$, so MLE (for parameter θ) is

$$\max_\theta \sum_{i=1}^{n} \log\left\{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - x_i^T\theta)^2}{2\sigma^2}\right)\right\},$$

or

$$\min_\theta \sum_{i=1}^{n} (y_i - x_i^T\theta)^2$$

(ordinary least squares).

Not surprisingly, OLS is not robust to deviations from normality of the error distribution G. We can see this by computing an influence function

$IF(x, y; T, F)$ (see below). Inspired by the MLE formulation (8), consider an M-estimator with a general loss function ρ:

$$\min_{\theta} \sum_{i=1}^{n} \rho(y_i - x_i^T \theta) \tag{9}$$

(assume for now that σ is known, or is a nuisance parameter). Then we have the estimating equation

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

5.1 Influence functions

Suppose $F = F_\theta$. Functional $T(F)$ is a vector that solves

$$0 = \mathbb{E}_{(x_i, y_i) \sim F}[\psi(y_i - x_i^T T(F)) x_i] = \int \psi(y - x^T T(F)) x dF(x, y).$$

It can be shown that

$$\begin{aligned} IF(x_0, y_0; T, F) &= \left(\int \psi'(y - x^T T(F)) x x^T dF(x, y) \right)^{-1} \left(\psi(y_0 - x_0^T T(F)) x_0 \right) \\ &:= M^{-1} \psi(y_0 - x_0^T T(F)) x_0. \end{aligned}$$

By independence, we have

$$M = \int \psi'(u) dG(u) \cdot \left(\int x x^T dK(x) \right).$$

If we write $y_0 = x_0^T T(F) + r_0$, we can think of the influence function as being broken into two factors:

$$IF(x_0, y_0; T, F) = \frac{\psi(r_0)}{\mathbb{E}_{u \sim G}[\psi'(u)]} \cdot \left(\left(\mathbb{E}_{x \sim K}[x x^T] \right)^{-1} x_0 \right) := IR(r_0; T, G) \cdot IP(x_0; T, K),$$

where IR is the *influence of the residual* and IP is the *influence of the position*. In particular, if we can guarantee boundedness of IF in response direction if ψ is bounded. (This is not the case for OLS.)

5.2 Optimality

Can we derive analogs of Huber's and Hampel's theories of optimality for robust regression estimators? Yes. The first step is to derive asymptotic normality results for solutions to estimating equations, and then minimize variance term (note that this is now a matrix).

5.2.1 Asymptotic normality

Huber (1973) analyzed M-estimator (9), for convex ρ, in fixed design setting, and derived asymptotic normality under suitable regularity conditions, in the regime $\frac{p^3}{n} \to 0$. What should the limiting variance be? When $p \to \infty$, we have to be a bit careful, since the estimate θ is a vector of growing dimensionality; so Huber showed asymptotic normality of projections $a^T\widehat{\theta}$, where $a \in \mathbb{R}^p$ is a fixed vector of contrasts.

Maronna & Yohai (1981) analyzed the setting we have described, with fixed p and where $n \to \infty$, showing that asymptotic covariance matrix is

$$\begin{aligned} V(T,F) &= \int IF(x,y;T,F)(IF(x,y;T,F))^T dF(x,y) \\ &= M^{-1}\left(\int \psi^2(y - x^T T(F))xx^T dF(x,y)\right) M^{-1} \\ &= M^{-1}\left(\int \psi^2(u)dG(u)\right)\left(\int xx^T dK(x)\right) M^{-1} \\ &= \frac{\int \psi^2(u)dG(u)}{\left(\int \psi'(u)dG(u)\right)^2}\left(\int xx^T dK(x)\right)^{-1}. \end{aligned}$$

In particular, minimizing $V(T,F)$ over the class of ψ functions then reduces to the familiar univariate problem of choosing ψ to minimize $\frac{\mathbb{E}_G[\psi^2(u)]}{\mathbb{E}_G[\psi'(u)]^2}$—in the case when $G = \Phi$, we again recover the Huber M-estimator as being minimax optimal in terms of variance.

5.2.2 B-robustness

Hampel's theory is a bit more complicated, due to the fact that we have to extract real-valued measures from vectors and matrices. For instance, we can derive gross error sensitivity

$$\gamma^*(T,F) = \sup_{x,y} \|IF(x,y;T,F)\|_2.$$

Can show parallels of univariate location estimation theory for monotone ψ functions.

However, for the family of M-estimators we have studied,

$$\gamma^*(T, F_\theta) = \sup_{x,y} \left\{ |\psi(y - x^T\theta)| \cdot \|M^{-1}x\|_2 \right\} = \infty.$$

Hence, optimality theory focuses on slightly broader class of M-estimators defined by

$$\mathbb{E}_{(x_i,u_i)\sim F} \left[w(x_i) \cdot \psi \left((y_i - x_i^T T(F)) \cdot v(x_i) \right) x_i \right] = 0. \tag{10}$$

Using same IF calculation above, we can compute

$$IF(x_0, y_0, T, F) = w(x_0)\psi \left((y_0 - x_0^T T(F)) \cdot v(x_0) \right) M^{-1}x_0,$$

where M is an appropriately defined population-level matrix. In particular, if $w(x)x$ is a bounded function of x (e.g., $w(x) = \frac{1}{\|Ax\|_2}$) and ψ is bounded, we can guarantee that $\gamma^*(T, F_\theta) < \infty$.

In the radially symmetric case, the optimal B-robust estimator corresponds to the *Hampel-Krasker estimator*, given by equation (10) with $v(x) = \|Ax\|_2 = \frac{1}{w(x)}$, where ψ is equal to the Huber function.

5.3 Unknown scale

So far, we have ignored the question of estimating the scale parameter σ. Back to the MLE when $x_i \overset{i.i.d.}{\sim} K$ and $u_i \overset{i.i.d.}{\sim} G_\sigma$, we want to maximize

$$\prod_{i=1}^{n} \left\{ k(x_i) \cdot \frac{1}{\sigma} g\left(\frac{y_i - x_i^T\theta}{\sigma} \right) \right\},$$

or

$$\min_\theta \sum_{i=1}^{n} \left(\rho\left(\frac{y_i - x_i^T\theta}{\sigma} \right) + \log \sigma \right),$$

where $\rho = -\log g$. If ρ is quadratic, we can ignore σ. However, if ρ is not quadratic, e.g., Huber loss, fixing a value of σ and minimizing only over θ could lead to different robustness properties (i.e., potentially large loss in efficiency) if the value of σ is chosen poorly. This is a much harder problem, and the theory is far from complete. We will mention a few different methods that have been proposed in robust statistics literature:

(a) Joint estimation: We could try to jointly optimize the objective with respect to (θ, σ). However, even if ρ is convex, the objective function is generally nonconvex.

A clever idea (introduced by Huber) is to jointly optimize

$$\min_{\theta,\sigma} \sum_{i=1}^{n} \left(\rho\left(\frac{y_i - x_i^T \theta}{\sigma} \right) + a \right) \sigma, \tag{11}$$

where $a \in \mathbb{R}$ is an appropriately chosen constant to make the resulting estimators consistent. In particular, this function is jointly convex in (θ, σ) when ρ is convex.

The estimating equations corresponding to equation (11) take the form (**??**), with

$$\psi(t) = \rho'(t), \qquad \chi(t) = t\rho'(t) - \rho(t) - a.$$

Note that if ρ is an even function and distribution of u_i is symmetric, we have

$$\mathbb{E}\left[\rho'\left(\frac{y_i - x_i^T \theta}{\sigma} \right) x_i \right] = 0,$$

but we need to include the constant a to ensure that

$$\mathbb{E}\left[\frac{u_i}{\sigma} \rho'\left(\frac{u_i}{\sigma} \right) - \rho\left(\frac{u_i}{\sigma} \right) - a \right] = 0.$$

In particular, the value of a should be set according to the distribution with which we want our method to be robust. (However, this depends on exact knowledge of G.)

A drawback of this method is that using nonconvex ρ may lead to more desirable properties from the point of view of robustness (i.e., high breakdown point, finite rejection point).

(b) MM-estimators: Introduced by Yohai (1987). Procedure is as follows:

1. Compute initial consistent estimate $\widehat{\theta}_0$ (e.g., using OLS or LAD).
2. Compute robust scale estimate $\widehat{\sigma}$ based on $\{y_i - x_i^T \widehat{\theta}_0\}_{i=1}^n$ (e.g., using M-estimator of scale).
3. Minimize $\sum_{i=1}^n \rho\left(\frac{y_i - x_i^T \theta}{\widehat{\sigma}} \right)$ with respect to θ.

Much of theory focuses on obtaining estimators with high BP and bounded IF. Asymptotic theory depends on assumption that $\widehat{\sigma}$ is sufficiently close to true scale parameter.

(c) Least trimmed squares (LTS): Introduced by Rousseeuw (1984). Optimize

$$\sum_{i=1}^{\lfloor \alpha n \rfloor} (r(\theta))^2_{(i)},$$

where $r_i(\theta) = y_i - x_i^T\theta$. However, the objective function is highly nonconvex, and theoretical properties of optimum are largely unknown. Output can also be used to obtain initial scale estimate $\widehat{\sigma}$ for MM-estimation algorithm.

6 Hypothesis testing

6.1 Huber theory

Reference: Huber & Strassen (1973)

Suppose $P_0 \neq P_1$ are two distributions, and we have i.i.d. samples $x_i \overset{i.i.d.}{\sim} P$. Our goal is to test the hypotheses

$$H_0 : P \in \mathcal{P}_\epsilon(P_0) := \mathcal{P}_0 \quad \text{vs.} \quad H_1 : P \in \mathcal{P}_\epsilon(P_1) := \mathcal{P}_1.$$

6.1.1 Optimality theory

We wish to find the "maximin" test: For a fixed level α, maximize the minimum power subject to an upper bound on the level:

$$\sup_{\varphi} \inf_{P \in \mathcal{P}_1} \mathbb{E}_P[\varphi(X)] \quad \text{s.t.} \quad \sup_{P \in \mathcal{P}_0} \mathbb{E}_P[\varphi(X)] \leq \alpha,$$

where $\varphi(X) \in \{0, 1\}$ defines the test.

The main result is that for sufficiently small ϵ, the optimal test φ^* is actually a likelihood ratio test between two "least favorable" distributions $Q_0 \in \mathcal{P}_0$ and $Q_1 \in \mathcal{P}_1$. In particular, for $j \in \{0, 1\}$, we have

$$\mathbb{P}_{Q_j'}(\varphi^*(X) \neq j) \leq \mathbb{P}_{Q_j}(\varphi^*(X) \neq j) \quad \text{for all } Q_j' \in \mathcal{P}_j.$$

Note that it is a non-asymptotic result, and provides the form of a maximin test for *any* value of n (and any value of ϵ):

Theorem 5. *Suppose P_0 and P_1 have densities p_0 and p_1, respectively. Let*

$$q_0(x) = \begin{cases} (1-\epsilon)p_0(x) & \text{if } \frac{p_1(x)}{p_0(x)} < b, \\ \frac{1-\epsilon}{b} p_1(x) & \text{if } \frac{p_1(x)}{p_0(x)} \geq b, \end{cases}$$

$$q_1(x) = \begin{cases} (1-\epsilon)p_1(x) & \text{if } \frac{p_1(x)}{p_0(x)} > a, \\ a(1-\epsilon)p_0(x) & \text{if } \frac{p_1(x)}{p_0(x)} \leq a. \end{cases}$$

For any $\epsilon \in (0,1)$, there exist unique $a < b$ such that q_0 and q_1 are valid pdfs. Furthermore, the respective distributions are in $\mathcal{P}_0$ and $\mathcal{P}_1$. If ϵ is sufficiently small, a maximin test is given by

$$\varphi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \frac{q_1(x_i)}{q_0(x_i)} \geq c, \\ 0 & \text{if } \prod_{i=1}^n \frac{q_1(x_i)}{q_0(x_i)} < c, \end{cases}$$

for an appropriate value of c.

Note that if we define $r(x) = \frac{p_1(x)}{p_0(x)}$, then

$$\frac{q_1(x)}{q_0(x)} = \begin{cases} a, & \text{if } r(x) \leq a, \\ r(x), & \text{if } a < r(x) < b, \\ b, & \text{if } r(x) \geq b. \end{cases}$$

Hence, the probability ratio test between P_0 and P_1 is replaced with a censored version.

Example. *Suppose $P_0 = N(\theta_0, 1)$ and $P_1 = N(\theta_1, 1)$, with $\theta_0 < \theta_1$. The usual LRT would consist of rejecting H_0 when $\sum_{i=1}^n x_i > c$. Instead, the censored LRT rejects H_0 when $\sum_{i=1}^n [x_i]_a^b > c'$.,*

6.1.2 Connections to differential privacy

Definition. *A randomized algorithm T taking inputs in $\mathcal{X}^*$ and returning outputs in a space with events set $\mathcal{S}$ satisfies ϵ-DP if, for all neighboring data sets $x, x' \in \mathcal{X}^n$, and all events $S \in \mathcal{S}$, we have*

$$\mathbb{P}(T(x) \in S) \leq e^{\epsilon} \mathbb{P}(T(x') \in S).$$

Private hypothesis testing: Suppose $x_1, \ldots, x_n$ are drawn i.i.d. from either P or Q. What is the minimum number of samples needed for an ϵ-DP test to reliably distinguish P from Q, and what are optimal private tests?

Canonne et al., "The structure of optimal private tests for simple hypothesis testing," STOC 2019, proposed a "clamped log-likelihood ratio test" and showed it is optimal up to constant factors:

$$T(X) = \sum_{i=1}^{n} \left[\log \frac{P(x_i)}{Q(x_i)} \right]_a^b + \text{Lap}\left(\frac{1}{\epsilon(b-a)} \right),$$

for an appropriate choice of (a, b), where we reject $H_0 : x_i \sim P$ if $T(X) < 0$.

This result was derived entirely independently of the robust statistics literature, although those who work in this area may not be surprised that a robust estimator can be used to construct a private estimator, e.g.:

- Dwork & Lei, STOC 2009
- Georgiev & Hopkins, NeurIPS 2022
- Asi, Ullman & Zakynthinou, ICML 2023

6.2 Hampel theory

Suppose we are interested in performing a parametric hypothesis test of the form

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta > \theta_0 \quad \text{(or two-sided version)}, \end{aligned}$$

based on a test statistic $T_n(x_1, \ldots, x_n)$. We will suppose that

$$T_n(x_1, \ldots, x_n) \xrightarrow{\mathbb{P}} T(F),$$

when $x_i \overset{i.i.d.}{\sim} F$.

6.2.1 Influence function of test

Recall that our discussion of Hampel's optimality theory used the fact that our functionals were Fisher consistent: $T(F_\theta) = \theta$. However, test statistics may not be Fisher consistent (e.g., test of variance for the $N(0, \sigma^2)$ family is

a χ^2-test based on sample variance, but scale parameter is σ). Accordingly, we define a map $\xi : \Theta \to \mathbb{R}$ such that $\xi(\theta) = T(F_\theta)$, and define the functional $U(F) = \xi^{-1}(T(F))$, so that

$$U(F_\theta) = \xi^{-1}(T(F_\theta)) = \xi^{-1}(\xi(\theta)) = \theta$$

is a Fisher consistent functional. Also assume that ξ is strictly monotone with nonvanishing derivative, so ξ^{-1} is well-defined.

Definition. *The* test influence function *of T at F is defined by*

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

Note that

$$IF_{\text{test}}(x; T, F_\theta) = \frac{d}{dt} U(F_t)\Big|_{t=0} = \frac{d}{dt} \xi^{-1}(T(F_t))\Big|_{t=0},$$

where $F_t = (1-t)F_\theta + t\Delta_x$, so using the chain rule, we have

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

Also note that if we replace the statistic $T(F)$ by any other statistic $\tilde{T}(F) = \eta(T(F))$, where η is monotonic transformation, the functional $U(F)$ will remain the same, since $\tilde{\xi}(\theta) = \tilde{T}(F_\theta) = \eta(T(F_\theta)) = \eta(\xi(\theta))$, so $\tilde{\xi} = \eta \circ \xi$, and

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

6.2.2 Level and power

We are interested in both:

- *Robustness of validity:* Stability of level of test under small deviations from null hypothesis.
- *Robustness of efficiency:* Stability of power of test under small deviations from alternative hypothesis.

We will study the influence of distributional deviations on the asymptotic level and power of tests. Accordingly, let $\theta_n = \theta_0 + \frac{\Delta}{\sqrt{n}}$, where $\Delta > 0$ is a constant. The asymptotic level of the test is

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

where $k_n(\alpha)$ is the critical threshold. (Thus, $\alpha(U, F) = \alpha$.) Here, $U_n := \xi^{-1}(T_n)$. Similarly, the asymptotic power is defined as

$$\beta(U, F) = \lim_{n\to\infty} \mathbb{P}_{\theta_n}(U_n \geq k_n(\alpha)). \tag{12}$$

We now introduce perturbations. Define

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

where $t_n = \frac{t}{\sqrt{n}}$. (We require the fraction of contamination to be converging to 0, since otherwise the difference between F_{θ_n} and F_{θ_0} would become negligible, and the two cdfs would converge together.)

We will study the *level influence function*

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

where $L_{n,t,x} = F^L_{n,t,x}(U_n \geq k_n(\alpha))$, and the *power influence function*

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

where $P_{n,t,x} = F^P_{n,t,x}(U_n \geq k_n(\alpha))$. It turns out that these influence functions are both multiples of $IF_{\text{test}}(x; T, F)$:

Theorem 6.

$$LIF(x; U, F) = \sqrt{E(T, F)}\varphi(\lambda_{1-\alpha})IF_{test}(x; T, F),$$
$$PIF(x; U, F, \Delta) = \sqrt{E(T, F)}\varphi\left(\lambda_{1-\alpha} - \Delta\sqrt{E(T, F)}\right) IF_{test}(x; T, F),$$

where $\lambda_{1-\alpha}$ *is the lower-*$(1-\alpha)$ *quantile of the standard normal distribution,* $\Phi^{-1}(1-\alpha)$*, and* $E(T, F) := \left(\int IF^2_{test}(y; T, F_{\theta_0})dF_{\theta_0}(y)\right)^{-1}$ *is the* asymptotic efficacy *of the test. (Note that* $\beta(U, F) = 1 - \Phi\left(\lambda_{1-\alpha} - \Delta\sqrt{E(T, F)}\right)$*, hence the terminology for asymptotic efficacy.)*

Thus, ensuring robustness of validity corresponds to bounding the LIF, whereas ensuring robustness of efficiency corresponds to bounding the PIF. Optimality theory concerns maximizing the asymptotic power of a test (12), subject to bounds on LIF and PIF. Can also be recast as minimizing asymptotic variance of T subject to an upper bound on the absolute value of the *self-standardized influence* $\sqrt{E(T,F)}IF_{\text{test}}(x;T,F)$.

Gives rise to tests based on truncated test statistics, censored likelihood ratio tests, etc.

7 Adversarial contamination

7.1 Setup

Huber's contamination model (which is also the framework for Hampel's theory on infinitesimal robustness) assumes that contaminated data come from an i.i.d. mixture $(1-\epsilon)F + \epsilon H$. However, what if we instead draw n i.i.d. data points $\{x_i\}_{i=1}^n$ from F, and then arbitrarily contaminate ϵn data points to obtain the final set $\{\widetilde{x}_i\}_{i=1}^n$ of observations?

We will work in the (nonasymptotic) probably approximately correct (PAC) framework. For a given $\delta > 0$, goal is to obtain an estimator $\widehat{\mu}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of distribution mean $\mu = \mathbb{E}_F[x_i]$ satisfying

$$\mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \leq t(n,\delta,\epsilon)\right) \geq 1-\delta,$$

where $t(n,\delta,\epsilon)$ is as small as possible. Clearly, all of the asymptotic theory and theory of robustness based on IFs is no longer applicable, since adversary is more powerful than Huber contamination model. Of course, the sample mean fails catastrophically in this framework: Since the adversary is omniscient and allowed to base the corrupted data on the other data points, if $\epsilon \geq \frac{1}{n}$, can always choose $\widetilde{x}_n$ such that $\|\widehat{\mu} - \mu\|_2$ is deterministically larger than any value.

Are medians any better? Yes!—and optimal. We establish the following lower bound on the error in 1 dimension:

Theorem 7. *Let $F_\mu = N(\mu, 1)$, and suppose $\delta < c$. Any location estimator $\widehat{\mu}$ must satisfy*

$$\sup_{\mu\in\mathbb{R}} \mathbb{P}_\mu \left(\sup_{\{\widetilde{x}_i\}} |\widehat{\mu}(\widetilde{x}_1, \ldots, \widetilde{x}_n) - \mu| > C\left(\epsilon + \sqrt{\frac{\log(1/\delta)}{n}}\right)\right) > \delta,$$

where the probability is taken with respect to $x_i \overset{i.i.d.}{\sim} F_\mu$ *and* $\{\widetilde{x}_i\}_{i=1}^n$ *are an (adversarial)* ϵ*-perturbation of* $\{\widetilde{x}_i\}_{i=1}^n$.

Proof. We consider two cases, where $\epsilon < \sqrt{\frac{\log(1/\delta)}{n}}$ and $\epsilon \geq \sqrt{\frac{\log(1/\delta)}{n}}$. In the first case, the bound is implied by the fact that the empirical mean satisfies

$$\mathbb{P}_\mu \left(|\widehat{\mu}(x_1, \ldots, x_n) - \mu| > \frac{1}{2}\sqrt{\frac{\log(1/\delta)}{n}} \right) > \delta$$

when δ is sufficiently small, which comes from a Gaussian tail bound and the fact that the empirical mean is the best possible estimator for clean data—the adversary can just do nothing. So suppose we are in the second case; we will show that

$$\sup_{\mu \in \mathbb{R}} \mathbb{P}_\mu \left(\sup_{\{\widetilde{x}_i\}} |\widehat{\mu}(\widetilde{x}_1, \ldots, \widetilde{x}_n) - \mu| > C\epsilon \right) \geq \frac{1}{2}.$$

In fact, we will show that a similar lower bound holds under Huber's ϵ-contamination model:

$$\sup_{\mu \in \mathbb{R}} \sup_{F \in \mathcal{P}_\epsilon(F_\mu)} \mathbb{P}_{x_i \overset{i.i.d.}{\sim} F} \left(|\widehat{\mu}(x_1, \ldots, x_n) - \mu| > C\epsilon \right) \geq \frac{1}{2}.$$

This implies the result for the adversary, since adversarial contamination is stronger than Huber contamination (and the existence of a stochastic contamination model which probabilistically chooses ϵn of the data points and incurs a certain error implies existence of an adversarial contamination model which incurs at least the same error, as well).

The key idea is that if $\mu_1 - \mu_2 = \epsilon$, then we can construct $G \in \mathcal{P}_\epsilon(F_{\mu_1}) \cap \mathcal{P}_\epsilon(F_{\mu_2})$. Assuming this construction, we can write

$$\begin{aligned}
&\sup_{\mu \in \mathbb{R}} \sup_{F \in \mathcal{P}_\epsilon(F_\mu)} \mathbb{P}_{x_i \overset{i.i.d.}{\sim} F} \left(|\widehat{\mu}(x_1, \ldots, x_n) - \mu| > \frac{\epsilon}{2} \right) \\
&\quad \geq \frac{1}{2}\mathbb{P}_{x_i \overset{i.i.d.}{\sim} G} \left(|\widehat{\mu}(x_1, \ldots, x_n) - \mu_1| > \frac{\epsilon}{2} \right) + \frac{1}{2}\mathbb{P}_{x_i \overset{i.i.d.}{\sim} G} \left(|\widehat{\mu}(x_1, \ldots, x_n) - \mu_2| > \frac{\epsilon}{2} \right) \\
&\quad \geq \frac{1}{2}\mathbb{P}_{x_i \overset{i.i.d.}{\sim} G} \left(\left\{ |\widehat{\mu} - \mu_1| \geq \frac{\epsilon}{2} \right\} \bigcup \left\{ |\widehat{\mu} - \mu_2| \geq \frac{\epsilon}{2} \right\} \right) \\
&\quad \geq \frac{1}{2}\mathbb{P}_{x_i \overset{i.i.d.}{\sim} G} \left(|\mu_1 - \mu_2| \geq \epsilon \right) = \frac{1}{2},
\end{aligned}$$

completing the proof. □

For an upper bound result, see the much more general result in Theorem 8 below.

7.2 Higher dimensions

What to do in $d \geq 1$ dimensions? This is a bit subtle, due to a multiplicity of higher-dimensional versions of the median, of which several give suboptimal rates.

The simplest idea is to use the coordinatewise median. However, this can be shown to incur an $O(\epsilon\sqrt{d})$ error. In contrast, we can derive a $O\left(\epsilon + \sqrt{\frac{d}{n}}\right)$ error using more complicated notions of medians.

Definition. *The* Tukey median *of a data set* $\{x_i\}_{i=1}^n$ *is defined as* $\widehat{\mu} = \arg\max_{\mu \in \mathbb{R}^d} \mathcal{D}(\mu, \{x_i\}_{i=1}^n)$*, where*

$$\mathcal{D}(\mu, \{x_i\}_{i=1}^n) := \inf_{\|u\|_2 = 1} \frac{1}{n} \sum_{i=1}^n 1\left\{u^T(x_i - \mu) \geq 0\right\}$$

is the Tukey depth function*.*

In other words, the Tukey depth at μ looks at all halfspaces cutting the recentered data set, and takes the one which cuts off the fewest points. The Tukey median then seeks to maximize this depth over all μ.

We have the following upper bound:

Theorem 8. *Suppose* $F = N(\mu, I_d)$*, the contamination level satisfies* $\epsilon < \frac{1}{8}$*, and the sample size is large enough so* $2C\sqrt{\frac{d+\log(1/\delta)}{n}} \leq \frac{1}{4}$*. The Tukey median satisfies*

$$t(n, \delta, \epsilon) \leq \Phi^{-1}\left(\frac{1}{2} + 2\epsilon + 2C\sqrt{\frac{d + \log(1/\delta)}{n}}\right).$$

(Here, $C > 0$ is a universal constant not necessarily agreeing with the one appearing in the earlier theorem.) This bound turns out to be minimax

optimal (up to constants) for small ϵ and large n when $F = N(\mu, I_d)$:

$$\begin{aligned}
&\Phi^{-1}\left(\frac{1}{2} + 2\epsilon + 2C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) \\
&\quad\approx \Phi^{-1}\left(\frac{1}{2}\right) + \frac{1}{\varphi(\Phi^{-1}(1/2))}\left(2\epsilon + 2C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) \\
&\quad= \frac{1}{\varphi(0)}\left(2\epsilon + 2C\sqrt{\frac{d+\log(1/\delta)}{n}}\right)
\end{aligned}$$

(compare with the lower bound rate in Theorem 7 in the 1D case).

Sketch of proof. We begin with a statement about the clean data, which follows from standard uniform concentration results from empirical process theory (e.g., using arguments based on VC dimension):

$$\mathbb{P}_{x_i \overset{i.i.d.}{\sim} N(\mu, I_d)}\left(\sup_{\|u\|_2=1, \eta\in\mathbb{R}^d}\left|\frac{1}{n}\sum_{i=1}^n 1\left\{u^T(x_i-\eta)\geq 0\right\} - \mathbb{P}_{x_i\sim N(\mu,I_d)}\left(u^T(x_i-\eta)\geq 0\right)\right| \geq C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) \leq \delta.$$

For the remainder of the argument, suppose we are in the "good" event with probability $1-\delta$, where all empirical depth measures are close to their population-level counterparts. Furthermore, note that

$$\mathbb{P}_{x_i\sim N(\mu,I_d)}\left(u^T(x_i-\eta)\geq 0\right) = \Phi\left(u^T(\mu-\eta)\right).$$

We now argue that the true mean has Tukey depth at least (approximately) $\frac{1}{2}-\epsilon$. Note that for any unit vector $u \in \mathbb{R}^d$, we have

$$\begin{aligned}
\sum_{i=1}^n 1\left\{u^T(\widetilde{x}_i-\mu)\geq 0\right\} &\geq \sum_{i=1}^n 1\left\{u^T(x_i-\mu)\geq 0\right\} - \epsilon n \\
&\geq n\left(\Phi(0) - C\sqrt{\frac{d+\log(1/\delta)}{n}} - \epsilon\right) \\
&= \left(\frac{1}{2} - \epsilon - C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) n,
\end{aligned}$$

where the second inequality follows from the fact that we are in the good event.

On the other hand, we can show that any point $\eta \in \mathbb{R}^d$ such that

$$\|\eta-\mu\|_2 > \Phi^{-1}\left(\frac{1}{2}+2\epsilon+2C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) = -\Phi^{-1}\left(\frac{1}{2}-2\epsilon-2C\sqrt{\frac{d+\log(1/\delta)}{n}}\right)$$

must have

$$\mathcal{D}(\eta, \{\widetilde{x}_i\}_{i=1}^n) < \frac{1}{2}-\epsilon-C\sqrt{\frac{d+\log(1/\delta)}{n}}.$$

Since the Tukey median maximizes the depth function, the result of the theorem will then hold. Indeed, suppose η is a far away point and let $v = \frac{\eta-\mu}{\|\eta-\mu\|_2}$. Then

$$\begin{aligned}
\sum_{i=1}^n 1\left\{v^T(\widetilde{x}_i-\eta)\geq 0\right\} &\leq \sum_{i=1}^n 1\left\{v^T(x_i-\eta)\geq 0\right\}+\epsilon n \\
&\leq n\left(\Phi(-\|\eta-\mu\|_2)+\epsilon+C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) \\
&< n\left(\left(\frac{1}{2}-2\epsilon-2C\sqrt{\frac{d+\log(1/\delta)}{n}}\right)+\epsilon+C\sqrt{\frac{d+\log(1/\delta)}{n}}\right) \\
&< \left(\frac{1}{2}-\epsilon-C\sqrt{\frac{d+\log(1/\delta)}{n}}\right)n.
\end{aligned}$$

This proves the desired result. □

However, computing the Tukey median is also difficult in high dimensions, with computational complexity $O(n^{d-1})$.

7.3 Spectral algorithms

We now discuss algorithmic alternatives which can overcome the $O(n^{d-1})$ computational complexity of the Tukey median and are also provably (nearly) optimal.

7.3.1 First algorithm

The first algorithm, due to Diakonikolas et al. Its success is roughly based on the intuition that outliers (measured in terms of their impact on the sample mean) can be detected based on the sample covariance. Here is the algorithm:

The algorithm proceeds by assigning weights $\{w_i\}_{i=1}^n$ to the data points. The final output is the weighted mean $\mu(w) = \sum_{i=1}^n w_i x_i$, for an iteratively determined vector of weights.

1. Define $w_i^{(0)} = \frac{1}{n}$ for all i.
2. For $t = 0, 1, 2, \ldots$, compute $\Sigma^{(t)} = \sum_{i=1}^n w_i^{(t)}(x_i - \mu(w^{(t)}))(x_i - \mu(w^{(t)}))^T$.
   (a) If $\|\Sigma^{(t)}\|_2 \leq C$, output $\widehat{\mu} = \mu(w^{(t)})$.
   (b) If $\|\Sigma^{(t)}\|_2 > C$, define scores
   $$\tau_i^{(t)} = \langle v^{(t)}, x_i - \mu(w^{(t)})\rangle^2$$
   where $v^{(t)}$ is the top eigenvector of $\Sigma^{(t)}$. For each i, define $w_i^{(t+1)} = \left(1 - \frac{\tau_i^{(t)}}{\tau_{\max}^{(t)}}\right) w_i^{(t)}$, where $\tau_{\max}^{(t)} = \max_{i:w_i^{(t)}>0} \tau_i^{(t)}$.

Thus, the second step effectively filters out one (or more) data points at a time.

The scores are defined as $\tau_i^{(t)} = \langle v^{(t)}, X_i - \mu(w^{(t)})\rangle^2$, where $v^{(t)}$ is the top eigenvector of the weighted sample covariance $\Sigma^{(t)} = \Sigma(w^{(t)})$. The initial weights $w^{(0)}$ are uniform over S, and the weighted mean and covariance are defined as $\mu(w) = \sum_i w_i x_i$ and $\Sigma(w) = \sum_i w_i(x_i - \mu(w))(x_i - \mu(w))^T$. The 1D filter algorithm takes as input two vectors (τ, w) and outputs the vector with component i equal to $\left(1 - \frac{\tau_i}{\tau_{\max}}\right) w_i$, where $\tau_{\max} = \max_{i:w_i>0} \tau_i$, and effectively truncates one (or more) components at a time.

In the contaminated Gaussian setting, error bounds are of the form $O(\epsilon\sqrt{\log(1/\epsilon)})$. Sample complexity is $n = \Omega(d \log d)$.

7.3.2 Second algorithm

We now introduce an algorithm due to Lai, Rao, and Vempala (2016). They consider an *additive* adversarial model (see Examples sheet for the difference between this and the strong adversarial contamination model considered earlier): $(1 - \epsilon)n$ points are drawn i.i.d., and then ϵn are chosen adversarially.

The idea is that we only need to figure out which direction in d-dimensional space has been affected by contamination, and estimate this direction robustly (e.g., using a median). The algorithm is described in a recursive manner.

Algorithm 3: AGNOSTICMEAN(S)

Input: $S \subset \mathbb{R}^n$, and a routine OUTLIERREMOVAL($\cdot$).
Output: $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$.

1. Let $(\widetilde{S}, \boldsymbol{w}) =$ OUTLIERREMOVAL(S) .
2. **if** $n = 1$:
   (a) **if** $\boldsymbol{w} = -1$, **Return** median($\widetilde{S}$). //Gaussian case
   (b) **else Return** mean($\widetilde{S}$). //General case
3. Let $\boldsymbol{\Sigma}_{\widetilde{S},\boldsymbol{w}}$ be the weighted covariance matrix of $\widetilde{S}$ with weights $\boldsymbol{w}$, and V be the span of the top $n/2$ principal components of $\boldsymbol{\Sigma}_{\widetilde{S},\boldsymbol{w}}$, and W be its complement.
4. Set $S_1 := \boldsymbol{P}_V(S)$ where $\boldsymbol{P}_V$ is the projection operation on to V.
5. Let $\widehat{\boldsymbol{\mu}}_V :=$ AGNOSTICMEAN(S_1) and $\widehat{\boldsymbol{\mu}}_W := \text{mean}(\boldsymbol{P}_W \widetilde{S})$.
6. Let $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$ be such that $\boldsymbol{P}_V \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_V$ and $\boldsymbol{P}_W \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_W$.
7. **Return** $\widehat{\boldsymbol{\mu}}$.

7.3.3 Linear regression with adversarial contamination

Consider n i.i.d. observations from the linear model

$$y_i = x_i^T \beta^* + z_i,$$

where x_i's are zero-mean, identity covariance, and z_i's are independent, zero-mean noise. We observe a contaminated dataset, where up to ϵn points are arbitrarily corrupted, in covariates and/or responses.

Previous/concurrent approaches: Methods based on sum-of-squares algorithms or iterative robust gradient descent.

Pensia, Jog & Loh, 2020: Apply filtering step to the covariates. Then apply Huber regression.

8 Heavy-tailed distributions

Interestingly enough, the same types of estimators described earlier for adversarial contamination can also be used for optimal estimation, with high probability, when we have i.i.d. data drawn from heavy-tailed distributions. Going back to our PAC framework, we want to find an estimator which achieves the minimal function $t(n, \delta)$ in the bound

$$\mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \leq t(n, \delta)\right) \geq 1 - \delta,$$

where the probability holds for i.i.d. data $\{x_i\}_{i=1}^n$ drawn from an appropriate class of distributions. (Here, we no longer have ϵ since there is no contamination—though of course, one could introduce contamination on top of the more general distributional assumptions.) If $x_i \sim N(\mu, \sigma^2)$, we can take $t(n, \delta) = C\sigma\sqrt{\frac{\log(1/\delta)}{n}}$, and the bound is tight.

What if we consider classes of distributions which only satisfy the condition that the variance is bounded by σ^2? How can we estimate μ with optimal rates? In 1 dimension, note that Chebyshev's inequality guarantees that the mean satisfies

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n x_i - \mu\right| \leq \sigma\sqrt{\frac{1}{n\delta}}\right) \geq 1 - \delta$$

(and the bound can also be shown to be tight, e.g., when x_i is drawn from a distribution which is supported on $\{-a, 0, a\}$). But this rate (n, δ) is far worse than the rate of Gaussian variables when δ is small.

To provide a flavor of existing results in this area, we state results for two estimators, and provide proof details for the bounds in 1 dimension. Similar types of results hold for estimators based on filtering and trimmed mean techniques, with subtle differences between the assumptions needed for the bounds to hold (and be optimal).

8.1 MOM estimator

Suppose for simplicity that $n = mk$ for some integer m. We have the following result:

Theorem 9. *Suppose $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a distribution with mean μ and variance σ^2. Then the MOM estimator with $k = \lceil 8\log(1/\delta)\rceil$ satisfies*

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \leq \sigma\sqrt{\frac{4\lceil 8\log(1/\delta)\rceil}{n}}\right) \geq 1 - \delta.$$

Proof. We will show that

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \leq \sigma\sqrt{\frac{4}{m}}\right) \geq 1 - \exp(-k/8),$$

from which the result follows by taking $k = \lceil 8\log(1/\delta)\rceil$.

Let $\{z_j\}_{j=1}^k$ denote the means within individual blocks. Applying Chebyshev's inequality shows that for each j, with probability at least $\frac{3}{4}$, we have

$$|z_j - \mu| \leq \sigma\sqrt{\frac{4}{m}}.$$

If $|\widehat{\mu} - \mu| > \sigma\sqrt{\frac{4}{m}}$, then at least $\frac{k}{2}$ of the z_j's must satisfy $|z_j - \mu| > \sigma\sqrt{\frac{4}{m}}$, so denoting $Y \sim Binomial\left(k, \frac{1}{4}\right)$, we have

$$\begin{aligned}\mathbb{P}\left(|\widehat{\mu} - \mu| > \sigma\sqrt{\frac{4}{m}}\right) &\leq \mathbb{P}\left(Y \geq \frac{k}{2}\right) \\ &= \mathbb{P}\left(Y - \mathbb{E}(Y) \geq \frac{k}{4}\right) \\ &\leq \exp(-k/8),\end{aligned}$$

by standard tail bounds (i.e., Hoeffding's inequality). □

In fact, analysis of MOM estimator can be generalized to settings when variance might not be finite, but $1+\alpha$ moments exist for some $\alpha \in (0, 1)$, etc. Note that k only depends on the error probability and *not* any properties of the distribution.

A bound on a multivariate version of the MOM estimator using the geometric median can be derived in a somewhat analogous manner. However, the simple geometric median does not yield optimal error rates. MOM-based procedures have also been derived for settings such as linear regression and, more generally, empirical risk minimization.