Robust statistics

Po-Ling Loh

University of Cambridge, Department of Pure Mathematics and Mathematical Statistics

CUSO Summer School
Jura, Switzerland

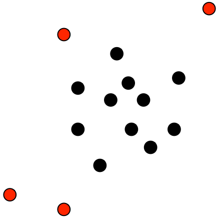
September 5-6, 2023

Introduction to robustness

- Deals with deviations from ideal models and their dangers for corresponding inference procedures
- Goal: 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

Robustness desiderata

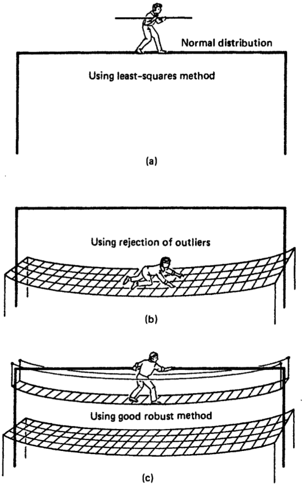


Figure 2. Various ways of analyzing data.

Hampel et al., *Robust Statistics*

Outline

1. Huber's perspective
   - Minimax bias
   - Minimax variance
2. Hampel's perspective
   - Influence functions
   - Optimal B-robust estimators
3. Extensions
   - Linear regression
   - Hypothesis testing
4. Modern perspectives
   - Adversarial contamination
   - Heavy-tailed distributions

Outline

1. Huber's perspective
   - Minimax bias
   - Minimax variance
2. Hampel's perspective
   - Influence functions
   - Optimal B-robust estimators
3. Extensions
   - Linear regression
   - Hypothesis testing
4. Modern perspectives
   - Adversarial contamination
   - Heavy-tailed distributions

Reference

- Huber & Ronchetti, "Robust Statististics"
- Huber, "Robust estimation of a location parameter," 1964

Location estimation

- Our goal is to estimate the *location parameter* of a distribution in one dimension:

$$x_1, \ldots, x_n \stackrel{i.i.d.}{\sim} F(t - \xi) = F_\xi(t),$$

where $F(t)$ is a cdf

- If the distribution corresponding to F is symmetric around 0, then $\mathbb{E}_{F_\xi}[x_i] = \xi$, so we could use the mean $\frac{1}{n}\sum_{i=1}^n x_i$—but 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.

Location estimation

- For $F \in \mathcal{P}_\epsilon(F_0)$, we have

$$\sup_t |F(t) - F_0(t)| = |(1-\epsilon)F_0(t) + \epsilon H(t) - F_0(t)|$$
$$= \epsilon \cdot \sup_t |H(t) - F_0(t)| \leq \epsilon,$$

  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

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

  implying the mean could be arbitrarily biased

- What about using the median?

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

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

is the *breakdown point* of T_n at X.

Breakdown point

- **Examples:**
  - The breakdown point of the mean is 0
  - The breakdown point of the median is $\frac{1}{n} \cdot \lfloor \frac{n-1}{2} \rfloor$
  - The median achieves the highest possible breakdown point among all translation-invariant estimators:

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

    for all $\{x_1, \dots, x_n\}$ and $a \in \mathbb{R}$
- However, this is a very rough notion, and has nothing to do with the distribution

Asymptotic bias

- Returning to the ϵ-contamination model, suppose F_0 is symmetric ($F_0 = \Phi$ for concreteness, though the arguments can be generalized)
- What is a bound on the (asymptotic) bias of the median?
- Clearly, worst case is when H concentrates all mass on one side of origin; median of $F \in \mathcal{P}_\epsilon$ is the solution to

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

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

Minimax bias

- Can we do better? Suppose $\{T_n\}$ is a sequence of estimators for a parameter $T(F_0)$, and define the *asymptotic bias* of a family of estimators $T = \{T_n\}$ as

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

- 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 the class $\mathcal{T}$ of translation-invariant estimators

Minimax bias

- An upper bound of b_0 can be achieved by the median
- To prove a lower bound, consider the distribution $F_+ \in P_\epsilon$ constructed as follows (shifted and centered around b_0):

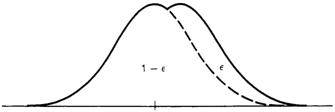


Exhibit 4.1 The distribution F_+ least favorable with respect to bias.

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

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

- Thus, the median is minimax optimal

Asymptotic variance

- Why do we use the sample mean as a location estimator anyway?

Theorem

Suppose the 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:

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

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

Minimax variance

- However, the situation may be more complicated when samples are from an ϵ-ball around some distribution
- Suppose $\sqrt{n}(T_n - T(F)) \stackrel{d}{\to} N(0, A(T, F))$, and consider the minimax problem

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

- Motivated by nice results in MLE theory, we restrict our attention to the class of M-estimators

Definition

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

Asymptotic theory

- For the following result, suppose $\psi = \rho'$ is nondecreasing

Theorem

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

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

Asymptotic theory

Corollary

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

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

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

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

Asymptotic theory

- In the corollary, symmetry of ρ and F implies $\mathbb{E}_F[\psi(x_i)] = 0$, so we can take $t_0 = 0$ in the theorem
- 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 the theorem in the case of the median

Contaminated distributions

- Now we 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 the conditions of the corollary
- Then
$$A(T, F) = \frac{\mathbb{E}_F[x_i^2]}{\mathbb{E}_F[1]^2} = (1-\epsilon) + \epsilon \mathbb{E}_H[x_i^2],$$
which can be arbitrarily large
- However, suppose we have a function ψ such that $\|\psi\|_\infty < k$ for some constant k; then
$$\frac{\mathbb{E}_F[\psi^2(x_i)]}{\mathbb{E}_F[\psi'(x_i)]^2} = \frac{(1-\epsilon)\mathbb{E}_\Phi[\psi^2(x_i)] + \epsilon \mathbb{E}_H[\psi^2(x_i)]}{\Big((1-\epsilon)\mathbb{E}_\Phi[\psi'(x_i)] + \epsilon \mathbb{E}_H[\psi'(x_i)]\Big)^2}$$
$$\leq \frac{(1-\epsilon)\mathbb{E}_\Phi[\psi^2(x_i)] + \epsilon k^2}{(1-\epsilon)^2 \mathbb{E}_\Phi[\psi'(x_i)]^2},$$
which is bounded as H ranges over different cdfs

Huber loss

- 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 $\psi(t) = \min\{k, \max(-k, t)\}$
- We could in theory try to minimize the upper bound with respect to k, though the derivation is rather tedious

Optimality of Huber loss

- In fact, the Huber estimator is actually minimax over *all* possible ψ
- The following result gives a constructive method for determining a saddlepoint solution to a generalized minimax problem

Theorem

Suppose G is the cdf of a log-concave symmetric distribution with twice continuously differentiable pdf g.

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

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

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.

Optimality of Huber loss

Theorem

(ii) *Furthermore, we have the explicit expressions*

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

and

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

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

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

Examples

- In the special case when $g(x) = \varphi(x)$, we can check that ψ_0 agrees with the Huber estimator
- If instead G is the cdf of a $\mathcal{N}(0, \sigma^2)$ distribution, we can derive

$$\psi_0(x) = \begin{cases} -k, & \text{if } x \leq -k\sigma^2, \\ \frac{x}{\sigma^2}, & \text{if } |x| < k\sigma^2, \\ k, & \text{if } x \geq k\sigma^2 \end{cases}$$

Kolmogorov neighborhood

- How much further can we push this theory? Consider the minimax variance problem when $\mathcal{P}_\epsilon^K(\Phi) = \{F : \sup_t |F(t) - \Phi(t)| < \epsilon\}$
- Recall that $\mathcal{P}_\epsilon(\Phi) \subseteq \mathcal{P}_\epsilon^K(\Phi)$
- 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

$$\psi_0(x) = \begin{cases} \omega \tan\left(\frac{\omega x}{2}\right), & \text{if } 0 \le x < x_0, \\ x, & \text{if } x_0 \le x \le x_1, \\ \lambda, & \text{if } x > x_1 \end{cases}$$

Outline

1. Huber's perspective
   - Minimax bias
   - Minimax variance
2. Hampel's perspective
   - Influence functions
   - Optimal B-robust estimators
3. Extensions
   - Linear regression
   - Hypothesis testing
4. Modern perspectives
   - Adversarial contamination
   - Heavy-tailed distributions

Reference

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

Hampel's approach

- Huber's approach relied heavily on nice form of normal density and symmetric contamination assumption; how can we "robustify" other estimation procedures?
- A second camp of robustness theory was developed by Hampel (1968) in his PhD thesis: "Contributions to the theory of robust estimation"

Hampel's approach

- Basic concepts are qualitative robustness (continuity of limiting functional), influence function (effect of infinitesimal perturbations), and breakdown point (distance to nearest singularity/asymptote)

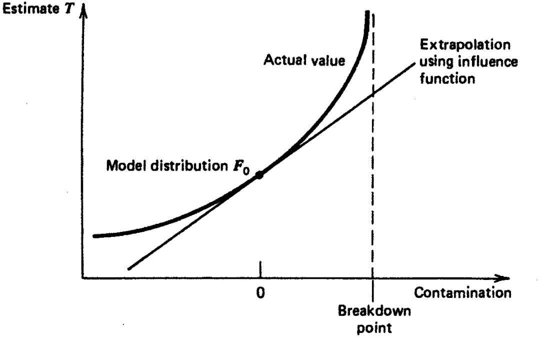


Figure 2. Extrapolation of a functional (estimator), using the infinitesimal approach. (Symbolic, using the analogue of an ordinary one-dimensional function.)

Influence functions

- Suppose we have a sequence of estimators satisfying $T_n(x_1, \ldots, x_n) \stackrel{P}{\to} T(F)$ when $x_i \sim F$

Definition

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

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

- This is a special case of a Gâteaux derivative of $T(F)$ in the direction of Δ_x
- In particular, we are interested in bounding the *gross-error sensitivity*

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

(analog of bounded derivative)

Influence functions

- Other quantities of interest include the local-shift sensitivity:

$$\lambda^*(T, F) := \sup_{x \neq y} \frac{|IF(y; T, F) - IF(x; T, F)|}{|y - x|},$$

- Rejection point:

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

- Change of variance functional: $CVF(x; T, F)$

Influence functions

- The influence function also relates to the asymptotic variance of T_n
- Under appropriate regularity conditions, when $x_i \stackrel{i.i.d.}{\sim} F$, we have

$$\begin{aligned}\sqrt{n}(T_n - T(F)) &\approx \sqrt{n}(T(F_n) - T(F)) \\ &\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(x_i; T, F) \\ &\stackrel{d}{\to} N(0, A(T, F)),\end{aligned}$$

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

Examples

- **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), $IF(x; T, F) = x$

Examples

- **Median:** We have

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

where $F_t := (1-t)F + t\Delta_x$, and differentiating the implicit equation

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

we can obtain

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

Examples

- **General M-estimators:** Since $T(F)$ is defined implicitly by

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

we can generalize the argument for the median to obtain

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

- 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 $\int IF(x, T, F)^2 dF(x)$

- Thus, the influence function is bounded if $\|\psi\|_\infty < \infty$

Optimality

- Hampel also derived optimality results with respect to the influence function
- 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 $\psi(y, \theta) = \psi(y - \theta)$)

Optimality

- 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

- Hampel studied the problem of minimizing the asymptotic variance $\int IF(x; T, F)^2 dF(x)$, subject to an upper bound on the gross error sensitivity $\gamma^*(T, F) = \sup_x |IF(x; T, F)|$

Optimality

Theorem

Suppose $F = F_\theta$ *(for a fixed* θ*) and* $I(F) = \int s(x,\theta)^2 dF(x) > 0$ *(this is the Fisher information). For any* $b > 0$*, there exists* $a \in \mathbb{R}$ *such that*

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

(truncated function) satisfies $\int \widetilde{\psi}(y) dF(y) = 0$ *and* $d := \int \widetilde{\psi}(y) s(y,\theta) dF(y) > 0$*. Furthermore,* $\widetilde{\psi}$ *uniquely 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* $\gamma^*(T,F) \leq c := \frac{b}{d}$.

Optimality

- 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 condition $\mathbb{E}_{F_\theta}[\psi(x_i - \theta)] = 0$
- Estimators that minimize the asymptotic variance subject to a bound on GES are *optimal B-robust estimators* (the B stands for "bias," whereas there are also V-robust estimators)
- Estimators such that $\gamma^*(T, F) < \infty$ are *B-robust*

Location M-estimators

- In this case,

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

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

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

Location M-estimators

- If $F = \Phi$, we have $\frac{-f'(y)}{f(y)} = y$, and we can take $a = 0$; this reduces to the Huber estimator with parameter b: $\widetilde{\psi}(y) = [y]_{-b}^{b}$!
- The finite-sample version solves $\sum_{i=1}^{n} [x_i - \theta]_{-b}^{b} = 0$
- Hence, the Huber estimator is also the optimal M-estimator for the location of a normal family with respect to B-robustness—different Huber parameters correspond to different bounds on γ^*

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)

- We can compute

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

so according to the theorem, the optimal B-robust estimator is

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

Scale M-estimators

- 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 solves

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

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

Most B-robust estimators

- Recall the optimality of the median according to Huber theory:

$$\min_{\{T_n\}} \max_{F \in \mathcal{P}_\epsilon(F_0)} b(\{T_n\}, F),$$

where F_0 is a symmetric, unimodal distribution

- We now provide an alternative result on optimality of the median according to Hampel's framework

Most B-robust estimators

- Assume F has a twice-differentiable density f which is symmetric around 0, log-concave, and satisfies $f(x) > 0$ for all x
- We restrict our attention to location M-estimators, where ψ ranges over a "nice" class of functions Ψ (smooth, except for a finite set of jumps $C(\psi)$)
- Hampel: "to our knowledge, Ψ covers all ψ-functions ever used for this estimation problem"

Definition

An estimator minimizing $\gamma^* := \sup_{x \in \mathbb{R} \setminus C(\psi)} |IF(x; \psi, F)|$ (for a fixed F, over a class of estimators Ψ) is called a **most** *B-robust* estimator.

Most B-robust estimators

Theorem

The median is the most B-robust estimator in Ψ. For all $\psi \in \Psi$, we have $\gamma^*(\psi, F) \geq \frac{1}{2f(0)}$, and equality holds if and only if ψ is the median estimator.

Reconciling Huber's and Hampel's approaches

- Minimax bias problem can be rephrased as

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

- For small ϵ, we make the approximation

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

Reconciling Huber's and Hampel's approaches

- Where (a) holds because

$$T((1-\epsilon)F + \epsilon\Delta_x) - T(F) \approx \epsilon \cdot IF(x; \psi, F),$$

and if T is linear, we can write

$$T((1-\epsilon)F + \epsilon H) - T(F) \approx \epsilon \cdot \int IF(x; \psi, F) dH(x)$$

- Hence, finding optimal ψ for minimax bias problem is (approximately) equivalent to solving

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

(resulting in median)

Reconciling Huber's and Hampel's approaches

- A connection can also be drawn between optimal B-robust and minimax variance estimators using influence function approximations
- Requires approximating the change-of-variance function, which is the change in asymptotic variance $V(\psi, F)$ when perturbed by a small mass at $(-x, x)$:

$$V\left(\psi, (1-\epsilon)F + \epsilon\left(\frac{1}{2}\Delta_x + \frac{1}{2}\Delta_{-x}\right)\right) - V(\psi, F) \approx \epsilon \cdot CVF(x; \psi, F)$$

Outline

1. Huber's perspective
   - Minimax bias
   - Minimax variance
2. Hampel's perspective
   - Influence functions
   - Optimal B-robust estimators
3. Extensions
   - Linear regression
   - Hypothesis testing
4. Modern perspectives
   - Adversarial contamination
   - Heavy-tailed distributions

Robust linear regression

- Analysis of multidimensional estimators becomes more complicated; however, results from the univariate case translate more easily into the context of linear regression
- Linear model:

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

where $x_i \overset{i.i.d.}{\sim} K$ 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

- 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

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

Robust linear regression

- When G_σ is cdf of $\mathcal{N}(0, \sigma^2)$, MLE corresponds to ordinary least squares, but OLS is not robust to deviations from normality
- To achieve robustness, consider regression M-estimator

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

(for now, assume σ is known)

- Estimating equation is

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

Influence functions

- By differentiating the implicit equation

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

  we can compute the influence function

$$IF(x_0, y_0; T, F) = M^{-1} \psi(y_0 - x_0^T T(F)) x_0,$$

  where

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

- Thus, we can guarantee boundedness of *IF* in response direction if ψ is bounded (this is not the case for OLS)

Asymptotic normality

- For fixed p and when $n \to \infty$, asymptotic covariance matrix is

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

- Minimizing $V(T,F)$ over the class of ψ functions then reduces to the familiar univariate problem of choosing ψ to minimize $\frac{\mathbb{E}_G[\psi^2(u)]}{\mathbb{E}_G[\psi'(u)]^2}$
- When $G = \Phi$, Huber M-estimator is again minimax optimal

B-robustness

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

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

- Since

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

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

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

B-robustness

- 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 $\gamma^*(T, F_\theta) < \infty$
- For this family of M-estimators, we have the lower bound

$$\gamma^*(T, F_\theta) \geq p\sqrt{\frac{\pi}{2}} \cdot \frac{1}{\mathbb{E}[\|x\|_2]}$$

  when $G = \Phi$

- Assuming radial symmetry of K, equality is achieved when $\psi(x) = \text{sign}(x)$, $w(x) = \frac{1}{\|x\|_2}$, and $v(x) = 1$, giving the most B-robust estimator
- In the radially symmetric case, the optimal B-robust estimator corresponds to the *Hampel-Krasker estimator*, with $v(x) = \|Ax\|_2 = \frac{1}{w(x)}$ and ψ equal to the Huber function

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 large loss in efficiency if σ is chosen poorly

Some approaches

- **Joint optimization:** We could jointly optimize the objective with respect to (θ, σ), but even if ρ is convex, the objective is generally nonconvex
- A clever idea by Huber is to jointly optimize

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

  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
- However, nonconvex ρ may lead to better robustness properties such as high breakdown point/finite rejection point

Some approaches

- *MM*-**estimators:**
  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 breakdown point and bounded influence function
- Asymptotic theory depends on assumption that $\widehat{\sigma}$ is sufficiently close to true scale parameter

Some approaches

- **Least trimmed squares (LTS):** Optimize

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

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

Robust hypothesis testing

- 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 $T_n(x_1, \dots, x_n)$
- Also suppose

$$T_n(x_1, \dots, x_n) \stackrel{\mathbb{P}}{\to} T(F),$$

  when $x_i \stackrel{i.i.d.}{\sim} F$
- We will define an influence function of a test, which is related to the influence function of the test statistic

Influence functions

- 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)
- 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

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

- Also assume 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

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

Influence functions

- In fact, by the chain rule, we can derive

$$IF_{\text{test}}(x; T, F_\theta) = \frac{1}{\xi'(\theta)} \cdot IF(x; T, F_\theta)$$

- 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 show how to characterize such types of stability using IF_{test}

Level and power

- 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 and $U_n := \xi^{-1}(T_n)$
- Similarly, the *asymptotic power* is

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

- Now define the perturbations

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

Level and power

- Finally, define 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 $IF_{\text{test}}(x; T, F)$

Level and power

Theorem

We have

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

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

where $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha)$ *and* $E(T, F) := \left(\int IF_{test}^2(y; T, F_{\theta_0}) dF_{\theta_0}(y)\right)^{-1}$.

- 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, subject to bounds on *LIF* and *PIF*
- Gives rise to tests based on truncated test statistics, censored likelihood ratio tests, etc.

Outline

1. Huber's perspective
   - Minimax bias
   - Minimax variance
2. Hampel's perspective
   - Influence functions
   - Optimal B-robust estimators
3. Extensions
   - Linear regression
   - Hypothesis testing
4. Modern perspectives
   - Adversarial contamination
   - Heavy-tailed distributions

References

- Lugosi & Mendelson, "Mean estimation and regression under heavy-tailed distributions: A survey"
- Diakonikolas & Kane, "Algorithmic high-dimensional robust statistics"
- Lerasle, "Selected topics on robust statistical learning theory"

Adversarial contamination

- Thus, far we assumed that contaminated data are drawn 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?

Adversarial contamination

- We will work in the (nonasymptotic) probably approximately correct (PAC) framework: Given $\delta > 0$, obtain an estimator $\widehat{\mu}(\widetilde{x}_1, \dots, \widetilde{x}_n)$ of $\mu = \mathbb{E}_F[x_i]$ 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
- The sample mean fails catastrophically in this framework: If $\epsilon \geq \frac{1}{n}$, the adversary 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

Lower bound

- We first give a lower bound for location estimation in one dimension

Theorem

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

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

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

- This is easily proven to be achievable by a median estimator

Upper bound

- In $d > 1$ dimensions, simplest idea is to take coordinatewise medians, but this gives $O(\epsilon\sqrt{d})$ error; we can achieve $O\left(\epsilon + \sqrt{\frac{d}{n}}\right)$ 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*.

- 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 maximizes this depth over all μ

Upper bound

Theorem

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*

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

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

Alternative methods

- Ongoing research tries to match error rate of Tukey median in general distributional families, without computational barriers
- Filtering algorithm (Diakonikolas et al.):
  - Iteratively flags outliers based on projections onto maximal principal components
  - For contaminated Gaussians, achieves $O(\epsilon\sqrt{\log(1/\epsilon})$ error with $n = \Omega(d \log d)$ samples

Alternative methods

- Trimmed means algorithm:
  - In one dimension: First split sample into two parts, one of which is used to determine trimming parameters (α, β) according to quantiles
  - Then take $\sum_{i=1}^n \phi_{\alpha,\beta}(y_i)$, where

$$\phi_{\alpha,\beta}(y) = \begin{cases} \beta & \text{if } y > \beta, \\ y & \text{if } \alpha \leq y \leq \beta, \\ \alpha & \text{if } y < \alpha \end{cases}$$

  - Extension to multiple dimensions is somewhat complicated, but roughly seeks an estimator which is close to trimmed mean of projected data in any direction $v \in \mathbb{R}^d$

Alternative methods

- Median of means (MOM) estimator:
  - Divide sample into k blocks, and compute sample mean within each block; then aggregate k values by taking a median
  - In high dimensions, correct notion of median is also not so straightforward (coordinatewise medians/geometric medians do not yield provable dimension-free rates for adversarial contmination)
  - Estimator with optimal rates can be obtained by finding an estimator close to the MOM estimator of the projected data in any direction $v \in \mathbb{R}^d$, as in the case of the trimmed mean, but is again computationally intractable

Heavy-tailed distributions

- Interestingly, the same types of estimators used for adversarial contamination can often be used for optimal estimation, w.h.p., for i.i.d. data drawn from heavy-tailed distributions
- Going back to the 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
- 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

Heavy-tailed distributions

- What if we consider classes of distributions which only satisfy the condition that the variance is bounded by σ^2?
- In one dimension, Chebyshev's inequality guarantees that the mean satisfies

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

  (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
- We will just give a flavor of results in 1 dimension

Median-of-means estimator

Theorem

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* $k = \lceil 8 \log(1/\delta) \rceil$ *bins satisfies*

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

- A multivariate version of the MOM estimator based on geometric medians does not quite yield optimal error rates in d

Catoni's *M*estimator

- Returning to the framework of classical M-estimation, take a parameter $\alpha > 0$ and define $\widehat{\mu}$ as the solution to the estimating equation

$$\sum_{i=1}^{n} \psi(\alpha(x_i - \xi)) = 0,$$

where ψ is a nondecreasing function satisfying

$$-\log\left(1 - t + \frac{t^2}{2}\right) \leq \psi(t) \leq \log\left(1 + t + \frac{t^2}{2}\right), \qquad \forall t \in \mathbb{R}$$

- The Huber ψ function does not quite satisfy these bounds

Catoni's M-estimator

Theorem

Suppose $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a distribution with mean μ and variance σ^2. Suppose $\delta > 0$ and $n > 2\log(2/\delta)$. Then Catoni's M-estimator with parameter

$$\alpha = \sqrt{\frac{2\log(2/\delta)}{n\sigma^2\left(1 + \frac{2\log(2/\delta)}{n-2\log(2/\delta)}\right)}},$$

satisfies

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \le \sqrt{\frac{2\sigma^2\log(2/\delta)}{n - 2\log(2/\delta)}}\right) \ge 1 - \delta.$$

- The proof proceeds by using Chernoff bounds and bounding mgfs
- A disadvantage is that α depends on σ, although adaptive choices of α exist when an upper bound on σ^2 is known a priori
- Multivariate versions of Catoni's M-estimator have also been derived

Summary

- **Huber's perspective:** Deriving minimax optimal estimators in an ϵ-ball around true distribution (asymptotic bias, asymptotic variance)
- **Hampel's perspective:** Deriving optimal estimators involving quantities related to influence functions (minimum GES, minimum asymptotic variance subject to bound on GES)
- Extensions to linear regression and hypothesis testing
- **Modern perspectives:** Nonasymptotic guarantees, new contamination models, computational feasibility in high dimensions

Thank you!