Robust estimation of a mean

in a multivariate Gaussian model: Part 1

Frejus, December 17, 2018

Arnak S. Dalalyan

ENSAE ParisTech / CREST

1. Various models of contamination

General notation

We first introduce the notation that are common to all the models of contamination considered in this talk.

- Number of observations : n.
- Dimension of the unknown parameter $\boldsymbol{\mu}^*$: p.
- Observations $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \sim \boldsymbol{P}_n$.
- Number of outliers (possibly random): $s \in \{1, \ldots, n\}$.
- Set of outliers: $S \subset \{1, \ldots, n\}$.
- Proportion of outliers: $\varepsilon = \mathbf{E}[s/n] = \mathbf{E}[|S|/n]$.

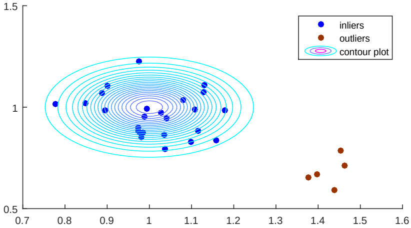
Setting (informal)

Among the n observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, there is a small number s of outliers. If we remove the outliers, all the other $\boldsymbol{X}_i$'s are iid drawn from a reference distribution $\boldsymbol{P}_{\boldsymbol{\mu}^*}$.

Gaussian model with unknown mean

Assumption (model for inliers)

Throughout this presentation, we assume that the reference distribution $\boldsymbol{P}_{\boldsymbol{\mu}^*}$ is p-variate Gaussian $\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$. The goal is to estimate the parameter $\boldsymbol{\mu}^* \in \mathbb{R}^p$.

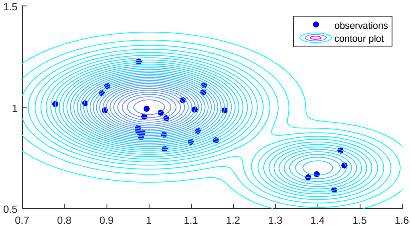


$$n = 30, \; s = 5, \; \boldsymbol{\mu}^* = [1, 1]^\top$$

Gaussian model with unknown mean

Assumption (model for inliers)

Throughout this presentation, we assume that the reference distribution $\boldsymbol{P}_{\boldsymbol{\mu}^*}$ is p-variate Gaussian $\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$. The goal is to estimate the parameter $\boldsymbol{\mu}^* \in \mathbb{R}^p$.



$$n = 30, \ s = 5, \ \boldsymbol{\mu}^* = [1, 1]^\top$$

Huber's contamination

Assumption (HC model for outliers)

There are unobserved iid random variables $Z_1, \ldots, Z_n \sim \mathscr{B}(\varepsilon)$ and a distribution $\boldsymbol{Q}$, such that

$$\mathscr{L}(\boldsymbol{X}_i | Z_i = 0) = \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p), \qquad \mathscr{L}(\boldsymbol{X}_i | Z_i = 1) = \boldsymbol{Q},$$

the observations X_i corresponding to different i's are independent.
This is equivalent to

$$\boldsymbol{P}_n = \big\{(1-\varepsilon)\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p) + \varepsilon\boldsymbol{Q}\big\}^{\otimes n}.$$

In this model,

$$\underbrace{S = \{i : Z_i = 1\}}_{\text{set of outliers}} \text{ and } \underbrace{s \sim \mathscr{B}(n, \varepsilon)}_{\text{nb of outliers}}$$

are both random.

Huber's contamination

Assumption (HC model for outliers)

There are unobserved iid random variables $Z_1, \ldots, Z_n \sim \mathscr{B}(\varepsilon)$ and a distribution $\boldsymbol{Q}$, such that

$$\mathscr{L}(\boldsymbol{X}_i | Z_i = 0) = \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p), \qquad \mathscr{L}(\boldsymbol{X}_i | Z_i = 1) = \boldsymbol{Q},$$

the observations X_i corresponding to different i's are independent. This is equivalent to

$$\boldsymbol{P}_n = \big\{(1-\varepsilon)\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p) + \varepsilon \boldsymbol{Q}\big\}^{\otimes n}.$$

In this model,

$$\underbrace{S = \{i : Z_i = 1\}}_{\text{set of outliers}} \text{ and } \underbrace{s \sim \mathscr{B}(n, \varepsilon)}_{\text{nb of outliers}}$$

are both random.

i	Z_i	$\boldsymbol{X}_i \sim$
1	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
2	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
3	1	$\boldsymbol{Q}$
4	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
5	1	$\boldsymbol{Q}$
6	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
7	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
$\vdots$	$\vdots$	$\vdots$
30	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
$\boldsymbol{s} =$	**5**	$\sim \mathscr{B}(30, 0.2)$

Huber's contamination

Assumption (HC model for outliers)

There are unobserved iid random variables $Z_1, \ldots, Z_n \sim \mathscr{B}(\varepsilon)$ and a distribution $\boldsymbol{Q}$, such that

$$\mathscr{L}(\boldsymbol{X}_i | Z_i = 0) = \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p), \qquad \mathscr{L}(\boldsymbol{X}_i | Z_i = 1) = \boldsymbol{Q},$$

the observations X_i corresponding to different i's are independent. This is equivalent to

$$\boldsymbol{P}_n = \big\{(1-\varepsilon)\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p) + \varepsilon \boldsymbol{Q}\big\}^{\otimes n}.$$

In this model,

$$\underbrace{S = \{i : Z_i = 1\}}_{\text{set of outliers}} \text{ and } \underbrace{s \sim \mathscr{B}(n, \varepsilon)}_{\text{nb of outliers}}$$

are both random.

i	Z_i	$\boldsymbol{X}_i \sim$
1	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
2	1	$\boldsymbol{Q}$
3	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
4	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
5	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
6	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
7	1	$\boldsymbol{Q}$
$\vdots$	$\vdots$	$\vdots$
30	0	$\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$
$\boldsymbol{s} =$	$\boldsymbol{6}$	$\sim \mathscr{B}(30, 0.2)$

Huber's contamination

Assumption (HC model for outliers)

There are unobserved iid random variables $Z_1, \ldots, Z_n \sim \mathscr{B}(\varepsilon)$ and a distribution $\boldsymbol{Q}$, such that

$$\mathscr{L}(\boldsymbol{X}_i | Z_i = 0) = \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p), \qquad \mathscr{L}(\boldsymbol{X}_i | Z_i = 1) = \boldsymbol{Q},$$

the observations X_i corresponding to different i's are independent. This is equivalent to

$$\boldsymbol{P}_n = \left\{(1-\varepsilon)\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p) + \varepsilon \boldsymbol{Q}\right\}^{\otimes n}.$$

In this model,

$$\underbrace{S = \{i : Z_i = 1\}}_{\text{set of outliers}} \text{ and } \underbrace{s \sim \mathscr{B}(n, \varepsilon)}_{\text{nb of outliers}}$$

are both random.

We write

$$\boxed{P_n \in \mathcal{M}_n^{\mathrm{HC}}(p, \varepsilon, \boldsymbol{\mu}^*).}$$

for the model of Huber's contamination.

Huber's deterministic contamination

Assumption (HDC model for outliers)

There is a set $S \subset \{1, \ldots, n\}$ of cardinality $s = [n\varepsilon]$ and a distribution $\boldsymbol{Q}$, such that

$$\{\boldsymbol{X}_i : i \in S^c\} \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p) \quad \perp\!\!\!\perp \quad \{\boldsymbol{X}_i : i \in S\} \overset{\text{iid}}{\sim} \boldsymbol{Q}.$$

- Similar to HC: the outliers are iid.
- Different from HC: the set of outliers is determenistic.

Remark *The number of outliers s should be smaller than $n/2$, otherwise $\boldsymbol{Q}$ would be the reference distribution and $\mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$ the contamination.*

We write $P_n \in \mathcal{M}_n^{\text{HDC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.

Parameter contamination

Assumption (PC model for outliers)

There is a set $S \subset \{1, \ldots, n\}$ of cardinality $s = [n\varepsilon]$ and a collection of vectors $\{\boldsymbol{\mu}_i : i \in S\}$, such that

$$\{\boldsymbol{X}_i : i \in S^c\} \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p) \quad \perp\!\!\!\perp \quad \{\boldsymbol{X}_i : i \in S\} \sim \bigotimes_{i \in S} \mathcal{N}_p(\boldsymbol{\mu}_i, \mathbf{I}_p).$$

- Similar to HC & HDC: the outliers are independent.
- Different from HC & HDC: the outliers might have different distributions.

We write $P_n \in \mathcal{M}_n^{\mathrm{PC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.

Adversarial contamination

Assumption (AC model for outliers)

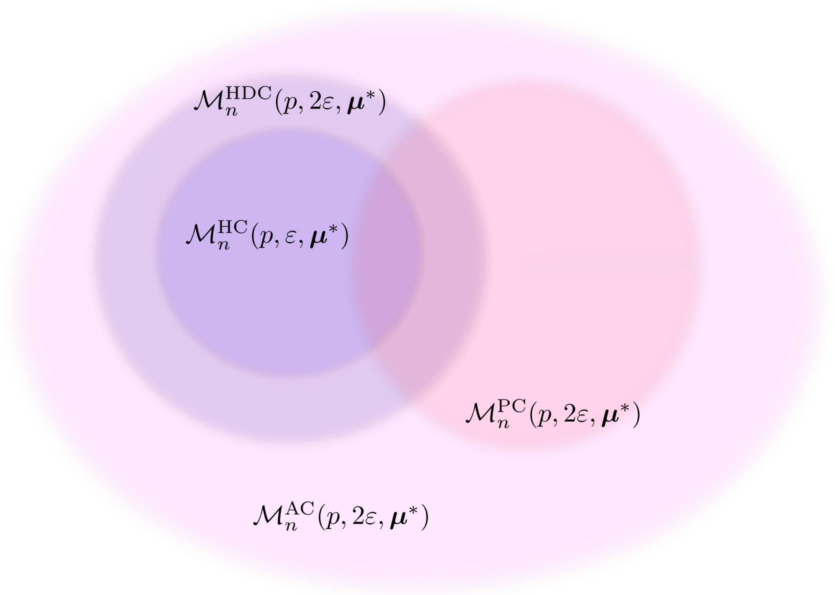
For a sequence $\boldsymbol{Y}_i \stackrel{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}^*, \mathbf{I}_p)$, $i = 1, \ldots, n$, and a random set $S \subset \{1, \ldots, n\}$ of cardinality $s = [n\varepsilon]$ we have

$$\boldsymbol{X}_i = \boldsymbol{Y}_i, \qquad \forall i \in S^c.$$

- The set S **is not** independent of $\{\boldsymbol{Y}_i : i = 1, \ldots, n\}$.
- The observations $\{\boldsymbol{X}_i : i \in S\}$ may have arbitrary dependence structure.

We write $P_n \in \mathcal{M}_n^{\mathrm{AC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.

Relation between the models



2. Problem formulation and overview of results

Historical approach

Breakdown point

- Assume the unknown parameter $\boldsymbol{\mu}^*$ is in $\mathbb{R}^p$.
- Let $\widehat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}^*$. Thus,

$$\widehat{\boldsymbol{\mu}} : \bigcup_{n=1}^{\infty} \mathcal{X}^n \to \mathbb{R}^p.$$

- The breakdown point ε_n^* of $\widehat{\boldsymbol{\mu}}$ is defined by

$$\varepsilon_n^* = \frac{1}{n} \min \Big\{ s \in \{1, \ldots, n\} : \sup_{y_1, \ldots, y_s} \|\widehat{\boldsymbol{\mu}}(\boldsymbol{x}_{1:(n-s)}, \boldsymbol{y}_{1:s})\| = +\infty \Big\}.$$

- Drawbacks:
  - does not take into account the impact of "mild" outliers,
  - meaningless if the parameter space is bounded,
  - does not depend on the norm under consideration,
  - ...

Minimax approach
In expectation

- A more informative way of quantifying the robustness is the evaluation of the worst-case risk and its comparison to the minimax risk.
- Worst-case risk of an estimator $\widehat{\boldsymbol{\mu}}_n$:

$$R^{\star}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n) = \sup_{\boldsymbol{\mu}^*} \sup_{\boldsymbol{P}_n \in \mathcal{M}^{\star}_n(p,\varepsilon,\boldsymbol{\mu}^*)} \mathbf{E}_{\mathbf{X}\sim \boldsymbol{P}_n}[\|\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) - \boldsymbol{\mu}^*\|_2^2].$$

  - Here, $\mathcal{M}^{\star}_n(p,\varepsilon,\boldsymbol{\mu}^*)$ is one of the 4 models of contamination considered in previous slides.
  - For instance, $R^{\mathrm{HC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$ is the minimax risk for Huber's contamination model.
- Minimax risk:

$$R^{\star}_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}_n} R^{\star}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n).$$

Minimax approach
In expectation

- A more informative way of quantifying the robustness is the evaluation of the worst-case risk and its comparison to the minimax risk.
- Worst-case risk of an estimator $\widehat{\boldsymbol{\mu}}_n$:

$$R^{\rm HC}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n) = \sup_{\boldsymbol{\mu}^*} \sup_{\boldsymbol{P}_n \in \mathcal{M}^{\rm HC}_n(p,\varepsilon,\boldsymbol{\mu}^*)} \mathbf{E}_{\mathbf{X}\sim \boldsymbol{P}_n}[\|\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) - \boldsymbol{\mu}^*\|_2^2].$$

  - Here, $\mathcal{M}^{\star}_n(p,\varepsilon,\boldsymbol{\mu}^*)$ is one of the 4 models of contamination considered in previous slides.
  - For instance, $R^{\rm HC}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$ is the minimax risk for Huber's contamination model.
- Minimax risk:

$$R^{\rm HC}_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}_n} R^{\rm HC}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n).$$

Minimax approach
In expectation

- A more informative way of quantifying the robustness is the evaluation of the worst-case risk and its comparison to the minimax risk.
- Worst-case risk of an estimator $\widehat{\boldsymbol{\mu}}_n$:

$$R_{n,p,\varepsilon}^{\mathrm{HDC}}(\widehat{\boldsymbol{\mu}}_n) = \sup_{\boldsymbol{\mu}^*} \sup_{\boldsymbol{P}_n \in \mathcal{M}_n^{\mathrm{HDC}}(p,\varepsilon,\boldsymbol{\mu}^*)} \mathbf{E}_{\mathbf{X}\sim \boldsymbol{P}_n}[\|\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) - \boldsymbol{\mu}^*\|_2^2].$$

  - Here, $\mathcal{M}_n^{\star}(p,\varepsilon,\boldsymbol{\mu}^*)$ is one of the 4 models of contamination considered in previous slides.
  - For instance, $R_{n,p,\varepsilon}^{\mathrm{HC}}(\widehat{\boldsymbol{\mu}}_n)$ is the minimax risk for Huber's contamination model.
- Minimax risk:

$$R_{n,p,\varepsilon}^{\mathrm{HDC}} = \inf_{\widehat{\boldsymbol{\mu}}_n} R_{n,p,\varepsilon}^{\mathrm{HDC}}(\widehat{\boldsymbol{\mu}}_n).$$

Minimax approach
In expectation

- A more informative way of quantifying the robustness is the evaluation of the worst-case risk and its comparison to the minimax risk.
- Worst-case risk of an estimator $\widehat{\boldsymbol{\mu}}_n$:

$$R^{\mathrm{PC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n) = \sup_{\boldsymbol{\mu}^*} \sup_{\boldsymbol{P}_n \in \mathcal{M}_n^{\mathrm{PC}}(p,\varepsilon,\boldsymbol{\mu}^*)} \mathbf{E}_{\mathbf{X}\sim\boldsymbol{P}_n}[\|\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) - \boldsymbol{\mu}^*\|_2^2].$$

  - Here, $\mathcal{M}_n^{\star}(p,\varepsilon,\boldsymbol{\mu}^*)$ is one of the 4 models of contamination considered in previous slides.
  - For instance, $R^{\mathrm{HC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$ is the minimax risk for Huber's contamination model.
- Minimax risk:

$$R^{\mathrm{PC}}_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}_n} R^{\mathrm{PC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n).$$

Minimax approach
In expectation

- A more informative way of quantifying the robustness is the evaluation of the worst-case risk and its comparison to the minimax risk.
- Worst-case risk of an estimator $\widehat{\boldsymbol{\mu}}_n$:

$$R^{\mathrm{AC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n) = \sup_{\boldsymbol{\mu}^*} \sup_{\boldsymbol{P}_n \in \mathcal{M}^{\mathrm{AC}}_n(p,\varepsilon,\boldsymbol{\mu}^*)} \mathbf{E}_{\mathbf{X}\sim \boldsymbol{P}_n}[\|\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) - \boldsymbol{\mu}^*\|_2^2].$$

  - Here, $\mathcal{M}^{\star}_n(p,\varepsilon,\boldsymbol{\mu}^*)$ is one of the 4 models of contamination considered in previous slides.
  - For instance, $R^{\mathrm{HC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$ is the minimax risk for Huber's contamination model.
- Minimax risk:

$$R^{\mathrm{AC}}_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}_n} R^{\mathrm{AC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n).$$

Minimax approach
In deviation

- Most results in the literature provide bounds on the deviation, not for the expectation.
- Fix a confidence level $\delta \in (0,1)$.
- Worst-case deviation of an estimator $\widehat{\boldsymbol{\mu}}_n$: $r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$ is solution to

$$\begin{aligned} \text{minimize} \quad & r \\ \text{subject to} \quad & \mathbf{P}_{\mathbf{X}\sim\boldsymbol{P}_n}\big(\|\widehat{\boldsymbol{\mu}}_n(\mathbf{X}) - \boldsymbol{\mu}^*\|_2^2 > r\big) \leq \delta \\ & \forall \boldsymbol{\mu}^* \in \mathbb{R}^p, \forall \boldsymbol{P}_n \in \mathcal{M}_n^\star(p,\varepsilon,\boldsymbol{\mu}^*). \end{aligned}$$

  Clearly, $r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$ depends on δ, but we will not be interested in this dependence.
- Minimax risk:

$$r^\star_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}_n} r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n).$$

- Tchebychev's inequality yields $\delta r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n) \leq R^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}_n)$.

Common robust estimators of the mean

- The most common robust estimators of the mean are perhaps the coordinatewise median, the geometric median and the Huber's estimator.
- All these estimators can be defined as an M-estimator:

$$\widehat{\boldsymbol{\mu}}_n \in \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^n \Psi(\boldsymbol{X}_i - \boldsymbol{\mu})$$

with

$$\Psi(\boldsymbol{x}) = \begin{cases} \|\boldsymbol{x}\|_1, & \text{coordinatewise median,} \\ \|\boldsymbol{x}\|_2, & \text{geometric median,} \\ \frac{\|\boldsymbol{x}\|_2^2}{2} \wedge \lambda(\|\boldsymbol{x}\|_2 - 0.5\lambda), & \text{Huber's estimator.} \end{cases}$$

- In all the three cases, the function Ψ is convex and the estimator is computable in polynomial time.

Overview of the results

Minimax rates in deviation

- Lower bound on the minimax risk (Chen et al., 2015):

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \geq c\big(\tfrac{p}{n} + \varepsilon^2\big)}$$

where c is a constant depending only on δ.

- Upper bound on the minimax risk (Chen et al., 2015):

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \leq C\big(\tfrac{p}{n} + \varepsilon^2\big)}$$

where C is a constant depending only on δ.

- It is attained by Tukey's median, which is not computationally tractable.
- The coordinatewise median, the geometric median and the Huber estimator are sub-optimal:

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}) \geq \bar{c}\big(\tfrac{p}{n} + p\varepsilon^2\big).}$$

Overview of the results

Minimax rates in deviation

- Lower bound on the minimax risk (Chen et al., 2015):

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \geq c\big(\tfrac{p}{n} + \varepsilon^2\big)}$$

  where c is a constant depending only on δ.

- Upper bound on the minimax risk (Chen et al., 2015):

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \leq C\big(\tfrac{p}{n} + \varepsilon^2\big)}$$

  where C is a constant depending only on δ.

- It is attained by Tukey's median, which is not computationally tractable.
- The coordinatewise median, the geometric median and the Huber estimator are sub-optimal:

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}) \geq \bar{c}\big(\tfrac{p}{n} + p\varepsilon^2\big).}$$

Overview of the results

Minimax rates in deviation

- Lower bound on the minimax risk (Chen et al., 2015):

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \geq c\left(\frac{p}{n} + \varepsilon^2\right)}$$

where c is a constant depending only on δ.

- Upper bound on the minimax risk (Chen et al., 2015):

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \leq C\left(\frac{p}{n} + \varepsilon^2\right)}$$

where C is a constant depending only

- It is attained by Tukey's median, which
tractable.

- The coordinatewise median, the geome r
estimator are sub-optimal:

$$\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}}) \geq \bar{c}\left(\frac{p}{n} + p\varepsilon^2\right).}$$

There is an extra factor $\boldsymbol{p}$.

We will present three tractable estimators that improve on the coordinatewise median.

1. The ellipsoid method (Diakonikolas et al., 2016).
2. The spectral method (Lai et al., 2016).
3. The iterative soft thresholding (Collier and Dalalyan, 2017).

3. The minimax rate

Minimax lower bound

Theorem 1 (Chen et al., 2015)

There is a constant $c > 0$ such that for every $\varepsilon \in [0,1]$ and every $\delta \in (0,1/2)$, it holds that

$$r^{\mathrm{HC}}_{n,p,\varepsilon} \geq c\left(\frac{p}{n} + \varepsilon^2\right).$$

Some remarks

- By Tchebychev's inequality, $\frac{p}{n} + \varepsilon^2$ is also a lower bound for the minimax risk in expectation.
- By inclusion, $\frac{p}{n} + \varepsilon^2$ is also a lower bound for the minimax risk in models HDC and AC.
- The same lower bound $\frac{p}{n} + \varepsilon^2$ holds true for the model PC.

Proof of the lower bound 1

1. From the classic parametric minimax theory: $\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \gtrsim \frac{p}{n}.}$

2. Thus, we need only to show that $\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \gtrsim \varepsilon^2.}$

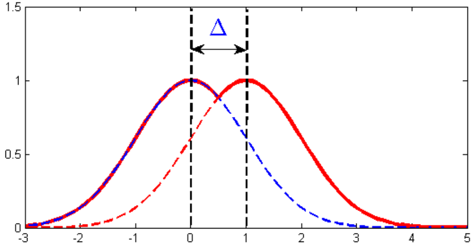
3. Main steps of the proof:
   - Reduction to dimension 1: $r^{\mathrm{HC}}_{n,\boldsymbol{p},\varepsilon} \geq r^{\mathrm{HC}}_{n,\mathbf{1},\varepsilon}$.
   - Construct a probability density function f_ε such that

$$\begin{aligned} f_\varepsilon^{\otimes n} &\in \mathcal{M}_n^{\mathrm{HC}}(1,\varepsilon,0) \\ f_\varepsilon^{\otimes n} &\in \mathcal{M}_n^{\mathrm{HC}}(1,\varepsilon,\Delta_\varepsilon) \end{aligned} \quad \text{with} \quad \Delta_\varepsilon \asymp \varepsilon.$$

   - Parameter values $\boldsymbol{\mu}^* = 0$ and $\boldsymbol{\mu}^* = \Delta_\varepsilon$ are indistinguishable from the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim f_\varepsilon^{\otimes n}$.
   - Therefore $\boxed{r^{\mathrm{HC}}_{n,p,\varepsilon} \gtrsim \|\Delta_\varepsilon - 0\|_2^2 \asymp \varepsilon^2.}$

Proof of the lower bound 2

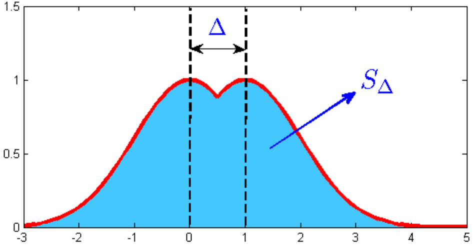
For a $\Delta > 0$, define $f^\circ_\Delta = \varphi_0 \vee \varphi_\Delta$.



Proof of the lower bound 2

For a $\Delta > 0$, define $f_\Delta^\circ = \varphi_0 \vee \varphi_\Delta$.

We have $S_\Delta = \int f_\Delta^\circ(x)\,dx = 1 + a\Delta + O(\Delta^2)$.

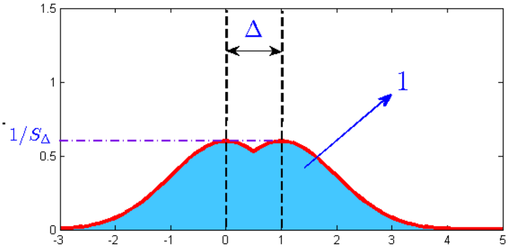


Proof of the lower bound 2

For a $\Delta > 0$, define $f_\Delta^\circ = \varphi_0 \vee \varphi_\Delta$.

We have $S_\Delta = \int f_\Delta^\circ(x)\,dx = 1 + a\Delta + O(\Delta^2)$.

Then, $f_\Delta = f_\Delta^\circ / S_\Delta$ is a pdf.



Proof of the lower bound 2

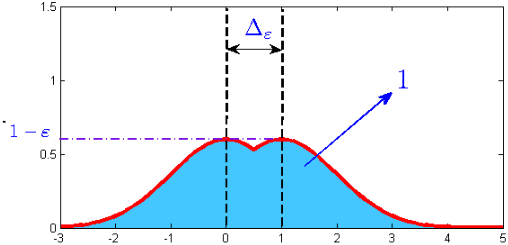
For a $\Delta > 0$, define $f^\circ_\Delta = \varphi_0 \vee \varphi_\Delta$.

We have $S_\Delta = \int f^\circ_\Delta(x)\,dx = 1 + a\Delta + O(\Delta^2)$.

Then, $f_\Delta = f^\circ_\Delta / S_\Delta$ is a pdf.

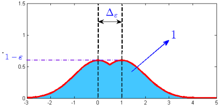
We choose Δ_ε so that $1/S_{\Delta_\varepsilon} = 1 - \varepsilon$

and set $f_\varepsilon = f_{\Delta_\varepsilon}$.



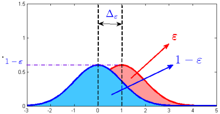
Proof of the lower bound 3

$$f_\varepsilon = (1-\varepsilon)(\varphi_0 \vee \varphi_{\Delta_\varepsilon})$$



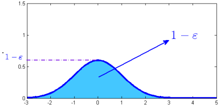
Proof of the lower bound 3

$$f_\varepsilon = (1-\varepsilon)(\varphi_0 \vee \varphi_{\Delta_\varepsilon})$$



$$\|$$

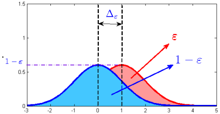
$$(1-\varepsilon)\varphi_0$$



$$+$$

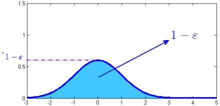
Proof of the lower bound 3

$$f_\varepsilon = (1-\varepsilon)(\varphi_0 \vee \varphi_{\Delta_\varepsilon})$$



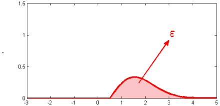
$$\parallel$$

$$(1-\varepsilon)\varphi_0$$



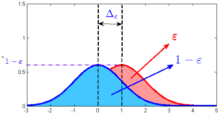
$$+$$

$$\varepsilon q$$



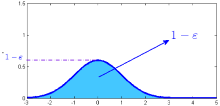
Proof of the lower bound 3

$$f_\varepsilon = (1-\varepsilon)(\varphi_0 \vee \varphi_{\Delta_\varepsilon})$$



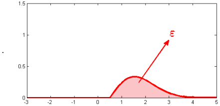
$$\|$$

$$(1-\varepsilon)\varphi_0$$



$$+$$

$$\varepsilon q$$



$$\boxed{f_\varepsilon = \begin{matrix}(1-\varepsilon)\varphi_0 & + & \varepsilon q\\ (1-\varepsilon)\varphi_{\Delta_\varepsilon} & + & \varepsilon q'\end{matrix}.}$$

Minimax upper bound

Theorem 2 (Chen et al., 2015)

There are two constants $C_1, C_2 > 0$ such that

- for every $\varepsilon \leq 1/5$
- for every $p \leq C_1 n$
- for every $\delta \geq e^{-C_1 n}$,

it holds that

$$r_{n,p,\varepsilon}^{\mathrm{HC}} \leq C_2\left(\frac{p}{n} + \varepsilon^2 + \frac{\log 1/\delta}{n}\right).$$

Some remarks:

- The upper bound is attained by Tukey's median.
- The condition $\varepsilon \leq 1/5$ can be replaced by $\varepsilon \leq 1/3 - c'$, with an arbitrarily small $c' > 0$.
- The estimator does not rely on the knowledge of ε.

Tukey's median

- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^n \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \le \boldsymbol{u}^\top \boldsymbol{\mu}).$$

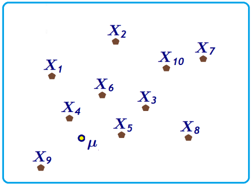
Tukey's median

- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \leq \boldsymbol{u}^\top \boldsymbol{\mu}).$$



depth of $\boldsymbol{\mu}$ in $\boldsymbol{X}_{1:10}$?

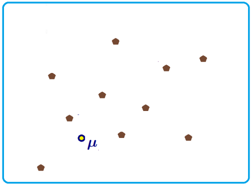
Tukey's median

- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \le \boldsymbol{u}^\top \boldsymbol{\mu}).$$



depth of $\boldsymbol{\mu}$ in $\boldsymbol{X}_{1:10}$?

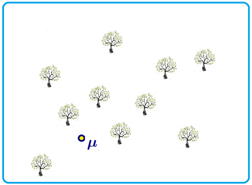
Tukey's median

- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \leq \boldsymbol{u}^\top \boldsymbol{\mu}).$$



depth of $\boldsymbol{\mu}$ in the forest ?

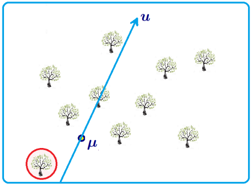
Tukey's median

- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \le \boldsymbol{u}^\top \boldsymbol{\mu}).$$



$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = 1.$$

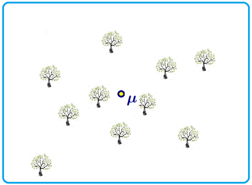
Tukey's median

- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^{n} \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \le \boldsymbol{u}^\top \boldsymbol{\mu}).$$



depth of $\boldsymbol{\mu}$ in the forest ?

Tukey's median

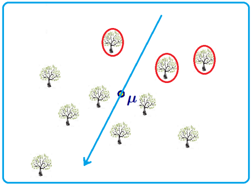
- The upper bound is attained by Tukey's median.
- Tukey's median is any maximaizer of Tukey's depth:

$$\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}} \in \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^p} \mathcal{D}(\boldsymbol{\mu}, \{\boldsymbol{X}_{1:n}\}).$$

- Tukey's (halfspace) depth is

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = \min_{\boldsymbol{u}\in\mathbb{S}_1} \sum_{i=1}^n \mathbb{1}(\boldsymbol{u}^\top \boldsymbol{X}_i \le \boldsymbol{u}^\top \boldsymbol{\mu}).$$

- $\widehat{\boldsymbol{\mu}}_n^{\mathrm{TM}}$ is computationally intractable for large p.



$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{X}_{1:n}) = 3$.

Summary

- We introduced four models of contamination by outliers:
  - Huber's contamination $\mathcal{M}_n^{\mathrm{HC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Huber's deterministic contamination $\mathcal{M}_n^{\mathrm{HDC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Parameter contamination $\mathcal{M}_n^{\mathrm{PC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Adversarial contamination $\mathcal{M}_n^{\mathrm{AC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
- We have defined the worst case risks in expectation and in deviation, $R^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$ and $r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.
- We have defined the minimax risks $R^\star_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}} R^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.

Summary

- We introduced four models of contamination by outliers:
  - Huber's contamination $\mathcal{M}_n^{\mathrm{HC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Huber's deterministic contamination $\mathcal{M}_n^{\mathrm{HDC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Parameter contamination $\mathcal{M}_n^{\mathrm{PC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Adversarial contamination $\mathcal{M}_n^{\mathrm{AC}}(p, \varepsilon, \boldsymbol{\mu}^*)$.
- We have defined the worst case risks in expectation and in deviation, $R^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$ and $r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.
- We have defined the minimax risks $r^\star_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}} r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.

Summary

- We introduced four models of contamination by outliers:
  - Huber's contamination $\mathcal{M}_n^{\rm HC}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Huber's deterministic contamination $\mathcal{M}_n^{\rm HDC}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Parameter contamination $\mathcal{M}_n^{\rm PC}(p, \varepsilon, \boldsymbol{\mu}^*)$.
  - Adversarial contamination $\mathcal{M}_n^{\rm AC}(p, \varepsilon, \boldsymbol{\mu}^*)$.
- We have defined the worst case risks in expectation and in deviation, $R^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$ and $r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.
- We have defined the minimax risks $r^\star_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}} r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.
- For every $\varepsilon < 1/3 - \Box$, we have $r^\star_{n,p,\varepsilon} \asymp \frac{p}{n} + \varepsilon^2$.
- This minimax rate is obtained by Tukey's median, which is hard to compute for large p.

Summary

- We introduced four models of contamination by outliers:
  - Huber's contamination $\mathcal{M}_n^{\mathrm{HC}}(p,\varepsilon,\boldsymbol{\mu}^*)$.
  - Huber's deterministic contamination $\mathcal{M}_n^{\mathrm{HDC}}(p,\varepsilon,\boldsymbol{\mu}^*)$.
  - Parameter contamination $\mathcal{M}_n^{\mathrm{PC}}(p,\varepsilon,\boldsymbol{\mu}^*)$.
  - Adversarial contamination $\mathcal{M}_n^{\mathrm{AC}}(p,\varepsilon,\boldsymbol{\mu}^*)$.
- We have defined the worst case risks in expectation and in deviation, $R^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$ and $r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.
- We have defined the minimax risks $r^\star_{n,p,\varepsilon} = \inf_{\widehat{\boldsymbol{\mu}}} r^\star_{n,p,\varepsilon}(\widehat{\boldsymbol{\mu}})$.
- For every $\varepsilon < 1/3 - \Box$, we have $r^\star_{n,p,\varepsilon} \asymp \frac{p}{n} + \varepsilon^2$.
- This minimax rate is obtained by Tukey's median, which is hard to compute for large p.

Question

What is the smallest rate of the worst-case risk that can be obtained by an estimator computable in poly$(n,p,1/\varepsilon)$ time?

References I

M. Chen, C. Gao, and Z. Ren. Robust Covariance and Scatter Matrix Estimation under Huber's Contamination Model. ArXiv e-prints, to appear in the Annals of Statistics, 2015.

Olivier Collier and Arnak S. Dalalyan. Minimax estimation of a multidimensional linear functional in sparse gaussian models and robust estimation of the mean. submitted 1712.05495, arXiv, December 2017. URL https://arxiv.org/abs/1712.05495.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, USA, pages 655–664, 2016.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 665–674, Oct 2016.