Robust Estimation

A. Dalalyan

Sept 6-7, 2018
Moscow

I. Introduction

- In these lectures, we consider that

  Robust = Robust to the presence of outliers in the data

- We will describe several models that are used for getting a mathematical framework with contaminated data.
- We will also present several methods of robust estimation. We will pay attention to the statistical optimality and computational tractability of these methods.
- The following recent papers will be discussed:

  [1] Chen, Gao, Ren (06/15) Robust Covariance...
  [2] Lai, Rao, Vempala (04/16) Agnostic Estimation...
  [3] Diakonikolas et al. (04/16) Robust Estimation...
  [4] Collier & Dalalyan (12/17) Minimax Estimation...

II. Modeling outliers

We present now 4 models, which are of interest in robust estimation.

A) Outlier-free model: $X_1, \dots, X_n \overset{iid}{\sim} P_\mu$ on $\mathbb{R}^k$
$\mu \in M \subset \mathbb{R}^p$ is the unknown parameter

$$\mathcal{M}_{OF} = \left\{ P_\mu^{\otimes n} : \mu \in M \right\}$$

B) Huber-contamination: $X_i \overset{iid}{\sim} (1-\varepsilon)P_\mu + \varepsilon Q$.

$$\mathcal{M}_{HC}(\varepsilon) = \left\{ \left[(1-\varepsilon)P_\mu + \varepsilon Q\right]^{\otimes n} : \mu \in M,\ Q \in \mathcal{P} \right\}$$

Here Q is the distribution of the outliers, so all the outliers are assumed to have the same

An equivalent formulation is that $\exists\, Z_1,\dots,Z_n \overset{iid}{\sim} \mathcal{B}(\varepsilon)$ such that (X_i, Z_i) are iid with

$$\mathbb{P}(X_i \in A \mid Z_i = 0) = \mathbb{P}_\mu(A) \quad \mathbb{P}(X_i \in A \mid Z_i = 1) = Q(A)$$

Then, $s = \sum_{i=1}^n Z_i$ is the number of outliers.

C) Parameter contamination : We fix some $s \in \{1,\dots,n\}$.

We assume that $X_i \overset{ind}{\sim} \mathbb{P}_{\mu_i}$ so that for some $S \subset \{1,\dots,n\}$, $\mathrm{Card}(S) \leq s$, we have

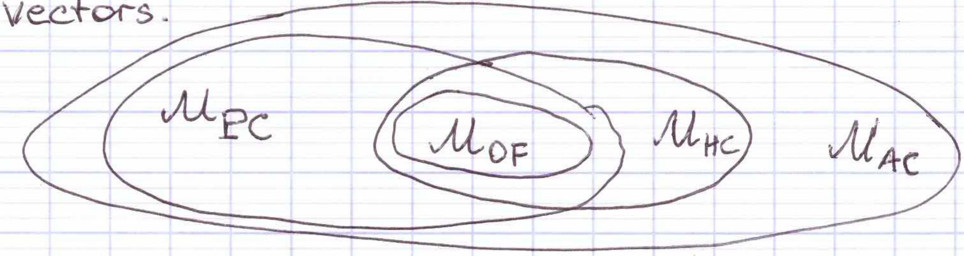
$$\mu_i = \mu \quad \forall\, i \in S^c.$$

$$\mathcal{M}_{PC}^{(s)} = \left\{ \mathbb{P}_{\mu_1} \otimes \dots \otimes \mathbb{P}_{\mu_n} : \mu_i \in M,\ \exists\, S \subset \{1,\dots,n\} \text{ s.t. } |S| \leq s \text{ and } \mu_i = \mu_j\ \forall\, i,j \notin S \right\}$$

D) Adversarial contamination: Fix $s \in \{1,\dots,n\}$.

$\mathbb{P} = \mathrm{distr.}(X_1,\dots,X_n) \in \mathcal{M}_{AC}(s)$ iff

$\exists\, S \subset \{1,\dots,n\}$, $|S| \leq s$ s.t. $\{X_i : i \notin S\} \overset{iid}{\sim} \mathbb{P}_\mu$ and $\{X_i : i \in S\}$ are arbitrary deterministic vectors.



<u>Remark</u> [1] deals with $\mathcal{M}_{HC}(\varepsilon)$

[2] deals with $\mathcal{M}_{AC}(s)$

[3] deals withe $\mathcal{M}_{AC}(s)$

[4] deals with $\mathcal{M}_{PC}(s)$

<u>Remark</u> s or ε can be known or unknown. Adaptation to unknown s or ε can be usually done by the Lepski method without much loss in statistical accuracy nor in computational complexity.

$\boxed{\text{Goal}}$ We wish to estimate μ and to quantify

$$r^{mmx}(\mathcal{M}) = \inf_{\bar{\mu}} \sup_{\mathbb{P}^{(n)} \in \mathcal{M}} \mathbb{E}\left[\|\bar{\mu} - \mu\|_2^2 \right].$$

Remark One can easily check that

$$\underbrace{r^{mmx}(\mathcal{M}_{OF})}_{\substack{\text{In regular models} \\ \text{is of order } \frac{p}{n}}} \leqslant \left[\begin{array}{l} r^{mmx}(\mathcal{M}_{PC}) \\ r^{mmx}(\mathcal{M}_{HC}(\frac{s}{n})) \end{array} \right. \leqslant r^{mmx}(\mathcal{M}_{AC})$$

Desired property: estimator $\hat{\mu}$ computable in polynomial time : polynomial in k, p, s, n etc.

$$\boxed{\begin{array}{l} \text{In what follows, we only consider the case} \\ \qquad P_\mu = \mathcal{N}_p(\mu, \sigma^2 I) \text{ with known } \sigma^2 \end{array}}$$

Disclaimer We do not perform outlier detection, which is a more difficult task. We merely look for an estimator that neglects harmful outliers.

III Summary of Chen, Gao & Ren (2015) [1]

Disclaimer We will not present ALL the results of papers [1-4], but only those that deal with the case $P_\mu = \mathcal{N}_p(\mu, \sigma^2 I)$.

[1] deals with the Huber contamination model

$$X_i \sim (1-\varepsilon)\mathcal{N}(\mu, \sigma^2 I) + \varepsilon Q.$$

Natural estimators of μ are the mean and the median. One can easily check that

Med$_n$ is the coordinatewise median of $X_1, \dots, X_n$

$$R(\bar{X}_n, \mathcal{M}_{OF}) = \frac{\sigma^2 p}{n} \qquad R(\bar{X}_n, \mathcal{M}_{HC}) = +\infty$$

$$R(\mathrm{Med}_n, \mathcal{M}_{OF}) \asymp \frac{\sigma^2 p}{n}$$

THEOREM 1. For every $\varepsilon \in (0,1)$, we have

$$R(\mathrm{Med}_n, \mathcal{M}_{HC}) \asymp \frac{\sigma^2 p}{n} + \sigma^2 \varepsilon^2 p$$

Question Is the order $\varepsilon^2 p$ optimal in minimax sense?

THEOREM 2 There are constants $c_0 \in [0,1]$ and $c_1 > 0$ such that for every $\varepsilon \leq c_0$, we have

$$r^{mmx}(\mathcal{M}_{HC}) \geq c_1 \sigma^2 \left(\frac{p}{n} + \varepsilon^2\right).$$

Comments

1) If the dimension is fixed when the sample size increases or the contamination rate decreases, that is $p = O(1)$, the sample median is mmx rate optimal. In addition, it is computationally tractable (probably the "cheapest" robust estim.).
2) When $p = p_n \xrightarrow[n\to\infty]{} +\infty$ or $p = p_\varepsilon \xrightarrow[\varepsilon\to 0]{} +\infty$, then there is a gap of order p between the lower bound of Thm 2 and the upper bound of Thm 1. Which one gives the optimal rate of the mmx risk?

THEOREM 3. There is an estimator $\hat{\mu}_n$, termed Tukey's median, satisfying the following property. There are constants $\bar{c}_0 \in [0,1]$ and $\bar{c}_1 > 0$ such that for every $\varepsilon \leq \bar{c}_0$ we have

$$R(\hat{\mu}_n, \mathcal{M}_{HC}(\varepsilon)) \leq \bar{c}_1 \sigma^2 \left(\frac{p}{n} + \varepsilon^2\right).$$

Thus, $\hat{\mu}$ is minimax-rate-optimal.

IV More details on [1] (if I have time)

In this section we give more details on Thm. 2 & 3. We start by defining $\hat{\mu}_n$, Tukey's depth, then we present a sketch of the proof of Thm. 2.

DEF Let $X_1, \dots, X_n$ be a sample from $\mathbb{R}^p$. Let $x_0 \in \mathbb{R}^p$ be any point. We call Tukey's depth of x_0 w.r.t. the sample $\{X_i\}$ the quantity

$$\mathcal{D}_n(x_0) = \inf_{u \in S^1} \sum_{i=1}^{n} \mathbb{1}(u^T X_i \leq u^T x_0)$$

We call Tukey's median the deepest point in $\mathbb{R}^p$:

We see that $\hat{\mu}_n$ is defined as the saddle point of a non-smooth non-concave-convex problem. Computing $\hat{\mu}_n$ is NP-hard. (I have never seen a formal proof of this claim, but it seems quite plausible).

Question What is the best rate that can be atteined by a poly-time algorithm?

Proof of Thm 2

Let us consider, w.l.o.g., that $\sigma = 1$ and set $P_1 = \mathcal{N}(0, I)$ and $P_2 = \mathcal{N}(\mu^*, I)$ with μ^* satisfying $\|\mu^*\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}$.

Then, we have

$$TV(P_1, P_2) \leq \frac{1}{\sqrt{2}} \sqrt{D_{KL}(P_1 \| P_2)} = \frac{\|\mu^*\|_2}{2} \leq \frac{\varepsilon}{1-\varepsilon}$$

Let $\varepsilon_1 \leq \varepsilon$ be such that

$$TV(P_1, P_2) = \varepsilon_1.$$

Define Q_1 and Q_2 by densities

$$q_1 = \left(1 - \frac{\varepsilon_1}{\varepsilon}\right) f_1 + \frac{\varepsilon_1}{\varepsilon} \times \frac{(f_2 - f_1)_+}{TV(P_1, P_2)}$$

$$q_2 = \left(1 - \frac{\varepsilon_1}{\varepsilon}\right) f_2 + \frac{\varepsilon_1}{\varepsilon} \times \frac{(f_1 - f_2)_+}{TV(P_1, P_2)}$$

One easily checks that q_1 and q_2 are densities and that

$$(1-\varepsilon) f_1 + \varepsilon q_1 = (1-\varepsilon) f_2 + \varepsilon q_2$$

Thus, the the distributions of two samples corresponding to parameters (μ_1, Q_1) and (μ_2, Q_2) are equal. This means that the mmx rate of estimation is at least $\|\mu_1 - \mu_2\|_2^2 = \frac{4\varepsilon^2}{(1-\varepsilon)^2} \geq 4\varepsilon^2$ ∎

THEOREM 2 There are constants $C_0 \in [0,1]$ and $C_1 > 0$ such that for every $\varepsilon \leq C_0$, we have

$$r^{mmx}(\mathcal{M}_{HC}) \geq C_1 \sigma^2 \left(\frac{p}{n} + \varepsilon^2\right).$$

Comments

1) If the dimension is fixed when the sample size increases or the contamination rate decreases, that is $p = O(1)$, the sample median is mmx rate optimal. In addition, it is computationally tractable (probably the "cheapest" robust estim.).
2) When $p = p_n \xrightarrow[n\to\infty]{} +\infty$ or $p = p_\varepsilon \xrightarrow[\varepsilon\to 0]{} +\infty$, then there is a gap of order p between the lower bound of Thm 2 and the upper bound of Thm 1. Which one gives the optimal rate of the mmx risk?

THEOREM 3. There is an estimator $\hat{\mu}_n$, termed Tukey's median, satisfying the following property. There are constants $\bar{C}_0 \in [0,1]$ and $\bar{C}_1 > 0$ such that for every $\varepsilon \leq \bar{C}_0$ we have

$$R(\hat{\mu}_n, \mathcal{M}_{HC}(\varepsilon)) \leq \bar{C}_1 \sigma^2 \left(\frac{p}{n} + \varepsilon^2\right).$$

Thus, $\hat{\mu}$ is minimax-rate-optimal.

IV More details on [1] (if I have time)

In this section we give more details on Thm. 2 & 3. We start by defining $\hat{\mu}_n$, Tukey's depth, then we present a sketch of the proof of Thm. 2.

DEF Let $X_1, \dots, X_n$ be a sample from $\mathbb{R}^p$. Let $x_0 \in \mathbb{R}^p$ be any point. We call Tukey's depth of x_0 w.r.t. the sample $\{X_i\}$ the quantity

$$\mathcal{D}_n(x_0) = \inf_{u \in S^1} \sum_{i=1}^n \mathbb{1}(u^T X_i \leq u^T x_0)$$

We call Tukey's median the deepest point in $\mathbb{R}^p$:

$$\hat{\mu}_n = \arg\max_{x^0 \in \mathbb{R}^p} \mathcal{D}_n(x_0)$$

An equivalent formulation is that $\exists\, Z_1,\dots,Z_n \overset{iid}{\sim} \mathcal{B}(\varepsilon)$ such that (X_i, Z_i) are iid with

$$\mathbb{P}(X_i \in A \mid Z_i = 0) = P_\mu(A) \quad \mathbb{P}(X_i \in A \mid Z_i = 1) = Q(A)$$

Then, $s = \sum_{i=1}^{n} Z_i$ is the number of outliers.

C) Parameter contamination : We fix some $s \in \{1,\dots,n\}$.

We assume that $X_i \overset{ind}{\sim} P_{\mu_i}$ so that for some $S \subset \{1,\dots,n\}$, $\mathrm{Card}(S) \leqslant s$, we have

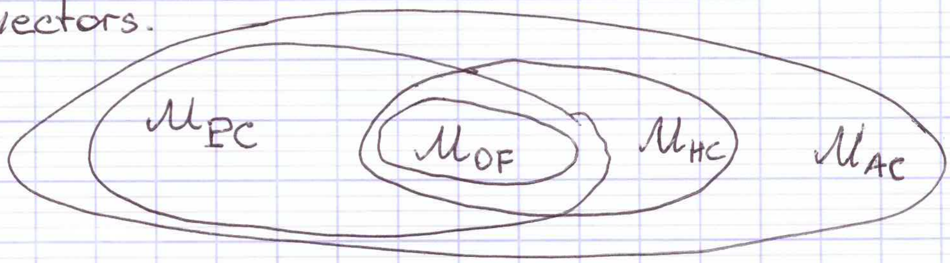
$$\mu_i = \mu \quad \forall\, i \in S^c.$$

$$\mathcal{M}_{PC}^{(s)} = \{ P_{\mu_1} \otimes \dots \otimes P_{\mu_n} : \mu_i \in M,\ \exists\, S \subset \{1,\dots,n\} \text{ s.t. } |S| \leqslant s \text{ and } \mu_i = \mu_j \ \forall\, i,j \notin S \}$$

D) Adversarial contamination: Fix $s \in \{1,\dots,n\}$.

$P = \mathrm{distr.}(X_1,\dots,X_n) \in \mathcal{M}_{AC}(s)$ iff

$\exists\, S \subset \{1,\dots,n\}$, $|S| \leqslant s$ s.t. $\{X_i : i \notin S\} \overset{iid}{\sim} P_\mu$ and $\{X_i : i \in S\}$ are arbitrary deterministic vectors.



Remark [1] deals with $\mathcal{M}_{HC}(\varepsilon)$
[2] deals with $\mathcal{M}_{AC}(s)$
[3] deals withe $\mathcal{M}_{AC}(s)$
[4] deals with $\mathcal{M}_{PC}(s)$

Remark s or ε can be known or unknown. Adaptation to unknown s or ε can be usually done by the Lepski method without much loss in statistical accuracy nor in computational complexity.

$\boxed{\text{Goal}}$ We wish to estimate μ and to quantify

$$r^{mmx}(\mathcal{M}) = \inf_{\bar{\mu}} \underbrace{\sup_{P^{(n)} \in \mathcal{M}} \mathbb{E}\left[\|\bar{\mu} - \mu\|_2^2\right]}_{R(\mathcal{M}, \bar{\mu})}.$$