Confidence regions and minimax rates in outlier-robust estimation on the probability simplex

Amir-Hossein Bateni

amirhossein.bateni@ensae.fr
ENSAE, IP Paris

Introduction

Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n i.i.d. random variables taking their values in the k-dimensional probability simplex

$$\Delta^{k-1} = \{\boldsymbol{v} \in \mathbb{R}_+^k : v_1 + \ldots + v_k = 1\}.$$

Our goal is to estimate the unknown vector

$$\boldsymbol{\theta} = \mathbf{E}[\boldsymbol{X}_i]$$

in the case where the observations are contaminated by outliers.
Particular case: The distribution of a discrete random variable $\boldsymbol{X}$ taking k distinct values ($\boldsymbol{X}_i$'s take values in $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$).

Minimax risk

We study the estimation error under three different metrics: total-variation, Hellinger and $\mathbb{L}^2$ distances

$$d_{\mathsf{TV}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := {}^1\!/_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1,$$
$$d_{\mathsf{H}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := {}^1\!/_{\sqrt{2}} \|\widehat{\boldsymbol{\theta}}^{1/2} - \boldsymbol{\theta}^{1/2}\|_2,$$
$$d_{\mathbb{L}^2}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2,$$

and we evaluate the minimax risk

$$\mathfrak{R}_\square(n, k, s, \varepsilon) := \inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{P}, \boldsymbol{Q}} \mathbf{E}[d_\square(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta})],$$

where the *inf* is over all estimators $\bar{\boldsymbol{\theta}}_n$ built upon the contaminated observations and the *sup* is over all distributions $\boldsymbol{P}$, $\boldsymbol{Q}$ on the probability simplex such that the mean $\boldsymbol{\theta}$ of $\boldsymbol{P}$ is s-sparse. The subscript $\square$ of $\mathfrak{R}$ above refers to the distance used in the risk, so that $\square$ is TV, H, or $\mathbb{L}^2$.

Various models of contamination

Huber's contamination $\big(\mathcal{M}_n^{\mathsf{HC}}(\varepsilon, \boldsymbol{\theta}^*)\big)$: The observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and drawn from the mixture distribution

$$\boldsymbol{P}_{\varepsilon, \boldsymbol{\theta}^*, \boldsymbol{Q}} := (1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}^*} + \varepsilon \boldsymbol{Q},$$

where $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ represents the reference distribution parametrized by $\boldsymbol{\theta}^*$, $\boldsymbol{Q}$ is the distribution of the outliers and ε is the fraction of the outliers.
Huber's deterministic contamination $\big(\mathcal{M}_n^{\mathsf{HDC}}(\varepsilon, \boldsymbol{\theta}^*)\big)$: The number of outliers is not random, and there are at most $n\varepsilon$ outliers following the contamination distribution.
Oblivious contamination $\big(\mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \boldsymbol{\theta}^*)\big)$: The outliers are not necessarily i.i.d. and the number and the joint distribution of outliers is determined in advance, possibly based on the knowledge of the reference distribution.
Parameter contamination $\big(\mathcal{M}_n^{\mathsf{OC}}(\varepsilon, \boldsymbol{\theta}^*)\big)$: The parameters of the distributions of some observations are contaminated. Each outlier $\boldsymbol{X}_i$ is drawn from a distribution $\boldsymbol{Q}_i = \boldsymbol{P}_{\boldsymbol{\theta}_i}$ belonging to the same family as the reference distribution, but corresponding to a contaminated parameter $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}^*$.
Adversarial contamination $\big(\mathcal{M}_n^{\mathsf{AC}}(\varepsilon, \boldsymbol{\theta}^*)\big)$: An adversary sees all the initial observations drawn from the reference distribution and replaces $n\varepsilon$ of them by arbitrary values.

Lower and upper bounds

To ease notation, we let $R_d^\square(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}})$ to be the worst-case risk of an estimator $\widehat{\boldsymbol{\theta}}$, where $\square$ is either HC or AC. More precisely, for $\mathcal{M}_n^\square(\varepsilon, \Theta) := \cup_{\boldsymbol{\theta} \in \Theta} \mathcal{M}_n^\square(\varepsilon, \boldsymbol{\theta})$, we set

$$R_d^\square(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}) := \sup_{P_n \in \mathcal{M}_n^\square(\varepsilon, \Theta)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)].$$

We denote by Δ_s^{k-1} the set of all $\boldsymbol{v} \in \Delta^{k-1}$ having at most s non-zero entries.

Theorem 1. *For every triple of positive integers (k, s, n) and for every $\varepsilon \in [0, 1]$, the sample mean $\bar{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ satisfies*

$$R_{TV}^{AC}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + 2\varepsilon,$$
$$R_{H}^{AC}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + \sqrt{2}\,\varepsilon^{1/2},$$
$$R_{\mathbb{L}^2}^{AC}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon.$$

Theorem 2. *There are universal constants $c > 0$ and n_0, such that for any integers $k \ge 3$, $s \le k \wedge n$, $n \ge n_0$ and for any $\varepsilon \in [0, 1]$, we have*

$$\inf_{\bar{\boldsymbol{\theta}}_n} R_{TV}^{HC}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \ge c\{(s/n)^{1/2} + \varepsilon\},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R_{H}^{HC}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \ge c\{(s/n)^{1/2} + \varepsilon^{1/2}\},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R_{\mathbb{L}^2}^{HC}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{\theta}}_n) \ge c\{(1/n)^{1/2} + \varepsilon\},$$

where $\inf_{\bar{\boldsymbol{\theta}}_n}$ stands for the infimum over all measurable functions $\bar{\boldsymbol{\theta}}_n$ from $(\Delta^{k-1})^n$ to Δ^{k-1}.

Thus, the rate obtained for the sample mean is minimax optimal.

Instance based bounds

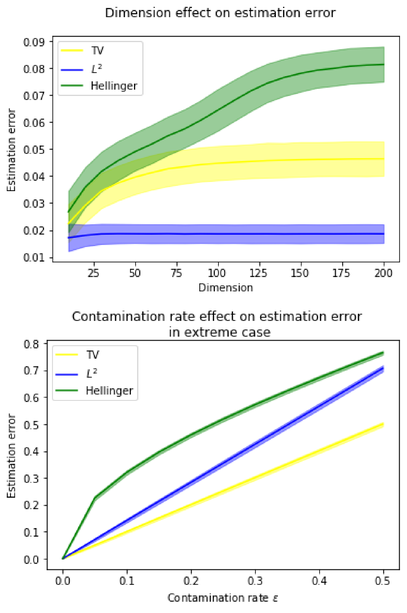
Theorem 4. *Suppose $\boldsymbol{X}_i$'s take value in $\{\boldsymbol{e}_1, \boldsymbol{e}_2 \ldots\}$, and for $j \in \mathbb{N}$, $\boldsymbol{e}_j$ occurs with probability $\boldsymbol{\theta}_j^*$. For every n and for every $\varepsilon \in [0, 1]$, the sample mean $\bar{\boldsymbol{X}}_n$ satisfies*

$$d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \le \frac{1}{\sqrt{n}} \bar{\boldsymbol{X}}_n^{1/2} + 2\varepsilon + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

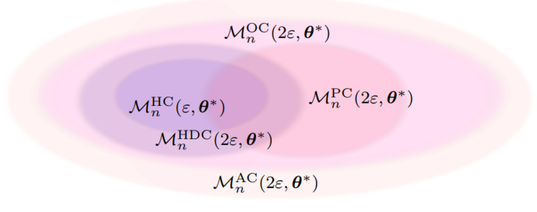
with probability at least $1 - \delta$, where $\delta \in (0, 1)$. We also have

$$\mathbf{E} d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) \le \frac{1}{\sqrt{n}} \mathbf{E} \bar{\boldsymbol{X}}_n^{1/2} + 2\varepsilon.$$

Illustration on a numerical example



Hierarchy between models



Confidence regions

Theorem 3. *Let $\delta \in (0, 1)$ be the tolerance level. If $\boldsymbol{\theta}^* \in \Delta_s^{k-1}$, then under any contamination model, the regions of Δ^{k-1} defined by each of the following inequalities*

$$d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon + \big(\log(1/\delta)/n\big)^{1/2},$$
$$d_{\mathsf{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \le (s/n)^{1/2} + 2\,\varepsilon + \big(2\log(1/\delta)/n\big)^{1/2},$$
$$d_{\mathsf{H}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \le \sqrt{5}\big((s/n)\log(2s/\delta)\big)^{1/2} + \varepsilon^{1/2} + \big((1/2n)\log(2/\delta)\big)^{1/2},$$

contain $\boldsymbol{\theta}^*$ with probability at least $1 - \delta$.

Conclusion

The choice of the distance has much stronger impact on the risk rate than the nature of contamination.

References

Bateni, A.-H. and Dalalyan, A. S. (2020). Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. Electron. J. Statist., 14(2):2653–2677.