Excess-Risk consistency of group-hard thresholding estimator in Robust Estimation of Gaussian Mean

Arshak Minasyan[*1]

[1] *Department of Mathematics and Mechanics*
Yerevan State University, YerevaNN Research Lab

February 23, 2022

In this work we introduce the notion of the excess risk in the setup of estimation of the Gaussian mean when the observations are corrupted by outliers. It is known that the sample mean loses its good properties in the presence of outliers (Huber, 1964, 1965). In addition, even the sample median is not minimax-rate-optimal in the multivariate setting. The optimal rate of the minimax risk in this setting was established by Chen et al. (2018). However, even these minimax-rate-optimality results do not quantify how fast the risk in the contaminated model approaches the risk in the uncontaminated model when the rate of contamination goes to zero. The present paper does a first step in filling this gap by showing that the group hard thresholding estimator has an excess risk that goes to zero when the corruption rate approaches zero.

Keywords: Robust estimation, Minimax estimation, Excess risk.

AMS subject classification: 62F35.

1 Introduction

In recent years, we witnessed a revival of interest in statistical methods that can efficiently deal with data sets corrupted by outliers. In particular, under the Huber contamination model in the problem of Gaussian mean estimation,

[*] minasyan@yerevann.com

Chen et al. (2018) established the minimax rate and showed that it is attained by Tukey's median. Furthermore, Chen et al. (2016) developed a general theory for obtaining the minimax rate (both upper and lower bounds) in a wide class of statistical models. These works are focused on statistical complexity of the estimators, without paying attention to the computational complexity. The latter has been addressed by Collier and Dalalyan (2019) and Cheng et al. (2018), who analyzed the risk of computationally tractable estimators. Interestingly, the results proved in these papers only provide the order of magnitude of the minimax risk and do not tell anything about how fast the risk in the corrupted setting get close to the risk in the uncorrupted setting.

In this paper, we introduce the notion of the excess risk, which is defined as the difference between the risks in the corrupted and uncorrupted settings. Then, we present an analysis of this risk for a procedure that we call group hard thresholding estimator. It can also be seen as a version of the trimmed mean estimator. Our main result shows that this excess risk goes to zero, as the rate of contamination goes to zero.

To be more precise, let us assume that we observe n random vectors $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ in $\mathbb{R}^p$, which are assumed to satisfy

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\theta}_i + \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_p). \tag{1}$$

In the above formula, $\boldsymbol{\mu}$ is the unknown mean we wish to estimate, $\{\boldsymbol{\theta}_i\}$ are arbitrary deterministic vectors measuring the outlyingness of each data point and $\boldsymbol{\xi}_i$ are random errors. In this paper, we assume that $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_n]$ is a column-wise sparse matrix. All the observations with indices $i \in \mathcal{O} = \{\ell : \|\boldsymbol{\theta}_\ell\|_2 > 0\}$ are considered as outliers, while all the other are called inliers. In the sequel, we use notation

$$o = \mathrm{Card}(\mathcal{O}), \quad \text{and} \quad \varepsilon = \frac{o}{n}.$$

The parameter ε, assumed to be strictly smaller than $1/2$, plays an important role in robust estimation. In particular, it is known that the minimax rate of estimation in model (1) is of order $\frac{p}{n} + \varepsilon^2$.

In this paper, we propose to consider a more precise measure of accuracy of an estimator, the excess risk. Recall that the risk of an estimator[1] $\widehat{\boldsymbol{\mu}}_n$ is given by

$$R[\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}; \boldsymbol{\Theta}] = [\mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Theta}} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2]^{1/2}.$$

[1] An estimator is any measurable function from $(\mathbb{R}^p)^n$ to $\mathbb{R}^p$

In the above formula and in the sequel, the notation $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Theta}}[h]$ stands for the expectation with respect to the distribution of $\{\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n\}$ as defined by Eq. (1) (it is implicitly assumed that the function h depends on the observations $\{\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n\}$). It is a well-known fact that in the outlier-free setup, where $\boldsymbol{\Theta} \equiv \mathbf{0}_{p\times n}$ the minimax risk satisfies

$$\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu}\in\mathbb{R}^p} R[\widehat{\boldsymbol{\mu}},\boldsymbol{\mu};\mathbf{0}] = \sup_{\boldsymbol{\mu}\in\mathbb{R}^p} R[\overline{\boldsymbol{Y}}_n,\boldsymbol{\mu};\mathbf{0}] = \sqrt{\frac{p}{n}}, \tag{2}$$

where $\overline{\boldsymbol{Y}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_i$ is the sample mean of the observed vectors. Let us define the mixed matrix norm

$$\|\boldsymbol{\Theta}\|_{0,2} = \sum_{i=1}^n \mathbb{1}(\|\boldsymbol{\theta}_i\|_2 > 0).$$

Based on (2), we define the worst-case excess risk of an estimator $\widehat{\boldsymbol{\mu}}$ by

$$\mathcal{E}(\widehat{\boldsymbol{\mu}}; n, p, \varepsilon) = \sup_{\boldsymbol{\mu}\in\mathbb{R}^p; \|\boldsymbol{\Theta}\|_{0,2}\le\varepsilon n} R[\widehat{\boldsymbol{\mu}},\boldsymbol{\mu};\boldsymbol{\Theta}] - \sqrt{\frac{p}{n}}$$

as well as the minimax excess risk

$$\mathcal{E}(n,p,\varepsilon) = \inf_{\widehat{\boldsymbol{\mu}}} \mathcal{E}(\widehat{\boldsymbol{\mu}}, n, p, \varepsilon),$$

where the infimum is over all possible estimators $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$. Note that according to the definition, the estimators considered in the above formula can depend on n, p and $\varepsilon = o/n$. The main result of this paper shows that the excess risk of the group hard thresholding estimator, introduced in the next section, tends to zero as $\varepsilon = \varepsilon_n \to 0$ as $n \to \infty$, as soon as $p = p_n$ is such that p_n/n is bounded by a constant.

2 Group Hard Thresholding

In this section we define the estimator $\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}$, called group hard thresholding estimator, and prove that this estimator has an excess risk that vanishes when the proportion of contamination ε tends to zero. Roughly speaking, $\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}$ is the arithmetic mean of a sample obtained from $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$ by replacing all the vectors that are at a large distance from the coordinatewise median by the latter.

More specifically, let $\widehat{\boldsymbol{\mu}}_{\mathrm{Med}} := \mathtt{Med}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n)$ be the coordinatewise median of the sample $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$. Let us fix a positive threshold $\lambda > 0$, which will be a tuning parameter of the method. For each $i \in \{1, \dots, n\}$, we put

$$\widehat{\boldsymbol{\theta}}_i = HT_\lambda(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}) := (\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}})\mathbb{1}(\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}\|_2 > \lambda) \tag{3}$$

$$\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\theta}}_i) := L_n(\boldsymbol{Y} - \widehat{\boldsymbol{\Theta}}). \tag{4}$$

Next, we formulate the main theorem of the paper showing that the excess risk of the GHT estimator tends to zero if the proportion of outliers $\varepsilon = \varepsilon_n$ tends to 0 fast enough so that $\varepsilon_n p_n^{1/4}$ goes to zero. Notice that this condition holds when p is fixed, however this setup allows the infinite dimensional case, i.e. $p = p_n \to \infty$ under the constraint $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$ as the sample size n goes to infinity.

Theorem 1. *For* $\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}$ *defined in* (4) *and* $\lambda^2 = p + 8\sqrt{p \log \varepsilon^{-1}} + 16 \log \varepsilon^{-1}$ *we have*

$$\overline{\lim_{n\to\infty}} \, \mathcal{E}(\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}}, n, p_n, \varepsilon_n) = 0$$

provided that $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$ *and* $p_n = O(n)$ *as* $n \to \infty$.

Proof. Let $s_i = \mathbb{1}(\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}\|_2 \le \lambda)$ and $\boldsymbol{\delta} = \widehat{\boldsymbol{\mu}}_{\mathrm{Med}} - \boldsymbol{\mu}^*$. Using the fact that $\boldsymbol{Y} = \boldsymbol{\mu}\mathbf{1}_n^\top + \boldsymbol{\Theta} + \boldsymbol{\Xi}$, we can write

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_{\mathrm{GHT}} - \boldsymbol{\mu} &= L_n(\boldsymbol{\Theta} + \boldsymbol{\Xi} - \widehat{\boldsymbol{\Theta}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i + T_1(n) + T_2(n) + T_3(n),\end{aligned} \tag{5}$$

where

$$T_1(n) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i(s_i - 1), \quad T_2(n) = \frac{\boldsymbol{\delta}}{n}\sum_{i=1}^{n}(1 - s_i),$$

$$T_3(n) = \frac{1}{n}\sum_{i\in\mathcal{O}}\boldsymbol{\theta}_i s_i.$$

For the ease of notation let us define the $\mathbb{L}_2$ norm of vector $\boldsymbol{V}$ as follows

$$\|\boldsymbol{V}\|_{\mathbb{L}_2} = \left(\mathbb{E}[\|\boldsymbol{V}\|_2^2]\right)^{1/2}.$$

Notice that it suffices to show that $T_i(n) \to 0$ whenever $\varepsilon \to 0$ for $i \in \{1,2,3\}$. Indeed,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \right\|_{\mathbb{L}_2}^2 = \frac{p}{n}.$$

Hence,

$$\begin{aligned} \mathcal{E}(\widehat{\boldsymbol{\mu}}; n, p, \varepsilon) &\le \|T_1(n) + T_2(n) + T_3(n)\|_{\mathbb{L}_2} \\ &\le \|T_1(n)\|_{\mathbb{L}_2} + \|T_2(n)\|_{\mathbb{L}_2} + \|T_3(n)\|_{\mathbb{L}_2}. \end{aligned}$$

One can check that for some constant C, it holds

$$\|\boldsymbol{\delta}\|_{\mathbb{L}_4} \le C\sqrt{p}\left(\frac{1}{\sqrt{n}} \vee \varepsilon\right) = o(p^{1/4}) \tag{6}$$

and

$$\|\boldsymbol{\xi}_i\|_{\mathbb{L}_4}^4 = \mathbb{E}\left(\sum_{j=1}^{p} \xi_{ij}^2\right)^2 = 3p + p(p-1) \le (p+1)^2. \tag{7}$$

We first bound $\mathbb{E}[1 - s_i]$ for all the inliers $i \in \mathcal{O}^c$. In view of (6), we have $\|\boldsymbol{\delta}\|_2 = o_{\mathbb{P}}\big(p^{1/4}\big)$ and, therefore,

$$\begin{aligned} \mathbb{E}[1 - s_i] &= \mathbb{P}(s_i = 0) = \mathbb{P}(\|\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{Med}}\|_2^2 > \lambda^2) = \mathbb{P}(\|\boldsymbol{\delta} + \boldsymbol{\xi}_i\|_2^2 > \lambda^2) \\ &\le \mathbb{P}\big(\|\boldsymbol{\xi}_i\|_2^2 > \lambda^2(1 - o(1))\big) \lesssim \varepsilon^8, \qquad \forall i \in \mathcal{O}^c, \end{aligned} \tag{8}$$

where the last inequality follows from χ_p^2 concentration bound due to the choice of λ^2. For $T_1(n)$, in view of the last display and (7), we have

$$\begin{aligned} \|T_1(n)\|_{\mathbb{L}_2} &\le \left\| \frac{1}{n} \sum_{i \in \mathcal{O}^c} \boldsymbol{\xi}_i (1 - s_i) \right\|_{\mathbb{L}_2} + \left\| \frac{1}{n} \sum_{i \in \mathcal{O}} \boldsymbol{\xi}_i (1 - s_i) \right\|_{\mathbb{L}_2} \\ &= O(\sqrt{p}\varepsilon^2) + \left\| \frac{1}{n} \sum_{i \in \mathcal{O}} \boldsymbol{\xi}_i (1 - s_i) \right\|_{\mathbb{L}_2}. \end{aligned}$$

On the other hand, the Cauchy-Schwarz inequality yields

$$\left\| \sum_{i \in \mathcal{O}} \boldsymbol{\xi}_i (1 - s_i) \right\|_2^2 \le \sum_{i \in \mathcal{O}} (1 - s_i) \left\| \sum_{i \in \mathcal{O}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\|_{\mathrm{op}} \le n\varepsilon \left\| \sum_{i \in \mathcal{O}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \right\|_{\mathrm{op}},$$

where the last norm corresponds to the operator norm (*i.e.*, the largest singular value). The well-known bound for the operator norm of the Gaussian matrix (see, for instance, Lemma 9 in Collier and Dalalyan (2019)) implies that

$$\Big\| \sum_{i\in\mathcal{O}} \boldsymbol{\xi}_i(1-s_i) \Big\|_{\mathbb{L}_2}^2 \le 3n\varepsilon(p+n\varepsilon+4),$$

This leads to

$$\|T_1(n)\|_{\mathbb{L}_2} \lesssim \sqrt{p}\,\varepsilon^2 + \frac{\sqrt{n\varepsilon p}+n\varepsilon}{n} = \sqrt{p}\,\varepsilon^2 + \sqrt{\varepsilon p/n} + \varepsilon = o(1). \tag{9}$$

For bounding $\|T_2(n)\|_{\mathbb{L}_2}$, we use the Cauchy-Schwarz inequality, in conjunction with (6) and (8), to get

$$\begin{aligned}\|T_2(n)\|_{\mathbb{L}_2} &= \Big\| \frac{\boldsymbol{\delta}}{n} \sum_{i=1}^n (1-s_i) \Big\|_{\mathbb{L}_2} \le \frac{\|\boldsymbol{\delta}\|_{\mathbb{L}_4}}{n} \cdot \sum_{i=1}^n \|1-s_i\|_{\mathbb{L}_4} \\ &= \frac{o(p^{1/4})}{n}(\varepsilon n + n(1-\varepsilon)\varepsilon^2) = \varepsilon o(p^{1/4}) = o(1), \end{aligned} \tag{10}$$

where in the third step we bounded $\|1-s_i\|_{\mathbb{L}_4} = \mathbb{E}^{1/4}[1-s_i]$ by 1 for $i \in \mathcal{O}$ and by ε^2 for $i \in \mathcal{O}^c$.

To bound the $\mathbb{L}_2$ norm of $T_3(n)$ we notice that whenever $s_i = 1$ we have an upper bound on $\|\boldsymbol{\theta}_i\|_2$. More precisely, one can check that the inequality $\|\boldsymbol{\theta}_i + \boldsymbol{\delta} + \boldsymbol{\xi}_i\|_2 \le \lambda$, equivalent to $s_i = 1$, implies that

$$\|\boldsymbol{\theta}_i\|_2 \le 2\|\boldsymbol{\delta}\|_2 + 2|\eta_i| + \big(\lambda^2 - \|\boldsymbol{\xi}_i\|_2^2\big)_+^{1/2} + (2|\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|)^{1/2}, \tag{11}$$

for standard Gaussian variables η_i for $i \in \{1,\dots,n\}$.

Using the Hölder inequality, one can show that

$$\begin{aligned}\sum_{i\in\mathcal{O}} |\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|^{1/2} &\le (n\varepsilon)^{3/4} \Big\{ \sum_{i\in\mathcal{O}} |\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|^2 \Big\}^{1/4} = (n\varepsilon)^{3/4} \Big\{ \boldsymbol{\delta}^\top \sum_{i\in\mathcal{O}} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top \boldsymbol{\delta} \Big\}^{1/4} \\ &\le (n\varepsilon)^{3/4} \|\boldsymbol{\delta}\|_2^{1/2} \|\boldsymbol{\xi}_O\|_{\mathrm{op}}^{1/2}, \end{aligned}$$

where $\|\boldsymbol{\xi}_O\|_{\mathrm{op}}^{1/2}$ is the spectral norm of the matrix obtained by concatenating

the vectors $\boldsymbol{\xi}_i$ for $i \in \mathcal{O}$. This implies that[2]

$$\begin{aligned}\Big\| \sum_{i\in\mathcal{O}} |\boldsymbol{\delta}^\top \boldsymbol{\xi}_i|^{1/2} \Big\|_{\mathbb{L}_2} &\le (n\varepsilon)^{3/4} \|\boldsymbol{\delta}\|_{\mathbb{L}_2}^{1/2} \|\boldsymbol{\xi}_O\|_{\mathbb{L}_2}^{1/2} \\ &= O\Big((n\varepsilon)^{3/4} (\varepsilon^{1/2} p^{1/4})((n\varepsilon)^{1/4} + p^{1/4})\Big) \\ &= O\Big(n\varepsilon p^{1/4} + (n\varepsilon)^{3/4}\sqrt{\varepsilon p}\Big).\end{aligned}$$

One can also show that

$$\mathbb{E}[(\lambda^2 - \|\boldsymbol{\xi}_i\|_2^2)_+]) \lesssim \sqrt{p \log \varepsilon^{-1}} + \log \varepsilon^{-1} \lesssim \sqrt{p}\, \log \varepsilon^{-1}$$

Using the triangle inequality for the $\mathbb{L}_2$-norm, we arrive at

$$\begin{aligned}\|T_3(n)\|_{\mathbb{L}_2} &\le \frac{1}{n} \sum_{i\in\mathcal{O}} \|\boldsymbol{\theta}_i \mathbb{1}(s_i = 1)\|_{\mathbb{L}_2} \\ &\lesssim \varepsilon(\|\boldsymbol{\delta}\|_{\mathbb{L}_2} + 1) + \varepsilon p^{1/4} \log^{1/2} \varepsilon^{-1} + \varepsilon p^{1/4} + \varepsilon^{5/4} p^{1/2} n^{-1/4} \\ &\lesssim \varepsilon p^{1/4} \log^{1/2} \varepsilon^{-1} + \varepsilon p^{1/4} + \varepsilon^{5/4} p^{1/4}.\end{aligned}$$

Since $\varepsilon_n = o(1)$ and $\varepsilon_n p_n^{1/4} \log^{1/2} \varepsilon_n^{-1} = o(1)$, we get that $\|T_3(n)\|_{\mathbb{L}_2} = o(1)$, concluding the proof of the theorem.

□

Acknowledgments. The author thanks Prof. Arnak Dalalyan for introduction to the topic, ongoing feedback and final revision of this work.

References

M. Chen, C. Gao, and Z. Ren. A general decision theory for huber's ε-contamination model. *Electronic Journal of Statistics, Vol. 10, 3752 – 3774*, 2016.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *Annals of Statistics, Volume 46, Number 5, 1932-1960*, 2018.

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. *arXiv:1811.09380*, 2018.

[2]For the sake of simplicity, we consider the case $n^{-1/2} \lesssim \varepsilon_n$.

Olivier Collier and Arnak S. Dalalyan. Rate-optimal estimation of p-dimensional linear functionals in a sparse gaussian model. *Electron. J. Statist.*, 13(2):2830–2864, 2019. URL https://projecteuclid.org/download/pdfview_1/euclid.ejs/1567065621.

P. J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics, 35(1) : 73–101*, 1964.

P. J. Huber. A robust version of the probability ratio test. *The annals of mathematical statistics, 36(6) :1 753–1758*, 1965.