All-In-One Robust Estimator of the Gaussian Mean

Arnak S. Dalalyan* Arshak Minasyan†

Abstract

The goal of this paper is to show that a single robust estimator of the mean of a multivariate Gaussian distribution can enjoy five desirable properties. First, it is computationally tractable in the sense that it can be computed in a time which is at most polynomial in dimension, sample size and the logarithm of the inverse of the contamination rate. Second, it is equivariant by translations and orthogonal transformations. Third, it has a high breakdown point equal to 0.5, and a nearly-minimax-rate-breakdown point approximately equal to 0.28. Fourth, it is minimax rate optimal, up to a logarithmic factor, when data consist of independent observations corrupted by adversarially chosen outliers. Fifth, it is asymptotically optimal when the rate of contamination tends to zero. The estimator is obtained by an iterative reweighting approach. Each sample point is assigned a weight that is iteratively updated using a convex optimization problem. We also establish a dimension-free non-asymptotic risk bound for the expected error of the proposed estimator. It is the first of this kind results in the literature and involves only the effective rank of the covariance matrix.

Key Words: robust estimation, breakdown point, minimax rate, computational tractability

1. Introduction

Robust estimation is one of the most fundamental problems in statistics. Its goal is to design efficient procedures capable of processing data sets contaminated by outliers, so that these outliers have little influence on the final result. The notion of outlier being hard to define for a single data point, it is also hard, inefficient and often impossible to clean data by removing the outliers. Instead, one can build methods that take as input the contaminated data set and provide as output an estimate which is not very sensitive to the contamination. Recent advances in data acquisition and computational power provoked a revival of interest in robust estimation and learning, with a focus on finite sample results and computationally tractable procedures. This was in contrast with more traditional studies analyzing asymptotic properties of statistical methods.

This paper builds on recent advances made in robust estimation and suggests a procedure that has attractive properties both from asymptotic and finite-sample points of view. Furthermore, it is computationally tractable and its statistical complexity depends optimally on the dimension. As a matter of fact, we even show that what really matters is the intrinsic dimension, defined in the Gaussian model as the effective rank of the covariance matrix.

Note that in the framework of robust estimation, the high-dimensional setting is qualitatively different from the one dimensional setting. This qualitative difference can be seen at two levels. First, from a computational point of view, the running time of several robust methods scales poorly with dimension. Second, from a statistical point of view, while a simple "remove than average" strategy might be successful in low-dimensional setting, it can easily be seen to fail in the high dimensional case. Indeed, assume that $\varepsilon \in (0, 1/2)$ and $n \in \mathbb{N}$ are two numbers and the data consist of $n(1-\varepsilon)$ points (inliers) drawn from a p-dimensional Gaussian distribution $\mathcal{N}_p(0, \mathbf{I}_p)$ (where $\mathbf{I}_p$ is the $p \times p$ identity matrix) and

*ENSAE Paris, CREST, IP Paris

†Yerevan State University and YerevaNN

εn points (outliers) equal to a given vector $\boldsymbol{u}$. A simple strategy consists in removing all the points of Euclidean norm larger than $2\sqrt{p}$ and averaging all the remaining points. If the norm of $\boldsymbol{u}$ is equal to $\sqrt{p}$, one can check that the distance between this estimator and the true mean $\boldsymbol{\mu} = 0$ is of order $\sqrt{p/n} + \varepsilon\|\boldsymbol{u}\|_2 = \sqrt{p/n} + \varepsilon\sqrt{p}$. This error rate is provably optimal in the small dimensional setting $p = O(1)$, but sub-optimal as compared to the optimal rate $\sqrt{p/n}+\varepsilon$. The reason of sub-optimality is that the individually harmless outliers, lying close to the point cloud, have a strong joint impact on the quality of estimation.

We postpone a review of the relevant prior work to Section 4 and provide here a summary of our contributions. In the context of a data set subject to a fully adversarial corruption, we introduce a new estimator of the Gaussian mean that enjoys the following properties (the precise meaning of these properties is given in the next section):

- it is computable in polynomial time,
- it is equivariant with respect to translations and orthogonal transformations,
- it has a high (minimax) breakdown point: $\varepsilon^* = (5 - \sqrt{5})/10 \approx 0.28$,
- it is minimax-rate-optimal, up to a logarithmic factor,
- it is asymptotically efficient when the rate of contamination tends to zero,
- for inhomogeneous covariance matrices, it achieves a better sample complexity than all the other previously studied methods.

2. Desirable properties of a robust estimator

We consider the setting in which the sample points are corrupted versions of independent and identically distributed random vectors drawn from a p-variate Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}$. In what follows, we will assume that the rate of contamination and the covariance matrix are known and, therefore, can be used for constructing an estimator of $\boldsymbol{\mu}^*$.

Definition 1. We say that the distribution $\mathbf{P}_n$ of data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ is Gaussian with adversarial contamination, denoted by $\mathbf{P}_n \in \mathrm{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$ with $\varepsilon \in (0, 1/2)$ and $\boldsymbol{\Sigma} \succeq 0$, if there is a set of n independent and identically distributed random vectors $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ drawn from $\mathcal{N}_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ satisfying $\big|\{i : \boldsymbol{X}_i \neq \boldsymbol{Y}_i\}\big| \leq \varepsilon n$.

The sample points $\boldsymbol{X}_i$ with indices in the set $\mathcal{O} = \{i : \boldsymbol{X}_i \neq \boldsymbol{Y}_i\}$ are called outliers, while all the other sample points are called inliers. Assumption GAC allows both the set of outliers $\mathcal{O}$ and the outliers themselves to be random and to depend arbitrarily on the values of $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$. The Statistician aims at estimating $\boldsymbol{\mu}^*$ as accurately as possible, the accuracy being measured by the expected estimation error:

$$R_{\mathbf{P}_n}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) = \|\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*\|_{\mathbb{L}_2(\mathbf{P}_n)} = \left(\sum_{j=1}^{p} \mathbf{E}_{\mathbf{P}_n}[(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)_j^2]\right)^{1/2}.$$

The goal of the Statistician is to find an estimator $\widehat{\boldsymbol{\mu}}_n$ that minimizes the worst-case risk

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\Sigma}, \varepsilon) = \sup_{\boldsymbol{\mu}^* \in \mathbb{R}^p} \sup_{\mathbf{P}_n \in \mathrm{GAC}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)} R_{\mathbf{P}_n}(\widehat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*).$$

Let $\mathbf{r}_{\mathbf{\Sigma}} = \mathrm{Tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_{\mathrm{op}}$ be the effective rank of $\mathbf{\Sigma}$. Using the theory developed by [Chen et al., 2016, 2018], one can check that

$$\inf_{\widehat{\boldsymbol{\mu}}_n} R_{\max}(\widehat{\boldsymbol{\mu}}_n, \mathbf{\Sigma}, \varepsilon) \geq c\|\mathbf{\Sigma}\|_{\mathrm{op}}^{1/2}\left(\sqrt{\frac{\mathbf{r}_{\mathbf{\Sigma}}}{n}} + \varepsilon\right) \tag{1}$$

for some constant $c > 0$, where the infimum is over all measurable functions of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. This naturally leads to the following definition.

Definition 2. We say that the estimator $\widehat{\boldsymbol{\mu}}_n$ is minimax rate optimal (in expectation), if there are universal constants c_1, c_2 and C such that

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \mathbf{\Sigma}, \varepsilon) \leq C\|\mathbf{\Sigma}\|_{\mathrm{op}}^{1/2}\left(\sqrt{\frac{\mathbf{r}_{\mathbf{\Sigma}}}{n}} + \varepsilon\right)$$

for every $(n, \mathbf{\Sigma}, \varepsilon)$ satisfying $\mathbf{r}_{\mathbf{\Sigma}} \leq c_1 n$ and $\varepsilon \leq c_2$.

The iteratively reweighted mean estimator, introduced in the next section, is not minimax rate optimal but is very close to being so. Indeed, it is provably minimax rate optimal up to a $\sqrt{\log(1/\varepsilon)}$ factor.

Definition 3. We say that $\widehat{\boldsymbol{\mu}}_n$ is an asymptotically efficient estimator of $\boldsymbol{\mu}^*$, if when $\varepsilon = \varepsilon_n$ tends to zero sufficiently fast, as n tends to infinity, we have

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \mathbf{\Sigma}, \varepsilon) \leq \|\mathbf{\Sigma}\|_{\mathrm{op}}^{1/2}\sqrt{\frac{\mathbf{r}_{\mathbf{\Sigma}}}{n}}\,\big(1 + o_n(1)\big).$$

One can infer from (1) that a necessary condition for the existence of asymptotically efficient estimator is $\varepsilon_n^2 = o(\mathbf{r}_{\mathbf{\Sigma}}/n)$. We show in the next section that this condition is almost sufficient, by proving that the iteratively reweighted mean estimator is asymptotically efficient provided that $\varepsilon_n^2 \log(1/\varepsilon_n) = o(\mathbf{r}_{\mathbf{\Sigma}}/n)$.

The last notion that we introduce in this section is the breakdown point, the term being coined by Hampel [1968], see also [Donoho and Huber, 1983]. Roughly speaking, the breakdown point of a given estimator is the largest proportion of outliers that the estimator can support without becoming infinitely large.

Definition 4. We say that $\varepsilon_n^* \in [0, 1/2]$ is the (finite-sample) breakdown point of the estimator $\widehat{\boldsymbol{\mu}}_n$, if

$$R_{\max}(\widehat{\boldsymbol{\mu}}_n, \mathbf{\Sigma}, \varepsilon) < +\infty, \qquad \forall \varepsilon < \varepsilon_n^*$$

and $R_{\max}(\widehat{\boldsymbol{\mu}}_n, \mathbf{\Sigma}, \varepsilon) = +\infty$, for every $\varepsilon > \varepsilon_n^*$.

One can check that the breakdown points of the componentwise median and the geometric median (see the definition of $\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}$ in (3) below) are equal to $1/2$. Unfortunately, the minimax rate of these methods is strongly sub-optimal, see [Chen et al., 2018, Prop. 2.1] and [Lai et al., 2016, Prop. 2.1]. Among all rate-optimal (up to a polylogarithmic factor) robust estimators, Tukey's median is the one with highest known breakdown point equal to $1/3$ [Donoho and Gasko, 1992]. This notion of breakdown point, well adapted to estimators that do not rely on the knowledge of ε, becomes less relevant in the context of known ε. Indeed, if a given estimator $\widehat{\boldsymbol{\mu}}_n(\varepsilon)$ is proved to have a breakdown point equal to 0.1, one can consider instead the estimator $\widetilde{\boldsymbol{\mu}}_n(\varepsilon) = \widehat{\boldsymbol{\mu}}_n(\varepsilon)\mathbb{1}(\varepsilon < 0.1) + \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}}\mathbb{1}(\varepsilon \geq 0.1)$, which will have a breakdown point equal to 0.5. For this reason, it appears more appealing to consider a different notion that we call rate-breakdown point, and which is of the same flavor as the δ-breakdown point defined in [Chen et al., 2016].

Algorithm 1: Iteratively reweighted mean estimator

Input: data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$, contamination rate ε and $\boldsymbol{\Sigma}$
Output: parameter estimate $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$
Initialize: compute $\boldsymbol{\mu}^0$ as a minimizer of $\sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2$
Set $K = 0 \vee \left\lceil \frac{\log(4\mathbf{r}_{\boldsymbol{\Sigma}}) - 2\log(\varepsilon(1-2\varepsilon))}{2\log(1-2\varepsilon) - \log\varepsilon - \log(1-\varepsilon)} \right\rceil$.
For $k = 1 : K$
  For $i = 1 : n$
    Set $\mathbf{M}_i = (\boldsymbol{X}_i - \boldsymbol{\mu}^{k-1})(\boldsymbol{X}_i - \boldsymbol{\mu}^{k-1})^\top$
  EndFor
  Compute current weights:

$$\boldsymbol{w} \in \operatorname*{arg\,min}_{(n-n\varepsilon)\|\boldsymbol{w}\|_\infty \le 1} \lambda_{\max}\Big(\sum_{i=1}^n w_i \mathbf{M}_i - \boldsymbol{\Sigma}\Big) \vee 0.$$

  Update the estimator: $\widehat{\boldsymbol{\mu}}^k = \sum_{i=1}^n w_i \boldsymbol{X}_i$.
EndFor
Return $\widehat{\boldsymbol{\mu}}^K$.

Definition 5. We say that $\varepsilon_r^* \in [0, 1/2]$ is the $r(n, \boldsymbol{\Sigma}, \varepsilon)$-breakdown point of the estimator $\widehat{\boldsymbol{\mu}}_n$ for a given function $r : \mathbb{N} \times \mathcal{S}_+^p \times [0, 1/2)$, if for every $\varepsilon < \varepsilon_r^*$,

$$\sup_{n,p} \frac{R_{\max}(\widehat{\boldsymbol{\mu}}_n(\varepsilon), \boldsymbol{\Sigma}, \varepsilon)}{r(n, \boldsymbol{\Sigma}, \varepsilon)} < +\infty.$$

In the context of Gaussian mean estimation, if the previous definition is applied with $r(n, \boldsymbol{\Sigma}, \varepsilon) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + \varepsilon\big)$, we call the corresponding value the minimax-rate-breakdown point. Similarly, if $r(n, \boldsymbol{\Sigma}, \varepsilon) = \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + \varepsilon\sqrt{\log(1/\varepsilon)}\big)$, we call the corresponding value the nearly-minimax-rate-breakdown point.

3. Iterative reweighting approach

In this section, we define the iterative reweighting estimator that will be later proved to enjoy all the desirable properties. To this end, we set

$$\bar{\boldsymbol{X}}_{\boldsymbol{w}} = \sum_{i=1}^n w_i \boldsymbol{X}_i, \qquad G(\boldsymbol{w}, \boldsymbol{\mu}) = \lambda_{\max,+}\Big(\sum_{i=1}^n w_i (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top - \boldsymbol{\Sigma}\Big) \tag{2}$$

for any pair of vectors $\boldsymbol{w} \in [0, 1]^n$ and $\boldsymbol{\mu} \in \mathbb{R}^p$. The main idea of the proposed methods is to find a weight vector $\widehat{\boldsymbol{w}}_n$ belonging to the probability simplex

$$\boldsymbol{\Delta}^{n-1} = \Big\{\boldsymbol{w} \in [0, 1]^n : w_1 + \ldots + w_n = 1\Big\}$$

that mimics the ideal weight vector $\boldsymbol{w}^*$ defined by $w_j^* = \mathbb{1}(j \in \mathcal{I})/|\mathcal{I}|$, so that the weighted average $\bar{\boldsymbol{X}}_{\widehat{\boldsymbol{w}}_n}$ is nearly as close to $\boldsymbol{\mu}^*$ as the average of the inliers.

The precise definition is as follows. We start from an arbitrary initial estimator $\widehat{\boldsymbol{\mu}}^0$ of $\boldsymbol{\mu}^*$. To give a concrete example, and also in order to guarantee equivariance by translations and orthogonal transformations, we assume that $\widehat{\boldsymbol{\mu}}^0$ is the geometric median:

$$\widehat{\boldsymbol{\mu}}^0 = \widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}} \in \operatorname*{arg\,min}_{\boldsymbol{\mu} \in \mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\mu}\|_2. \tag{3}$$

Definition 6. We call iteratively reweighted mean estimator, denoted by $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$, the K-th element of the sequence $\{\widehat{\boldsymbol{\mu}}^k; k = 0, 1, \ldots\}$ starting from $\widehat{\boldsymbol{\mu}}^0$ in (3) and defined by the recursion

$$\widehat{\boldsymbol{w}}^k \in \arg \min_{(n-n\varepsilon)\|\boldsymbol{w}\|_\infty \leq 1} G(\boldsymbol{w}, \widehat{\boldsymbol{\mu}}^{k-1}), \qquad \widehat{\boldsymbol{\mu}}^k = \bar{\boldsymbol{X}}_{\widehat{\boldsymbol{w}}^k}, \tag{4}$$

where the minimum is over all weight vectors $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ satisfying $\max_j w_j \leq 1/(n-n\varepsilon)$ and the number of iteration is

$$K = 0 \bigvee \left\lceil \frac{\log(4\mathbf{r}_{\boldsymbol{\Sigma}}) - 2\log(\varepsilon(1-2\varepsilon))}{2\log(1-2\varepsilon) - \log\varepsilon - \log(1-\varepsilon)} \right\rceil. \tag{5}$$

The idea of computing a weighted mean, with weights measuring the outlyingness of the observations goes back at least to [Donoho, 1982, Stahel, 1981]. Perhaps the first idea similar to that of minimizing the largest eigenvalue of the covariance matrix was that of minimizing the determinant of the sample covariance matrix over all subsamples of a given cardinality [Rousseeuw, 1985, 1984]. It was also observed in [Lopuhaä and Rousseeuw, 1991] that one can improve the estimator by iteratively updating the weights. An overview of these results can be found in [Rousseeuw and Hubert, 2013].

Note that the value of K provided above is tailored to the case where the initial estimator is the geometric median. Clearly, K depends only logarithmically on the dimension and $K = K_\varepsilon$ tends to 2 when ε goes to zero. The rest of this section is devoted to showing that the iteratively reweighted estimator enjoys all the desirable properties announced in the introduction.

Fact 1

The estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is computationally tractable.

In order to check computational tractability, it suffices to prove that each iteration of the algorithm can be performed in polynomial time. Since the number of iterations depends logarithmically on p, this will suffice. Note now that the optimization problem in (4) is convex and can be cast into a semi-definite program. Indeed, it is equivalent to minimizing a real value $t \geq 0$ over all the pairs $(t, \boldsymbol{w})$ satisfying the constraints

$$\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}, \quad \|\boldsymbol{w}\|_\infty \leq \frac{1}{n(1-\varepsilon)}, \quad \sum_{i=1}^n w_i(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{k-1})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{k-1})^\top \preceq \boldsymbol{\Sigma} + t\mathbf{I}_p.$$

The first two constraints can be rewritten as a set of linear inequalities, while the third constraint is a linear matrix inequality.

Fact 2

The estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is translation and orthogonal transformation equivariant.

The equivariance mentioned in this statement should be understood as follows. If we denote by $\widehat{\boldsymbol{\mu}}_{n,X}^{\mathrm{IR}}$ the estimator computed for data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and by $\widehat{\boldsymbol{\mu}}_{n,X'}^{\mathrm{IR}}$ the one computed for

data $\boldsymbol{X}'_1, \ldots, \boldsymbol{X}'_n$ with $\boldsymbol{X}'_i = \boldsymbol{a} + \mathbf{U}\boldsymbol{X}_i$, where $\boldsymbol{a} \in \mathbb{R}^p$ and $\mathbf{U}$ is a $p \times p$ orthogonal matrix, then $\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_{n,X'} = \boldsymbol{a} + \mathbf{U}\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_{n,X}$. To prove this property, we first note that

$$\min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}'_i - \boldsymbol{\mu}\|_2 = \min_{\boldsymbol{\mu}\in\mathbb{R}^p} \sum_{i=1}^n \|\boldsymbol{X}_i - \mathbf{U}^\top(\boldsymbol{\mu} - \boldsymbol{a})\|_2.$$

This implies that $\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X} = \mathbf{U}^\top(\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X'} - \boldsymbol{a})$, which is equivalent to $\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X'} = \boldsymbol{a} + \mathbf{U}\widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_{n,X}$. Therefore, the initial value of the recursion is equivariant. If we add to this the fact that[1] $G_X(\boldsymbol{w}, \boldsymbol{\mu}) = G_{X'}(\boldsymbol{w}, \boldsymbol{a} + \mathbf{U}\boldsymbol{\mu})$ for every $(\boldsymbol{w}, \boldsymbol{\mu})$, we get the equivariance of $\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_n$.

Fact 3

The breakdown point ε^*_n and the nearly-minimax-rate-breakdown point $\varepsilon^*_{\mathrm{r}}$ of $\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_n$ satisfy, respectively $\varepsilon^*_n = 0.5$ and $\varepsilon^*_{\mathrm{r}} \geq (5 - \sqrt{5})/10 \approx 0.28$.

It can be proved that if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ satisfy GAC$(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}, \varepsilon)$, there is a random variable Ξ depending only on $\boldsymbol{\zeta}_i = \boldsymbol{Y}_i - \boldsymbol{\mu}^*$, $i = 1, \ldots, n$, such that

$$\|\bar{\boldsymbol{X}}_{\boldsymbol{w}} - \boldsymbol{\mu}^*\|_2 \leq \frac{\sqrt{\varepsilon(1-\varepsilon)}}{1-2\varepsilon}\, G(\boldsymbol{w}, \boldsymbol{\mu})^{1/2} + \Xi, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^p, \tag{6}$$

for every $\boldsymbol{w} \in \boldsymbol{\Delta}^{n-1}$ such that $n(1-\varepsilon)\|\boldsymbol{w}\|_\infty \leq 1$. Inequality (6) is one of the main building blocks of the proof of Facts 3 to 5. This inequality, as well as inequalities (8) and (9) below will be formally stated and proved in subsequent sections. To check Fact 3, we set $\alpha_\varepsilon = \sqrt{\varepsilon(1-\varepsilon)}/(1-2\varepsilon)$ and note that

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}^k - \boldsymbol{\mu}^*\|_2 = \|\bar{\boldsymbol{X}}_{\widehat{\boldsymbol{w}}^k} - \boldsymbol{\mu}^*\|_2 &\leq \alpha_\varepsilon\, G(\widehat{\boldsymbol{w}}^k, \widehat{\boldsymbol{\mu}}^{k-1})^{1/2} + \Xi \\
&\leq \alpha_\varepsilon\, G(\boldsymbol{w}^*, \widehat{\boldsymbol{\mu}}^{k-1})^{1/2} + \Xi \\
&\leq \alpha_\varepsilon \left(G(\boldsymbol{w}^*, \bar{\boldsymbol{X}}_{\boldsymbol{w}^*}) + \|\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1}\|_2^2\right)^{1/2} + \Xi \\
&\leq \alpha_\varepsilon \left(G(\boldsymbol{w}^*, \boldsymbol{\mu}^*) + \|\bar{\boldsymbol{X}}_{\boldsymbol{w}^*} - \widehat{\boldsymbol{\mu}}^{k-1}\|_2^2\right)^{1/2} + \Xi \\
&\leq \alpha_\varepsilon\, \|\widehat{\boldsymbol{\mu}}^{k-1} - \boldsymbol{\mu}^*\|_2 + \widetilde{\Xi},
\end{aligned}$$

where $\widetilde{\Xi} = \alpha_\varepsilon \left(G(\boldsymbol{w}^*, \boldsymbol{\mu}^*)^{1/2} + \|\bar{\boldsymbol{\xi}}_{\boldsymbol{w}^*}\|_2\right) + \Xi$. Unfolding this recursion, we get [2]

$$\|\widehat{\boldsymbol{\mu}}^{\mathrm{IR}}_n - \boldsymbol{\mu}^*\|_2 = \|\widehat{\boldsymbol{\mu}}^K - \boldsymbol{\mu}^*\|_2 \leq \alpha_\varepsilon^K\, \|\widehat{\boldsymbol{\mu}}^0 - \boldsymbol{\mu}^*\|_2 + \frac{\widetilde{\Xi}}{1-\alpha_\varepsilon}. \tag{7}$$

The geometric median $\widehat{\boldsymbol{\mu}}^0 = \widehat{\boldsymbol{\mu}}^{\mathrm{GM}}_n$ having a breakdown point equal to $1/2$, we infer from the last display that the error of the iteratively reweighted estimator remains bounded after altering ε-fraction of data points provided that $\alpha_\varepsilon < 1$. This implies that the breakdown point is at least equal to the solution of the equation $\sqrt{\varepsilon(1-\varepsilon)} = 1 - 2\varepsilon$, which yields $\varepsilon^* \geq (5 - \sqrt{5})/10$. Moreover, if $\varepsilon \in [(5 - \sqrt{5})/10, 1/2]$, then the number of iterations K equals zero and the iteratively reweighted mean coincides with the geometric median. Therefore, its breakdown point is $1/2$.

[1] We use here the notation $G_X(\boldsymbol{w}, \boldsymbol{\mu})$ to make clear the dependence of G in (2) on $\boldsymbol{X}_i$s. We also stress that when the estimator is computed for the transformed data $\boldsymbol{X}'_i$, the matrix $\boldsymbol{\Sigma}$ is naturally replaced by $\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$.

[2] Here and in the sequel α_ε^K stands for K-th power of α_ε.

Fact 4

The estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is nearly minimax rate optimal, in the sense that its worst-case risk is of order $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}\big(\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + \varepsilon\sqrt{\log(1/\varepsilon)}\big)$.

Without loss of generality, we assume that $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = 1$ so that $\mathbf{r}_{\boldsymbol{\Sigma}} = \mathrm{Tr}(\boldsymbol{\Sigma})$. We can always reduce ourselves to this case by considering scaled data points $\boldsymbol{X}_i/\|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2}$ instead of $\boldsymbol{X}_i$. Combining (7) and the triangle inequality, we get

$$\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}} - \boldsymbol{\mu}^*\|_{\mathbb{L}_2} \le \alpha_\varepsilon^K \, \|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}} - \boldsymbol{\mu}^*\|_{\mathbb{L}_2} + \frac{\|\widetilde{\Xi}\|_{\mathbb{L}_2}}{1-\alpha_\varepsilon}.$$

It is not hard to check that $\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{GM}} - \boldsymbol{\mu}^*\|_{\mathbb{L}_2} \le 2\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}}/(1-2\varepsilon)$. Furthermore, the choice of K in (5) is made in such a way that $2\alpha_\varepsilon^K\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}} \le \varepsilon(1-2\varepsilon)$. This implies that

$$\|\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}} - \boldsymbol{\mu}^*\|_{\mathbb{L}_2} \le \varepsilon + \frac{\|\widetilde{\Xi}\|_{\mathbb{L}_2}}{1-\alpha_\varepsilon}.$$

The last two building blocs of the proof are the following[3] inequalities:

$$\mathbf{E}[G(\boldsymbol{w}^*, \boldsymbol{\mu}^*)] \le C\big(1 + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\big)\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}, \tag{8}$$

$$\|\Xi\|_{\mathbb{L}_2} \le \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}(1 + C\sqrt{\varepsilon}) + C\sqrt{\varepsilon}(\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4} + C\varepsilon\sqrt{\log(1/\varepsilon)}, \tag{9}$$

where $C > 0$ is a universal constant. In what follows, the value of C may change from one line to the other. We have

$$\begin{aligned}
\|\widetilde{\Xi}\|_{\mathbb{L}_2} &\le \alpha_\varepsilon\big(\|G(\boldsymbol{w}^*, \boldsymbol{\mu}^*)^{1/2}\|_{\mathbb{L}_2} + \|\bar{\boldsymbol{\xi}}_{\boldsymbol{w}^*}\|_{\mathbb{L}_2}\big) + \|\Xi\|_{\mathbb{L}_2} \\
&\le C\sqrt{\varepsilon}\,\big(\mathbf{E}^{1/2}[G(\boldsymbol{w}^*, \boldsymbol{\mu}^*)] + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\big) + \|\Xi\|_{\mathbb{L}_2}^2 \\
&\le C\varepsilon\,\big((\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4} + \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}\big) + \|\Xi\|_{\mathbb{L}_2} \\
&\le \sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n}(1 + C\sqrt{\varepsilon}) + C\sqrt{\varepsilon}(\mathbf{r}_{\boldsymbol{\Sigma}}/n)^{1/4} + C\varepsilon\sqrt{\log(1/\varepsilon)} \\
&\le C\sqrt{\mathbf{r}_{\boldsymbol{\Sigma}}/n} + C\varepsilon\sqrt{\log(1/\varepsilon)}.
\end{aligned} \tag{10}$$

Returning to (7) and combining it with (10), we get the claim of Fact 4 for every $\varepsilon \le \varepsilon_0$, where ε_0 is any positive number strictly smaller than $(5-\sqrt{5})/10$. This also proves the second claim of Fact 3.

Fact 5

In the setting $\varepsilon = \varepsilon_n \to 0$ so that $\varepsilon^2\log(1/\varepsilon) = o(\mathbf{r}_{\boldsymbol{\Sigma}}/n)$ when $n \to \infty$, the estimator $\widehat{\boldsymbol{\mu}}_n^{\mathrm{IR}}$ is asymptotically efficient.

The proof of this fact follows from (7) and (9). Indeed, if $\varepsilon^2\log(1/\varepsilon) = o(\mathbf{r}_{\boldsymbol{\Sigma}}/n)$, (9) implies that

$$\|\widetilde{\Xi}\|_{\mathbb{L}_2}^2 \le \frac{\mathbf{r}_{\boldsymbol{\Sigma}}}{n}\big(1 + o(1)\big).$$

Injecting this bound in (7) and using the fact that ε tends to zero, we get the claim of Fact 5.

[3] Inequality (8) is [Koltchinskii and Lounici, 2017, Theorem 4].

4. Relation to prior work and discussion

Robust estimation of a mean is a statistical problem studied by many authors since at least sixty years. It is impossible to give an overview of all existing results and we will not try to do it here. The interested reader may refer to the books [Maronna et al., 2006] and [Huber and Ronchetti, 2009]. We will rather focus here on some recent results that are the most closely related to the present work. Let us just recall that Huber and Ronchetti [2009] enumerates three desirable properties of a statistical procedure: efficiency, stability and breakdown. We showed here that iteratively reweighted mean estimator possesses these features and, in addition, is equivariant and computationally tractable.

To the best of our knowledge, the form $\sqrt{p/n}+\varepsilon$ of the minimax risk in the Gaussian mean estimation problem has been first obtained by Chen et al. [2018]. They proved that this rate holds with high probability for the Tukey median, which is known to be computationally intractable in the high-dimensional setting. The first nearly-rate-optimal and computationally tractable estimators have been proposed by Lai et al. [2016] and Diakonikolas et al. [2016][4]. The methods analyzed in these papers are different, but they share the same idea: If for a subsample of points the empirical covariance matrix is sufficiently close to the theoretical one, then the arithmetic mean of this subsample is a good estimator of the theoretical mean. Our method is based on this idea as well, which is mathematically formalized in (6).

Further improvements in running times—up to obtaining a linear in np computational complexity in the case of a constant ε—are presented in [Cheng et al., 2019]. Some lower bounds suggesting that the log-factor in the term $\varepsilon\sqrt{\log(1/\varepsilon)}$ cannot be removed from the rate of computationally tractable estimators are established in [Diakonikolas et al., 2017]. In a slightly weaker model of corruption, [Diakonikolas et al., 2018] propose an iterative filtering algorithm that achieves the optimal rate ε without the extra factor $\sqrt{\log(1/\varepsilon)}$. On a related note [Collier and Dalalyan, 2019] shows that in a weaker contamination model termed as parametric contamination, the carefully trimmed mean can achieve a better rate than that of the coordinatewise/geometric median.

An overview of the recent advances on robust estimation with a focus on computational aspects can be found in [Diakonikolas and Kane, 2019]. Extensions of these methods to the sparse mean estimation are developed in [Balakrishnan et al., 2017, Diakonikolas et al., 2019b]. All these results are proved to hold on an event with a prescribed probability, see [Bateni and Dalalyan, 2019] for a relation between results in expectation and those with high probability, as well as for the definitions of various types of contamination.

REFERENCES

S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *COLT 2017*, pages 169–212, 2017.

A.-H. Bateni and A. S. Dalalyan. Confidence regions and minimax rates in outlier-robust estimation on the probability simplex, 2019.

M. Chen, C. Gao, and Z. Ren. A general decision theory for Huber's ϵ-contamination model. *Electron. J. Statist.*, 10(2):3752–3774, 2016.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 10 2018.

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. In *SODA 2019*, pages 2755–2771, 2019.

[4]See [Diakonikolas et al., 2019a] for the extended version

O. Collier and A. S. Dalalyan. Multidimensional linear functional estimation in sparse gaussian models and robust estimation of the mean. *Electron. J. Statist.*, 13(2):2830–2864, 2019.

I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *FOCS 2016*, pages 655–664, 2016.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *FOCS 2017*, pages 73–84, 2017.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *SODA 2018*, pages 2683–2702, 2018.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019a.

I. Diakonikolas, D. Kane, S. Karmalkar, E. Price, and A. Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *NeurIPS 2019*, pages 10688–10699, 2019b.

D. Donoho. Breakdown properties of multivariate location estimators, 1982.

D. Donoho and P. J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. 1983.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

F. R. Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California, Berkeley, 1968.

P. J. Huber and E. M. Ronchetti. *Robust Statistics, Second Edition*. Wiley Series in Probability and Statistics. Wiley, 2009.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

K. A. Lai, A. B. Rao, and S. S. Vempala. Agnostic estimation of mean and covariance. In *FOCS 2016*, pages 665–674, 2016.

H. P. Lopuhaä and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19(1):229–248, 1991.

R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. Wiley, 2006.

P. Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pages 283–297. Reidel, Dordrecht, 1985.

P. Rousseeuw and M. Hubert. High-breakdown estimators of multivariate location and scatter. In *Robustness and complex data structures*, pages 49–66. Springer, Heidelberg, 2013.

P. J. Rousseeuw. Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388): 871–880, 1984.

W. Stahel. Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen, 1981.