Fast and simple algorithms for robust mean estimation

Jules Depersin JULES.DEPERSIN@ENSAE.FR
ENSAE,CREST
5 avenue Henry Le Chatelier
91120 Palaiseau. France

Guillaume Lecué GUILLAUME.LECUE@ENSAE.FR
ENSAE,CREST
5 avenue Henry Le Chatelier
91120 Palaiseau. France

Editor: ...

Abstract

We construct fast and simple algorithms that estimate the mean of a vector valued random variable from a dataset that may have been corrupted by outliers and under only a second moment assumption. The resulting algorithms achieve the same statistical bound as Minsker's geometric median-of-means Minsker (2015). Using our approach we also construct a simplified version of the geometric median by replacing the entire search space by a simple finite set without losing the statistical property of the geometric median.

Keywords: Robust mean estimation, median-of-means estimators.

1. Introduction

Let $\mathcal{X}$ be a Hilbert space. Let X be a random vector taking its values in $\mathcal{X}$, with mean $\mu = \mathbb{E}X$. We are given data $(X_i)_{i=1}^N$ taking their values in $\mathcal{X}$, our aim is to estimate μ using these data. We are in particular interested in situations where the dataset contains a fraction of outliers data $(X_i)_{i\in\mathcal{O}}$ that has nothing to do with X and the remaining data $(X_i)_{i\in\mathcal{I}}$ where $\mathcal{I} = [N]\backslash\mathcal{O}$ are only assumed to have the same mean as X and a second moment (without assuming that they are i.i.d.).

If the data $(X_i)_{i=1}^N$ were known to be i.i.d. having the same distribution as X then the classical approach to that problem would be to use the empirical mean $\hat{\mu}_N = (1/N)\sum_{i=1}^N X_i$ to estimate μ, but this solution has some limitations. First, it is not robust to adversarial data: we can make the empirical mean as large as we want just by changing a single data. So even one adversarial outliers is enough for this estimator to fail. Another limitation is that it is not robust to heavy-tailed data: if the data are only assumed to have a second moment, then, for a given confidence $1-\delta \in (0,1)$, one cannot expect a confidence interval for μ centered in $\hat{\mu}_N$ of lenght smaller than $\sqrt{\mathrm{Tr}(\Sigma)/(N\delta)}$, up to a multiplicative constant, where $\mathrm{Tr}(\Sigma)$ is the trace of the covariance operator associated with X.

There has been an important renewal of the problem of mean estimation for heavy-tailed data and corrupted databases during the last decade starting with Catoni (2012b); Lerasle and Oliveira (2011). For instance, in the one-dimensional case, one can get exponential confidence intervals:

$$\mathbb{P}\left(|\mu - \hat{\mu}| > L\sigma\sqrt{\frac{t}{N}}\right) < \exp(-t) \tag{1}$$

where σ^2 is the variance of the distribution and L is an absolute constant (see Devroye et al. (2016b) or Catoni (2012a)), where the only assumption on the data is finite second moment Lugosi et al. (2019). Following this trend, some multi-dimensional robust estimation strategies have been investigated. We can basically identify three lines of researches on this topic:

- constructing statistical procedures achieving the *sub-gaussian* rate (which is the rate achieved by the empirical mean when the data are i.i.d. Gaussian $\mathcal{N}(\mu, \Sigma)$ where Σ is the covariance matrix). This is the approach from Lugosi et al. (2019); Catoni and Giulini (2017); M. Lerasle and Lecué (2017); Chen et al. (2018). In this approach only the statistical properties matter and algorithmic performance are not studied.

- constructing algorithms with optimal statistical properties (such as achieving the information-theoretically optimal error in terms of the ϵ-proportion of outliers or the subgaussian rate of estimation) together with a control of the running time of the algorithm. Only polynomial time algorithms are constructed and linear-time algorithms, that is algorithms running in $\mathcal{O}(Nd)$, being the optimal desired constructions. This is the line of researches followed by Hopkins (2018); Cherapanamjeri et al. (2019b); Depersin and Lecué (2019); Cheng et al. (2019). Most of these algorithms do not come with efficient code even the one proved to run in (nearly) linear time (constructing the empirical mean also takes $\mathcal{O}(Nd)$). This is due, in part, to the use of some SDP relaxation such as the Sum-of-Squares approach or some SDP solvers which have not yet being coded efficiently or do not scale efficiently (see Peng and Tangwongsan (2012)).

- constructing computationally tractable algorithms. Here the aim is to provide efficient code and then, if possible, to prove some statistical and algorithmic properties (even sub-optimal ones). It is for instance the case of Minsker's geometric median Minsker (2015) which obtain sub-optimal convergence rate for an efficient tractable linear-time algorithm coming with the geometric median (see Cohen et al. (2016)). This is also the approach of Diakonikolas et al. (2018, 2017).

Our work focuses on algorithms coming with tractable implementation and is therefore in the third line of research mentioned above. In this paper, we construct a robust algorithm for estimation of the mean μ, that achieve exponentially likely confidence intervals and is computationally efficient under the only following assumption:

Assumption 1 *There exists a partition $\mathcal{O} \sqcup \mathcal{I}$ of $\{1, \ldots, N\}$ such that $(X_i)_{i \in \mathcal{I}}$ are independent and for all $i \in \mathcal{I}$, $\mathbb{E} X_i = \mu$ and $\mathbb{E}\left\|X_i - \mu\right\|_2^2 \leq \mathbb{E}\left\|X - \mu\right\|_2^2 = \mathrm{Tr}(\Sigma)$ where Σ is the covariance operator of X.*

In particular, no assumption is made on the data $(X_i)_{i \in \mathcal{O}}$ – the one indexed by $\mathcal{O}$ – which can therefore be seen as outliers or adversarial data. Moreover, the informativre data $(X_i)_{i \in \mathcal{I}}$ are only assumed to have a second moment. The framework given by Assumption 1 encompasses the two type of robustness that have been considered recently: robustness w.r.t. heavy-tailed data and robustness w.r.t. to data corruption by adversarial outliers.

We first present our procedures and algorithms, then we give their statistical guarantees. In a third part we show how to simplify the geometric median technique developed in Minsker (2015), without losing its statistical properties. We then state how to make our estimation technique adaptive in the number of outliers (meaning one does not have to know the proportion of outliers to construct the estimator). We finally present some simulations, in order to show that our algorithms are computationally tractable and that their empirical behavior match the theoretical bounds.

Notations: The cardinality of a set A is denoted by $|A|$. Given K real numbers $a_1, \ldots, a_K$ we define their median as $\mathrm{Med}\{a_1, \ldots, a_K\} = \inf_{\substack{J \subset \{1,\ldots,K\} \\ |J| \geq \lceil K/2 \rceil}} \sup_{j \in J} a_j$.

2. A search algorithm in the mutual distance matrix of block means

The so-called *Median-of-Mean* (MOM) approach Nemirovsky and Yudin (1983); Alon et al. (1999); Jerrum et al. (1986), widely investigated in the last few years Bubeck et al. (2013); Lerasle and Oliveira (2011); Devroye et al. (2016a); Minsker and Strawn (2017), often yields robust estimation strategies. Let us give the general idea behind that approach: we first randomly split the data into K equal-size blocks B_1, ..., B_K

(if K does not divide N, we just remove some data). We then compute the empirical mean within each block:

$$\bar{X}_k = \frac{1}{|B_k|} \sum_{i \in B_k} X_i$$

for $k = 1, \dots, K$, where we set $|B_k| = \mathrm{Card}(B_k) = N/K$. In the one-dimensional case, we then take the median of the latter K empirical means to construct an estimator of the mean. It is more complicated in the multi-dimensional case, where there is no "definitive" equivalent of the one dimensional median but several candidates: coordinate-wise median, the geometric median (also known as Fermat point), the Tukey Median, among many others (see Small (1990)). The strength of this approach is the robustness of the median operator, which leads to good statistical properties even on corrupted databases. But some of them can be very hard to compute Johnson and Preparata (1978).

The geometric median-of-mean has received an important attention recently starting with the work from Minsker (2015). It is defined as

$$\hat{\mu}_K^{geo} \in \operatorname*{argmin}_{u \in H} \sum_{k=1}^{K} \left\| \bar{X}_k - u \right\|_2 . \tag{2}$$

It is proved in Minsker (2015) that with probability at least $1 - 2\exp(-K)$,

$$\left\| \hat{\mu}_K^{geo} - \mu \right\|_2 \leq \sqrt{\frac{\mathrm{Tr}(\Sigma) K}{N}}. \tag{3}$$

The latter rate has been improved in several recent papers Lugosi et al. (2019); M. Lerasle and Lecué (2017); Catoni and Giulini (2017); Hopkins (2018); Cherapanamjeri et al. (2019a); Depersin and Lecué (2019) where it is shown that the "subgaussian rate"

$$\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} K}{N}}$$

can be achieved by polynomial time algorithms Hopkins (2018); Cherapanamjeri et al. (2019a) or even nearly linear time algorithms such as in Depersin and Lecué (2019). But none of the latter algorithms come with actual codes and they should be looked at more as theoretical than practical results.

Our aim here is more on the practical side even though we prove theoretical bounds such as the one in (3). We therefore provide simple codes available at MOMpower github page and a Simulation section. Here we take advantage of the particular features of this problem in order to propose an other estimator inspired by the MOM-approach and somehow very easy to compute.

Our first approach is to consider the procedure

$$\hat{\mu}_K^{(0)} = \operatorname*{argmin}_{a \in H} \ \mathrm{Med}\{\left\| a - \bar{X}_k \right\|_2 : 1 \leq k \leq K\}. \tag{4}$$

We will prove below that it achieves the same statistical performance as the geometric median in (3). But from a computational point-of-view, there is no advantage to use this procedure compare to the geometric median, on the contrary since the objective function in (4) is not convex whereas it is convex in the geometric median. Nevertheless, (4) paved the way toward a procedure which comes with a very simple search algorithm. We show below that (4) can be simplified a lot by reducing the search space H to a finite set made of only K elements, the K empirical means $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_K$. This drastic simplification comes at no price from a statistical point of view since the resulting estimator

$$\hat{\mu}_K^{(1)} = \operatorname*{argmin}_{a \in \{\bar{X}_1, \bar{X}_2, \dots, \bar{X}_K\}} \ \mathrm{Med}\{\left\| a - \bar{X}_k \right\|_2 : 1 \leq k \leq K\} \tag{5}$$

```
input  : $X_1,\dots,X_N$: $N$ data points in $\mathcal{X}$, $K$: a number of blocks
output: estimator of the mean of $X$
Construct an equipartition $B_1 \sqcup \dots \sqcup B_K = \{1,\dots,N\}$
Construct the $K$ empirical means $\bar{X}_k = (N/K)\sum_{i\in B_k} X_i$
Construct the $K \times K$ matrix $\left(\|\bar{X}_{k_1} - \bar{X}_{k_2}\|_2\right)_{1\leq k_1,k_2\leq K}$
for $k_1 \in \{1,\dots,K\}$ do
    Compute $M(k_1) = \text{Med}\{\|\bar{X}_{k_1} - \bar{X}_{k_2}\|_2 : k_2 \in \{1,\dots,K\}\}$
end
Find $\hat{k}_1 \in \text{argmin}_{k_1\in\{1,\dots,K\}} M(k_1)$
Return $\bar{X}_{\hat{k}_1} = \hat{\mu}_K^{(1)}$
```

Algorithm 1: A search algorithm in the mutual distance matrix of the block means $\bar{X}_1,\dots,\bar{X}_K$ for robust estimation of the mean

satisfies exactly the same statistical bound as $\hat{\mu}_K^{(0)}$ (and also the geometric median-of-means from (2)).

The algorithmic complexity of Algorithm 1 is $O(Nd + K^2 d + K^2\log(K))$ where $O(Nd)$ is the cost for computing the K empirical means $\bar{X}_1,\dots,\bar{X}_K$, $O(K^2 d)$ is the cost for computing the mutual distance matrix and $O(K^2\log K)$ is the cost for finding the median of each row of the mutual distance matrix. In particular, when $K = O(\sqrt{N}/\log(N))$, the algorithmic cost for computing $\hat{\mu}_K^{(1)}$ is of the same order as the one of the empirical mean, that is in $O(Nd)$. But, unlike the empirical mean, $\hat{\mu}_K^{(0)}$ and $\hat{\mu}_K^{(1)}$ are robust to outliers and heavy-tailed data; a property that we show in the next section.

3. Statistical guarantee for $\hat{\mu}_K^{(0)}$ and $\hat{\mu}_K^{(1)}$

The aim of this section is to show that the two estimators $\hat{\mu}_K^{(0)}$ and $\hat{\mu}_K^{(1)}$ satisfy the same theoretical bound as the geometric median-of-means from Minsker (2015).

Theorem 1 *Grant Assumption 1. Let $K \in [8|\mathcal{O}|/3, N]$. With probability at least $1 - 2^{-K/8+6}$,*

$$\left\|\hat{\mu}_K^{(0)} - \mu\right\|_2 \leq 8\sqrt{\frac{\text{Tr}(\Sigma)K}{N}} \text{ and } \left\|\hat{\mu}_K^{(1)} - \mu\right\|_2 \leq 8\sqrt{\frac{\text{Tr}(\Sigma)K}{N}}.$$

Proof. Let $\mathcal{K} = \{k \in \{1,\dots,K\} : B_k \cap \mathcal{O} = \emptyset\}$ be the set of indices of blocks of data containing no outliers. Since $K \geq 8|\mathcal{O}|/3$, we have $|\mathcal{K}| \geq K - |\mathcal{O}| \geq 5K/8$. Let $k \in \mathcal{K}$ and $\epsilon \in (0, 1/2)$. It follows from Assumption 1 and Markov inequality that with probability at least $1-\epsilon$,

$$\left\|\bar{X}_k - \mu\right\|_2 \leq \sqrt{\frac{\text{Tr}(\Sigma)K}{N\epsilon}} := R_\epsilon. \tag{6}$$

Let N_ϵ denote the number of empirical means $\bar{X}_k, k = 1,\dots,K$ that are outside of the ℓ_2-ball $B_2(\mu, R_\epsilon)$ centered at μ with radius R_ϵ. Given that $(\bar{X}_k)_{k\in\mathcal{K}}$ are independent under Assumption 1, it follows from (6) that

$$\begin{aligned}\mathbb{P}\left(N_\epsilon \geq \frac{K}{2} - 1\right) &\leq \mathbb{P}\left(\sum_{k\in\mathcal{K}} I(\left\|\bar{X}_k - \mu\right\|_2 > R_\epsilon) \geq \frac{K}{2} - 1 - |\mathcal{K}^c|\right)\\ &\leq \sum_{i=K/2-1-|\mathcal{K}^c|}^{|\mathcal{K}|} \binom{|\mathcal{K}|}{i}\epsilon^i \leq 2^{|\mathcal{K}|}\epsilon^{K/2-1-|\mathcal{K}^c|} \leq (2\epsilon)^{5K/8}\epsilon^{-1-K/2}\end{aligned}$$

where we used that $\epsilon < 1/2$ and $|\mathcal{K}| \geq 5K/8$.

Therefore, there exists an event Ω_ϵ of probability measure at least $1-(2\epsilon)^{5K/8}\epsilon^{-1-K/2}$ and a subset $A_\epsilon \subset \{1,\dots,K\}$ of cardinality at least $K/2+1$ such that on the event Ω_ϵ for all $k \in A_\epsilon$, $\left\|\bar{X}_k-\mu\right\|_2 \leq R_\epsilon$. It is then obvious that, on Ω_ϵ, for every $j \in A_\epsilon$, $\mathrm{Med}\{\left\|\bar{X}_j-\bar{X}_k\right\|_2 : 1 \leq k \leq K\} \leq 2R_\epsilon$, so that for $\hat{\mu} \in \{\hat{\mu}_K^{(0)}, \hat{\mu}_K^{(1)}\}$ we have

$$\mathrm{Med}\left\{\left\|\hat{\mu}-\bar{X}_k\right\|_2 : 1 \leq k \leq K\right\} \leq 2R_\epsilon \tag{7}$$

because

$$\mathrm{Med}\left\{\left\|\hat{\mu}-\bar{X}_k\right\|_2 : 1 \leq k \leq K\right\} \leq \min_{a \in \{\bar{X}_1,\bar{X}_2,\cdots,\bar{X}_K\}} \mathrm{Med}\left\{\left\|a-\bar{X}_k\right\|_2 : 1 \leq k \leq K\right\}.$$

Moreover, when (7) holds, if we note $A_{\hat{\mu}}$ the set containing the $\lceil K/2 \rceil$ block-means $\bar{X}_k$ that are the closest to $\hat{\mu}$ then for all $k \in A_{\hat{\mu}}$ we have $\left\|\hat{\mu}-\bar{X}_k\right\|_2 \leq 2R_\epsilon$.

By cardinality $A_\epsilon \cap A_{\hat{\mu}} \neq \emptyset$, so for $k \in A_\epsilon \cap A_{\hat{\mu}}$, we obtain

$$\|\hat{\mu}-\mu\|_2 \leq \left\|\hat{\mu}-\bar{X}_k\right\|_2 + \left\|\bar{X}_k-\mu\right\|_2 \leq 3R_\epsilon.$$

By taking for instance $\epsilon = 2^{-6}$ we get that, with probability $> 1-2^{-K/8+6}$, $\left\|\hat{\mu}-\bar{X}_k\right\|_2 \leq 8\sqrt{\mathrm{Tr}(\Sigma)K/N}$. ■

This bound, even if it has subgaussian deviations, is not optimal: in Catoni and Giulini (2017); Lugosi et al. (2019); M. Lerasle and Lecué (2017); Hopkins (2018); Cherapanamjeri et al. (2019a); Depersin and Lecué (2019) the authors achieve a bound of order of the *sub-gaussian* rate $\sqrt{\mathrm{Tr}(\Sigma)/N} + \sqrt{||\Sigma||_{op}K/N}$ (where Σ is the covariance matrix of X) with the same deviation as in Theorem 1. While, in Theorem 1 the achieved rate is the same rate as the one of the geometric median $\sqrt{\mathrm{Tr}(\Sigma)K/N}$. However the three estimators in Catoni and Giulini (2017); Lugosi et al. (2019); M. Lerasle and Lecué (2017) have not yet been proved to be computationally feasible. In Hopkins (2018), a SDP estimator achieving the optimal rate has been constructed but it is based on the Sum-of-squares approach and is therefore not computationally tractable yet. Similar observations hold for the two papers Cherapanamjeri et al. (2019a); Depersin and Lecué (2019) which construct intractable algorithms even though they are proved to run in polynomial and even (nearly) linear times. As announced in the Introduction, the main interest of Theorem 1 is that the procedure $\hat{\mu}^{(1)}$ can be efficiently implemented and that the proof of its convergence analysis is simple. The price we pay for this simplicity is on the rate of convergence since we do not recover the optimal subgaussian rate in Theorem 1.

4. Simplification of the geometric median

Following the main idea from Section 2, we show that the geometric median (2) can also be simplified in the statistical framework that we consider. The key idea is that the (possibly infinite dimensional) search space H used in (2) can be replaced by the finite set $\{\bar{X}_1, \bar{X}_2, \dots, \bar{X}_K\}$ without any loss. This yields to the procedure

$$\hat{\mu}_K^{(2)} \in \underset{a \in \{\bar{X}_1,\bar{X}_2,\dots,\bar{X}_K\}}{\mathrm{argmin}} \sum_{k=1}^{K} \left\|a-\bar{X}_k\right\|_2. \tag{8}$$

Moreover, the Weiszfeld's algorithm used to approach the geometric median can also be reduced into a simple search algorithm similar to the one from Algorithm 1.

Contrary to the Weiszfeld algorithm (which is the computation of a weighted mean, where the weights are updated at each step), one does not have to worry about convergence issues or starting point or about the pace of its convergence with Algorithm 2. Moreover, the next theorem states that, with this procedure, we do not lose anything on the statistical point of view compared with what is presented in Minsker (2015).

Theorem 2 *Grant Assumption 1. Let $K \in [8|\mathcal{O}|/3, N]$. With probability at least $1-2^{-K/8+6}$,*

$$\left\|\hat{\mu}_K^{(2)}-\mu\right\|_2 \leq 52\sqrt{\frac{\mathrm{Tr}(\Sigma)K}{N}}.$$

```
input  : $X_1, \dots, X_N$: $N$ data points in $H$, $K$: a number of blocks
output: estimator of the mean of $X$
Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$
Construct the $K$ empirical means $\bar{X}_k = (N/K)\sum_{i \in B_k} X_i$
Construct the $K \times K$ matrix $\left(\left\|\bar{X}_{k_1} - \bar{X}_{k_2}\right\|_2\right)_{1 \leq k_1, k_2 \leq K}$
for $k_1 \in \{1, \cdots, K\}$ do
    Compute $\Sigma(k_1) = \sum_{k_2=1}^K \left\|\bar{X}_{k_1} - \bar{X}_{k_2}\right\|_2$
end
Find $\hat{k}_1 \in \operatorname{argmin}_{k_1 \in \{1, \cdots, K\}} \Sigma(k_1)$
Return $\bar{X}_{\hat{k}_1} = \hat{\mu}_K^{(2)}$
```

Algorithm 2: A simple geometric median

Proof. Using the same argument as in the proof of Theorem 1, there exists an event Ω and a subset $\hat{\mathcal{K}} \subset \{1, \dots, K\}$ such that $\mathbb{P}(\Omega) \geq 1 - 2^{-K/8+6}$, $|\hat{\mathcal{K}}| \geq K/2 + 1$ and on the event Ω, for all $k \in \mathcal{K}$, $\left\|\bar{X}_k - \mu\right\|_2 \leq 8\sqrt{\operatorname{Tr}(\Sigma)K/N} := R$.

We now place ourselves on the event Ω. Let $k_0 \in \hat{\mathcal{K}}$. It follows from the definition of $\hat{\mu}_K^{(2)}$ that

$$\sum_{k=1}^K \left\|\hat{\mu}_K^{(2)} - \bar{X}_k\right\|_2 \leq \sum_{k=1}^K \left\|\bar{X}_{k_0} - \bar{X}_k\right\|_2 . \tag{9}$$

On one side, since $k_0 \in \hat{\mathcal{K}}$, we have

$$\begin{aligned} \sum_{k=1}^K \left\|\bar{X}_{k_0} - \bar{X}_k\right\|_2 &= \sum_{k \in \hat{\mathcal{K}}} \left\|\bar{X}_{k_0} - \bar{X}_k\right\|_2 + \sum_{k \in \hat{\mathcal{K}}^c} \left\|\bar{X}_{k_0} - \bar{X}_k\right\|_2 \\ &\leq 2|\hat{\mathcal{K}}|R + \sum_{k \in \hat{\mathcal{K}}^c} \left\|\bar{X}_k - \mu\right\|_2 + |\hat{\mathcal{K}}^c|R. \end{aligned} \tag{10}$$

On the other side, we have

$$\begin{aligned} \sum_{k=1}^K \left\|\hat{\mu}_K^{(2)} - \bar{X}_k\right\|_2 &= \sum_{k \in \hat{\mathcal{K}}} \left\|\hat{\mu}_K^{(2)} - \bar{X}_k\right\|_2 + \sum_{k \in \hat{\mathcal{K}}^c} \left\|\hat{\mu}_K^{(2)} - \bar{X}_k\right\|_2 \\ &\geq |\hat{\mathcal{K}}|\left(\left\|\hat{\mu}_K^{(2)} - \mu\right\|_2 - R\right) + \sum_{k \in \hat{\mathcal{K}}^c} \left\|\bar{X}_k - \mu\right\|_2 - |\hat{\mathcal{K}}^c| \left\|\hat{\mu}_K^{(2)} - \mu\right\|_2 . \end{aligned} \tag{11}$$

Combining (10) and (11) in (9), we obtain

$$\left(|\hat{\mathcal{K}}| - |\hat{\mathcal{K}}^c|\right)\left\|\hat{\mu}_K^{(2)} - \mu\right\|_2 \leq \left(3|\hat{\mathcal{K}}| + |\hat{\mathcal{K}}^c|\right)R.$$

This shows that, on the event Ω, we have

$$\left\|\hat{\mu}_K^{(2)} - \mu\right\|_2 \leq \left(\frac{3|\hat{\mathcal{K}}| + |\hat{\mathcal{K}}^c|}{|\hat{\mathcal{K}}| - |\hat{\mathcal{K}}^c|}\right) R \leq 52\sqrt{\frac{\operatorname{Tr}(\Sigma)K}{N}}$$

because $|\hat{\mathcal{K}}| \geq 5K/8$. ■

The simplification of the geometric median that we propose in this section only makes sense in a statistical framework and may be seen as an other instance of the trade-off between statistical properties and computational trade-off: given that we are asked to approximate the mean μ up to a statistical error term of the order of $\sqrt{\mathrm{Tr}(\Sigma)K/N}$ no need to find exactly the geometric median but only an approximation for it is enough. It appears that to find such an approximating solution, the entire search space H as used in the Wieszfeld algorithm can simply be replaced by a finite set of cardinality K.

Remark 1 *In dimension $d = 1$, that is for $H = \mathbb{R}$, the geometric median and $\hat{\mu}_K^{(3)}$ coincides.*

5. Robust mean estimator and algorithm adaptive to the number of outliers

Given K, the number of blocks, procedures $\hat{\mu}_K^{(1)}$ and $\hat{\mu}_K^{(2)}$ can be efficiently computed using Algorithms 1 and 2 but the choice of K has to be done beforehand and should satisfies $K \geq 8|\mathcal{O}|/3$ according to Theorem 1 and Theorem 2, where $|\mathcal{O}|$ is the number of outliers. Even though it is usually admit that most of real databases are corrupted up to 5% and so one can usually assume that $|\mathcal{O}| \leq 5\% \times N$ and therefore take $K = (4/30)N$, it is better to make no such assumption, in particular, in situations where the number of outliers is much smaller. In such cases, confidence sets can be much smaller. Another possible choice for K is to take it equal to N but it makes the confidence interval usually bigger than necessary. One way to solve this issue is to construct an estimator adaptive to K, for instance, using Lepsky's method Lepskiĭ (1990, 1991).

Let us assume that for all $K \in \{8|\mathcal{O}|/3, \ldots, N\}$, we know how to construct an estimator $\hat{\sigma}_K^2$ of $\mathrm{Tr}(\Sigma)$ such that $(1/2)\hat{\sigma}_K^2 \leq \mathrm{Tr}(\Sigma) \leq 2\hat{\sigma}_K^2$ with probability at least $1-2^{-K/8+6}$ (we provide an explicit construction of such an estimator later). In that case, for every integer $K \in \{1, \ldots, N\}$ and for $\hat{\mu}_K$ being either $\hat{\mu}_K^{(1)}$ or $\hat{\mu}_K^{(2)}$ the confidence sets

$$\hat{I}_K = B_2\left(\hat{\mu}_K, C_0\hat{\sigma}_K\sqrt{\frac{2K}{N}}\right) = \left\{a \in H : \|a - \hat{\mu}_K\|_2 \leq C_0\hat{\sigma}_K\sqrt{\frac{2K}{N}}\right\} \tag{12}$$

can be constructed from the dataset (where C_0 is either $C_0 = 8$ for $\hat{\mu}_K = \hat{\mu}_K^{(1)}$ or $C_0 = 52$ for $\hat{\mu}_K = \hat{\mu}_K^{(2)}$). Using the latter confidence sets, an adaptive choice of K can be done as follows:

$$\hat{K} = \inf\left(K \in \{1, \ldots, N\} : \bigcap_{k=K}^{N} \hat{I}_k \neq \emptyset\right). \tag{13}$$

The fully data-driven estimator $\tilde{\mu}$ that we consider is any element $\tilde{\mu} \in \bigcap_{k=\hat{K}}^N \hat{I}_k$. We now show that $\tilde{\mu}$ satisfies the same statistical bounds as the $\hat{\mu}_K$'s for all $K \in [8|\mathcal{O}|/3, N]$.

Theorem 3 *Grant Assumption 1 and assume that for all $K \in \{8|\mathcal{O}|/3, \ldots, N\}$, we know how to construct an estimator $\hat{\sigma}_K^2$ such that $(1/2)\hat{\sigma}_K^2 \leq \mathrm{Tr}(\Sigma) \leq 2\hat{\sigma}_K^2$ with probability at least $1 - 2^{-K/8+6}$. Then, for all $K \in [8|\mathcal{O}|/3, N]$, with probability at least $1 - (2^{57/8}/(2^{1/8} - 1))2^{-K/8}$,*

$$\|\tilde{\mu} - \mu\|_2 \leq 2C_0\sqrt{\frac{2\mathrm{Tr}(\Sigma)K}{N}}$$

where C_0 is either $C_0 = 8$ for $\hat{\mu}_K = \hat{\mu}_K^{(1)}$ or $C_0 = 52$ for $\hat{\mu}_K = \hat{\mu}_K^{(2)}$.

Proof. For all $K \in [8|\mathcal{O}|/3, N]$, consider the following events. Denote by $\Omega_{\Sigma,K}$ the event onto which $(1/2)\hat{\sigma}_K^2 \leq \mathrm{Tr}(\Sigma) \leq 2\hat{\sigma}_K^2$. By assumption, we have $\mathbb{P}[\Omega_{\Sigma,K}] \geq 1 - 2^{-K/8+6}$. Denote by Ω_K the event onto which $\|\hat{\mu}_K - \mu\|_2 \leq C_0\sqrt{\mathrm{Tr}(\Sigma)K/N}$. It follows from Theorem 1 and 2 that $\mathbb{P}[\Omega_K] \geq 1 - 2^{-K/8+6}$.

Let $K \in [8|\mathcal{O}|/3, N]$. On the event $\Omega^{(K)} := \bigcap_{k=K}^N \Omega_k \cap \Omega_{\Sigma,k}$, we have $\mu \in \bigcap_{k=K}^N \hat{I}_k$ because for all $k = K, \dots, N$

$$\|\hat{\mu}_k - \mu\|_2 \leq C_0 \sqrt{\frac{\mathrm{Tr}(\Sigma) k}{N}} \leq C_0 \hat{\sigma}_k \sqrt{\frac{2k}{N}},$$

meaning that $\mu \in \hat{I}_k$. Hence, by definition, $\hat{K} \leq K$ and, in particular, $\tilde{\mu} \in \hat{I}_K$, therefore, $\|\tilde{\mu} - \hat{\mu}_K\|_2 \leq C_0 \hat{\sigma}_K \sqrt{K/N}$. Moreover, $\|\hat{\mu}_K - \mu\|_2 \leq C_0 \sqrt{\mathrm{Tr}(\Sigma) K/N}$ and $\hat{\sigma}_K^2 \leq 2\mathrm{Tr}(\Sigma)$. Therefore, on the event $\Omega^{(K)}$, we have

$$\|\tilde{\mu} - \mu\|_2 \leq 2C_0 \sqrt{\frac{2\mathrm{Tr}(\Sigma) K}{N}}.$$

Finally, we have

$$\mathbb{P}[\Omega^{(K)}] \geq 1 - 2 \sum_{k=K}^N 2^{-k/8+6} \geq 1 - \left(\frac{2^{57/8}}{2^{1/8} - 1} \right) 2^{-K/8}.$$

■

Let us now construct robust estimators $\hat{\sigma}_K^2$ for all $K \in \{8|\mathcal{O}|/3, \dots, N\}$ of $\mathrm{Tr}(\Sigma)$ satisfying the required "isomorphic" property "$(1/2)\hat{\sigma}_K^2 \leq \mathrm{Tr}(\Sigma) \leq 2\hat{\sigma}_K^2$" in Theorem 3. Our starting point is to write $\mathrm{Tr}(\Sigma) = (1/2)\mathbb{E}\|X - X'\|_2^2$ when X and X' are independent random variables with mean μ and covariance matrix Σ (no need to have the same distribution). It is therefore possible to look at $\mathrm{Tr}(\Sigma)$ as the mean of a random variable $Z = \|X - X'\|_2^2/2$ and so, use robust mean-estimators such as $\hat{\mu}_K^{(1)}$ or $\mu_K^{(2)}$ to estimate $\mathbb{E}Z$.

We remark that we need two X_i's data to construct one Z_i data. We therefore need to couple data in the dataset $\{X_1, \dots, X_N\}$. For instance, we consider the coupling $\{(X_{2i}, X_{2i+1}) : i = 1, \dots, \lfloor N/2 \rfloor\}$ and for all $i = 1, \dots, \lfloor N/2 \rfloor$, we define $Z_i = \|X_{2i} - X_{2i+1}\|_2^2/2$. Next, for a given $K \in \{1, \dots, \lfloor N/2 \rfloor\}$, we consider an equipartition $B_1 \sqcup \dots \sqcup B_K$ of $\{1, \dots, \lfloor N/2 \rfloor\}$ and construct the K associated empirical means $\bar{Z}_1, \dots, \bar{Z}_K$, where $\bar{Z}_k = (1/|B_k|) \sum_{i \in B_k} Z_i$, for all $k = 1, \dots, K$. Finally, we define an estimator $\hat{\sigma}_K^2$ of $\mathrm{Tr}(\Sigma)$ as any element in

$$\underset{a \in \hat{\mathcal{Z}}_K}{\mathrm{argmin}} \ \mathrm{Med}\{|a - \bar{Z}_k| : 1 \leq k \leq K\} \text{ or } \underset{a \in \hat{\mathcal{Z}}_K}{\mathrm{argmin}} \sum_{k=1}^K |a - \bar{Z}_k|. \tag{14}$$

where $\hat{\mathcal{Z}}_K = \{\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_K\}$. Remark that if $\hat{\sigma}_K^2$ is taken in the right-hand side set in (14) then it is simply the median of $\{\bar{Z}_1, \dots, \bar{Z}_K\}$.

We now turn to the statistical analysis of $\hat{\sigma}_K^2$. We need slightly stronger assumption than in Assumption 1 to obtain an "isomorphic result" such as "$(1/2)\hat{\sigma}_K^2 \leq \mathrm{Tr}(\Sigma) \leq 2\hat{\sigma}_K^2$". In particular, we need to compare the first and second moment of the Z_i's variables such as in a Bernstein / Margin condition (see Bartlett and Mendelson (2006); Mammen and Tsybakov (1999); Tsybakov (2004)) which translates into a L_4/L_2 moment equivalence assumption on the random variables $\|X_{2i} - X_{2i+1}\|_2$ when $2i, 2i+1 \in \mathcal{I}$. Such assumption have become popular since the introduction of the small ball assumption in Koltchinskii and Mendelson (to appear) see also van de Geer and Muro (2014) and Oliveira (2016) for various examples.

Assumption 2 *There exists a partition $\mathcal{O} \sqcup \mathcal{I}$ of $\{1, \dots, N\}$ such that $(X_i)_{i \in \mathcal{I}}$ are independent and for all $i \in \mathcal{I}$, $\mathbb{E}X_i = \mu$ and $\mathbb{E}\|X_i - \mu\|_2^2 = \mathbb{E}\|X - \mu\|_2^2 = \mathrm{Tr}(\Sigma)$ where Σ is the covariance operator of X. Moreover, there exists a constant A such that $\|\|X_{2i} - X_{2i+1}\|_2\|_{L_4} \leq A \|\|X_{2i} - X_{2i+1}\|_2\|_{L_2}$ for all $i \in \{1, \dots, \lfloor N/2 \rfloor\}$ such that $2i, 2i+1 \in \mathcal{I}$.*

Theorem 4 *Grant Assumption 2 for some $A \geq 1$. Let $K \in [8|\mathcal{O}|/3, \lfloor N/2 \rfloor/(2C_0^2 A^4)]$ where $C_0 = 8$ if $\hat{\sigma}_K^2$ is taken in the left-hand side set in (14) or $C_0 = 58$ if it is taken in the right-hand side set. With probability at least $1 - 2^{-K/8+6}$,*

$$(1/2)\hat{\sigma}_K \leq \mathrm{Tr}(\Sigma) \leq 2\hat{\sigma}_K. \tag{15}$$

Proof. Let $K \in [8|\mathcal{O}|/3, \lfloor N/2 \rfloor/(2C_0^2A^4)]$. Following the same arguments as in the proof of Theorem 1 and Theorem 2, we can prove that with probability at least $1-2^{-K/8+6}$,

$$|\hat{\sigma}_K^2 - \mathrm{Tr}(\Sigma)| \leq C_0 \sigma \sqrt{\frac{K}{\lfloor N/2 \rfloor}} \tag{16}$$

where $\sigma^2 = \max_{i:2i,2i+1\in\mathcal{I}} \mathrm{Var}(Z_i)$ for $Z_i = \|X_{2i} - X_{2i+1}\|_2^2/2$.

Next, it follows from the L_4/L_2 assumption in Assumption 2 that for all $i \in \{1, \ldots, \lfloor N/2 \rfloor\}$ such that $2i, 2i{+}1 \in \mathcal{I}$, $\sqrt{\mathrm{Var}(Z_i)} \leq \|Z_i\|_{L_2} \leq A^2 \|Z_i\|_{L_1} = A^2\mathbb{E}Z_i = A^2\mathrm{Tr}(\Sigma)$. Therefore, since $K \leq \lfloor N/2 \rfloor/(2C_0^2A^4)$ then $|\hat{\sigma}_K^2 - \mathrm{Tr}(\Sigma)| \leq (1/2)\mathrm{Tr}(\Sigma)$ follows from (16). ■

Remark that we don't have the "isomorphic result" (15) for all values of K up to N but only up to $\lfloor N/2 \rfloor/(2C_0^2A^4)$. Therefore, we can only apply Theorem 3 up to $\lfloor N/2 \rfloor/(2C_0^2A^4)$ which is a minor modification of the result.

Finally, we describe a fully data-driven algorithm for the robust estimation of the mean of a random variable which does not assume the knowledge of any upper bound on $|\mathcal{O}|$ for its construction (except for the one that $8|\mathcal{O}|/3 \leq \lfloor N/2 \rfloor$). In the following we use the modified geometric median estimator from (8) for both estimation of the mean and the variance (in particular, we set $C_0 = 52$). A similar algorithm follows by using the $\hat{\mu}_K^{(1)}$ procedure from (5) instead of the simplified geometric median.

```
   input  : $X_1, \ldots, X_N$: $N$ data points in $H$
   output : robust estimator of the mean $\mu$ of $X$
   init   : $K = \lfloor N/2 \rfloor$
1  while $\|\hat{\mu}_K - \hat{\mu}_k\|_2 \leq \hat{r}_k + \hat{r}_K, k = 2K, 4K, \ldots, \lfloor N/2 \rfloor$ do
2      Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, \lfloor N/2 \rfloor\}$
3      Construct the $K$ empirical means $\bar{X}_k = (2\lfloor N/2 \rfloor/K)\sum_{i\in B_k} X_{2i} + X_{2i+1}$
4      Construct the $K \times K$ matrix of mutual distances $\left(\|\bar{X}_{k_1} - \bar{X}_{k_2}\|_2\right)_{1\leq k_1,k_2\leq K}$
5      for $k_1 \in \{1, \ldots, K\}$ do
6          Compute $\Sigma(k_1) = \sum_{k_2=1}^K \|\bar{X}_{k_1} - \bar{X}_{k_2}\|_2$
7      end
8      Find $\hat{k}_1 \in \mathrm{argmin}_{k_1\in\{1,\cdots,K\}} \Sigma(k_1)$
9      Set $\hat{\mu}_K = \bar{X}_{\hat{k}_1}$
10     Construct the $K$ empirical variances $\bar{Z}_k = (\lfloor N/2 \rfloor/K)\sum_{i\in B_k} \|X_{2i} - X_{2i+1}\|_2^2/2$
11     Construct the empirical median $\hat{\sigma}_K = \mathrm{Med}(\bar{Z}_1, \ldots, \bar{Z}_K)$
12     Set $\hat{r}_K = 52\hat{\sigma}_K\sqrt{2K/N}$
13     $K \leftarrow \lfloor K/2 \rfloor$
14 end
15 Return $\hat{\mu}_K$
```

Algorithm 3: A data-driven robust estimator of the mean adaptive to the number of outliers.

Note that instead of checking that the intersection of the Euclidean balls $\bigcap_{k=K}^{\lfloor N/2 \rfloor} B_2(\hat{\mu}_k, \hat{r}_k)$ is none empty we check that $\hat{\mu}_K \in \bigcap_{k=K+1}^{\lfloor N/2 \rfloor} B_2(\hat{\mu}_k, \hat{r}_k + \hat{r}_K)$ in step **1** of Algorithm 3, which is a way easier condition to check (in fact the first condition is algorithmically hard to check). Moreover, we consider values of K in the geometric grid $\{1, 2, 4, \ldots, \lfloor N/4 \rfloor, \lfloor N/2 \rfloor\}$ to fasten the algorithm. One can check that the result from Theorem 3 still holds with this choice of $\hat{K}$, by adjusting the proof *mutatis mutandis* and the constants by a factor ≤ 8.

6. Simulation study

We now present a simulation performed using the adaptive algorithm described in Algorithm 3 (code is available in MOMpower github page). For this simulation we chose the following setting: the variable in $\mathcal{I}$ are distributed according to a standard normal distribution in dimension $d = 20$. The size of $\mathcal{I}$ is 100000. To these data, we add $|\mathcal{O}|$ outliers where $|\mathcal{O}|$ goes from 0 to 120. We repeat each experiment 20 times. We then let our algorithm choose $\hat{K}$ by itself, and we draw, in Figure 1, the medium, the maximum and the minimum $\hat{K}$ picked by our algorithm. As our data are Gaussian, their empirical mean is well concentrated, so when there are no outliers, it is expected for $\hat{K}$ to be equal to 1 – this is indeed the case. When outliers are added, it seems natural (as our inlier data are gaussian) that $\hat{K}$ should not be more than $2\mathcal{O} + 1$, which represents the worst case scenario where there is exactly one outliers in half the $K = 2\mathcal{O}$ blocks when we set $K = 2|\mathcal{O}|$. We do find this trend in Figure 1. Here we took as outliers the vector $(10000, 10000, \cdots, 10000) \in \mathbb{R}^{20}$ repeated a number $|\mathcal{O}|$ of times.

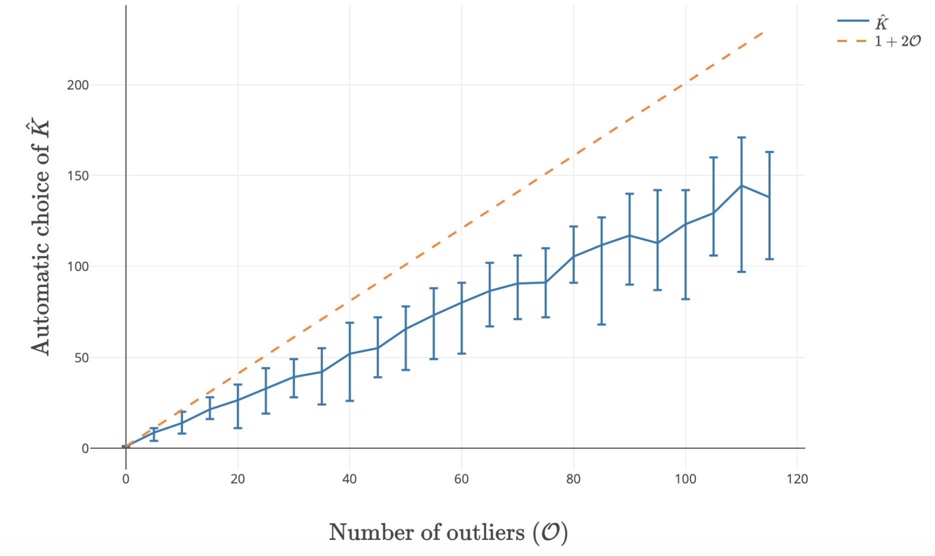

Figure 1: Adaptively selected $\hat{K}$ against the number of outliers.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147, 1999. ISSN 0022-0000. doi: 10.1006/jcss.1997.1545. URL https://doi.org/10.1006/jcss.1997.1545. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3): 311–334, 2006. ISSN 0178-8051. doi: 10.1007/s00440-005-0462-3. URL https://doi.org/10.1007/s00440-005-0462-3.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans. Inform. Theory*, 59(11):7711–7717, 2013. ISSN 0018-9448. doi: 10.1109/TIT.2013.2277869. URL https://doi.org/10.1109/TIT.2013.2277869.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'I.H.P. Probabilités et statistiques*, 48(4):1148–1185, 2012a. doi: 10.1214/11-AIHP454. URL http://www.numdam.org/item/AIHPB_2012__48_4_1148_0.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012b. ISSN 0246-0203. doi: 10.1214/11-AIHP454. URL https://doi.org/10.1214/11-AIHP454.

Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. Technical report, CNRS and LSPM, 2017.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1607. URL https://doi.org/10.1214/17-AOS1607.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates, 2019a.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806, Phoenix, USA, 25–28 Jun 2019b. PMLR. URL http://proceedings.mlr.press/v99/cherapanamjeri19b.html.

Michael B. Cohen, Yin Tat Lee, Gary L. Miller, Jakub Pachocki, and Aaron Sidford. Geometric median in nearly linear time. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 9–21, 2016. doi: 10.1145/2897518.2897647. URL https://doi.org/10.1145/2897518.2897647.

Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. Technical report, ENSAE - CREST, 2019.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016a. ISSN 0090-5364. doi: 10.1214/16-AOS1440. URL https://doi.org/10.1214/16-AOS1440.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 12 2016b. doi: 10.1214/16-AOS1440. URL https://doi.org/10.1214/16-AOS1440.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

Samuel B Hopkins. Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986. ISSN 0304-3975. doi: 10.1016/0304-3975(86)90174-X. URL https://doi.org/10.1016/0304-3975(86)90174-X.

David S Johnson and Franco P Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Notices*, to appear. arXiv:1312.3580.

O. V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35 (3):459–470, 1990. ISSN 0040-361X. doi: 10.1137/1135065. URL https://doi.org/10.1137/1135065.

O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991. ISSN 0040-361X. doi: 10.1137/1136085. URL https://doi.org/10.1137/1136085.

M. Lerasle and R. Oliveira. Robust empirical mean estimators. Technical report, IMPA and CNRS, 2011.

Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

Z. Szabo M. Lerasle, T. Matthieu and G. Lecué. Monk – outliers-robust mean embedding estimation by median-of-means. Technical report, CNRS, University of Paris 11, Ecole Polytechnique and CREST, 2017.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939240. URL https://doi.org/10.1214/aos/1017939240.

S Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. Technical report, arXiv: 1704.02658, 2017.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

A. S. Nemirovsky and D. B. and Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3-4):1175–1194, 2016. ISSN 0178-8051. doi: 10.1007/s00440-016-0738-9. URL https://doi.org/10.1007/s00440-016-0738-9.

Richard Peng and Kanat Tangwongsan. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. In *Proceedings of the twenty-fourth annual ACM symposium on Parallelism in algorithms and architectures*, pages 101–108. ACM, 2012.

Christopher G Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277, 1990.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1): 135–166, 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120131. URL https://doi.org/10.1214/aos/1079120131.

Sara van de Geer and Alan Muro. On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.*, 8(2):3031–3061, 2014. ISSN 1935-7524. doi: 10.1214/15-EJS983. URL https://doi.org/10.1214/15-EJS983.