MEAN ESTIMATION IN HIGH DIMENSION

GÁBOR LUGOSI

ABSTRACT

In this note we discuss the statistical problem of estimating the mean of a random vector based on independent, identically distributed data. This classical problem has recently attracted a lot of attention both in mathematical statistics and in theoretical computer science and numerous intricacies have been revealed. We discuss some of the recent advances, focusing on high-dimensional aspects.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 62G05; Secondary 62G35, 68Q32

KEYWORDS

Mean estimation, robustness, high-dimensional statistics

INTERNATIONAL CONGRESS
OF MATHEMATICIANS
2022 JULY 6–14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5500–5514
DOI 10.4171/ICM2022/75

1. INTRODUCTION

We consider the statistical problem of estimating the mean of a random vector based on independent, identically distributed data. This seemingly innocent classical problem has drawn renewed attention both in mathematical statistics and theoretical computer science.

The problem is formulated as follows: let $X_1, \dots, X_n$ be independent, identically distributed random vectors taking values in $\mathbb{R}^d$ such that their mean $\mu = \mathbb{E}X_1$ exists. Upon observing these random variables, one would like to estimate the vector μ. An estimator $\widehat{\mu}_n = \widehat{\mu}_n(X_1, \dots, X_n)$ is simply a measurable function of the "data" $X_1, \dots, X_n$, taking values in $\mathbb{R}^d$.

Naturally, the standard empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the first estimate that comes to mind. Indeed, the strong law of large numbers guarantees that $\overline{\mu}_n$ converges to μ almost surely without any further conditions on the distribution. However, here we are interested in the finite-sample behavior of mean estimators and for any meaningful statement one needs to make further assumptions on the distribution. Throughout this note, we assume that the covariance matrix $\Sigma = \mathbb{E}(X_1 - \mu)(X_1 - \mu)^T$ exists.

The empirical mean is known to be sensitive to "outliers" that are inevitably present in the data when the distribution may be *heavy-tailed*. This concern gave rise to the area of *robust statistics*. Classical references include Huber [19], Huber and Ronchetti [20], Hampel, Ronchetti, Rousseeuw, and Stahel [14], Tukey [44].

The quality of an estimator may be measured in various ways. While most of the early statistical work focused on expected risk measures such as the *mean-squared error*

$$\mathbb{E}\big[\|\widehat{\mu}_n - \mu\|^2\big]$$

(with $\|\cdot\|$ denoting the Euclidean norm), such risk measures may be misleading. Indeed, if the distance $\|\widehat{\mu}_n - \mu\|$ is not sufficiently concentrated, the expected value does not necessarily reflect the "typical" behavior of the error. For such reasons, estimators $\widehat{\mu}_n$ that are close to μ *with high probability* are desirable.

Thus, our aim is to understand, for any given sample size n and confidence parameter $\delta \in (0,1)$, the smallest possible value $\varepsilon = \varepsilon(n, \delta)$ such that

$$\mathbb{P}\big\{\|\widehat{\mu}_n - \mu\| > \varepsilon\big\} \le \delta.$$

In Section 2 we briefly discuss the one-dimensional case and lay out some of the basic ideas behind the more complex high-dimensional estimators. In Section 3 we present so-called sub-Gaussian estimators that guarantee the optimal order of magnitude for the accuracy $\varepsilon(n, \delta)$. Finally, in Section 4 we discuss the more refined requirement of estimators being close to the mean in each direction.

Bibliographic remark. It is beyond the scope of this note to offer an exhaustive bibliography of the topic. We refer the reader to the recent—though already somewhat outdated—survey of Lugosi and Mendelson [27].

2. BASIC IDEAS: THE ONE-DIMENSIONAL CASE

First consider the case $d = 1$, that is, when the X_i are real-valued random variables. In this case, if σ^2 denotes the variance of X_1, then the central limit theorem guarantees that the empirical mean satisfies

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mu}_n - \mu| > \frac{\sigma\Phi^{-1}(1-\delta/2)}{\sqrt{n}}\right\} = \delta,$$

where $\Phi(x) = \mathbb{P}\{G \le x\}$ is the cumulative distribution function of a standard normal random variable G. This implies the slightly loose asymptotic inequality

$$\lim_{n\to\infty} \mathbb{P}\left\{|\overline{\mu}_n - \mu| > \frac{\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right\} \le \delta.$$

Motivated by this property, we introduce a corresponding nonasymptotic notion as follows: for a given sample size n and confidence level δ, we say that a mean estimator $\widehat{\mu}_n$ is *L-sub-Gaussian* if there is a constant $L > 0$, such that, with probability at least $1 - \delta$,

$$|\widehat{\mu}_n - \mu| \le \frac{L\sigma\sqrt{\log(2/\delta)}}{\sqrt{n}}.$$

As it is pointed out in [11], if one considers the class of distributions with finite variance, the best accuracy one can hope for is of the order $\sqrt{\log(1/\delta)/n}$ and in this sense sub-Gaussian estimators are optimal. Perhaps surprisingly, sub-Gaussian estimators exist under the only assumption that the X_i have a finite second moment.

One such estimator is the so-called *median-of-means* estimator. It has been proposed in different forms in various papers, see Nemirovsky and Yudin [41], Jerrum, Valiant, and Vazirani [21], Alon, Matias, and Szegedy [1].

The definition of the median-of-means estimator calls for partitioning the data into k groups of roughly equal size, computing the empirical mean in each group, and taking the median of the obtained values.

Formally, recall that the median of k real numbers $x_1, \dots, x_k \in \mathbb{R}$ is defined as $M(x_1, \dots, x_k) = x_i$ where x_i is such that

$$\left|\{j \in [k] : x_j \le x_i\}\right| \ge \frac{k}{2} \quad \text{and} \quad \left|\{j \in [k] : x_j \ge x_i\}\right| \ge \frac{k}{2}.$$

(If several indices i fit the above description, we take the smallest one.)

Now let $1 \le k \le n$ and partition $[n] = \{1, \dots, n\}$ into k blocks $B_1, \dots, B_k$, each of size $|B_i| \ge \lfloor n/k \rfloor \ge 2$.

Given $X_1, \dots, X_n$, compute the sample mean in each block

$$Z_j = \frac{1}{|B_j|}\sum_{i\in B_j} X_i$$

and define the median-of-means estimator by $\widehat{\mu}_n = M(Z_1, \dots, Z_k)$.

To grasp intuitively why this estimator works, note that for each block, the empirical mean is an unbiased estimator of the mean, with controlled standard deviation $\sigma/\sqrt{n/k}$. Hence, the median of the distribution of the blockwise empirical mean lies within $\sigma/\sqrt{n/k}$ from the expectation. Now the empirical median is a highly concentrated estimator of this median. Now it is easy to derive the following performance bound. For simplicity, assume that n is divisible by k so that each block has $m = n/k$ elements.

Let $X_1, \dots, X_n$ be independent, identically distributed random variables with mean μ and variance σ^2. For any $\delta \in (0,1)$, if $k = \lceil 8 \log(1/\delta) \rceil$, and $n = mk$, then, with probability at least $1 - \delta$, the median-of-means estimator $\widehat{\mu}_n$ satisfies

$$|\widehat{\mu}_n - \mu| \leq \sigma \sqrt{\frac{32 \log(1/\delta)}{n}} .$$

In other words, the median-of-means estimator has a sub-Gaussian performance with $L = \sqrt{32}$ for all distributions with a finite variance.

An even more natural mean estimator is based on removing possible outliers using a truncation of X. Indeed, the so-called *trimmed-mean* (or *truncated*-mean) estimator is defined by removing a fraction of the sample, consisting of the εn largest and smallest points for some parameter $\varepsilon \in (0,1)$, and then averaging over the rest. This idea is one of the most classical tools in robust statistics, see, Tukey and McLaughlin [45], Huber and Ronchetti [20], Bickel [3], Stigler [43] for early work on the trimmed-mean estimator. The nonasymptotic sub-Gaussian property of the trimmed mean was established recently by Oliveira and Orenstein [42] who proved that if ε is chosen proportionally to $\log(1/\delta)/n$, then the trimmed-mean estimator has a sub-Gaussian performance for all distributions with a finite variance (see also [27]).

A quite different approach was introduced and analyzed by Catoni [4]. Catoni's idea is based on the fact that the empirical mean $\overline{\mu}_n$ is the solution $y \in \mathbb{R}$ of the equation

$$\sum_{i=1}^{n} (X_i - y) = 0.$$

Catoni proposed to replace the left-hand side of the equation above by another strictly decreasing function of y of the form

$$\sum_{i=1}^{n} \psi\big(\alpha(X_i - y)\big),$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is an antisymmetric increasing function and $\alpha \in \mathbb{R}$ is a parameter. The idea is that if $\psi(x)$ increases much slower than x, then the effect of "outliers" present due to heavy tails is diminished. Catoni offers a whole range of "influence" functions ψ and proves that by an appropriate choice of ψ the estimator has a sub-Gaussian performance.

We close this section by noting that in a recent work Lee and Valiant [23] construct a sub-Gaussian estimator with the (almost) optimal constant $L = \sqrt{2} + o(1)$. Their estimator builds on a clever combination of median of means, trimmed mean, and Catoni's estimator. A different approach was proposed by Minsker and Ndaoud [39]. Just like median of means, their mean estimator also starts by computing empirical averages on disjoint blocks of the data. Then they reweight the block averages in function of their empirical standard deviation. Using nontrivial properties of self-normalized sums, they obtain an estimator that is not only sub-Gaussian but it is also asymptotically efficient, in the sense that the estimator is asymptotically normal with an asymptotic variance that is as small as possible in the minimax sense.

3. MULTIVARIATE SUB-GAUSSIAN ESTIMATORS

Next we discuss the substantially more complex multivariate problem. Recall that X is a random vector taking values in $\mathbb{R}^d$ with mean $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$. Given n independent, identically distributed samples $X_1, \dots, X_n$ drawn from the distribution of X, one wishes to estimate the mean vector μ.

In order to obtain guidance of what a desirable performance is for a mean estimator, it is instructive to consider the properties of the empirical mean $\overline{\mu}_n$ when X has a multivariate normal distribution. In that case, it is not difficult to see that the Gaussian concentration inequality implies that for $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\overline{\mu}_n - \mu\| \le \sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}} .$$

where $\operatorname{Tr}(\Sigma)$ and $\lambda_{\max}$ denote the trace and spectral norm of the covariance matrix Σ. Inspired by this, we may generalize the definition of a sub-Gaussian mean estimator to the multivariate setting as follows: we say that for a given confidence level $\delta \in (0,1)$ and sample size n, a mean estimator $\widehat{\mu}_n$ is *sub-Gaussian* if there exists a constant C such that, for all distributions whose covariance matrix exists, with probability at least $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \le C\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max}\log(1/\delta)}{n}}\right). \tag{3.1}$$

Naive attempts to generalize the one-dimensional median-of-means estimator do not necessarily achieve the desired sub-Gaussian property. For example, one may define the *geometric median-of-means* estimator defined as follows (see Minsker [37], Hsu and Sabato [18], Lerasle and Oliveira [25]): we start by partitioning $[n] = \{1, \dots, n\}$ into k blocks $B_1, \dots, B_k$, each of size $|B_i| \ge \lfloor n/k \rfloor \ge 2$, where $k \sim \log(1/\delta)$. Just like in the univariate case, we compute the sample mean within each block: for $j = 1, \dots, k$, let

$$Z_j = \frac{1}{m}\sum_{i \in B_j} X_i .$$

The estimator may be defined as the geometric median of the Z_j, defined as

$$\widehat{\mu}_n = \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{j=1}^{k} \|Z_i - m\|.$$

This estimator was proposed by Minsker [37] and independently by Hsu and Sabato [18] (see also Lerasle and Oliveira [25]). Minsker [37] proved that there exists a constant C such that, whenever the covariance matrix exists, with probability at least $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \le C\sqrt{\frac{\operatorname{Tr}(\Sigma)\log(1/\delta)}{n}}.$$

This is quite nice since the inequality does not require any assumption other than the existence of the covariance matrix. However, it is not quite a sub-Gaussian bound as in (3.1). An important advantage of the geometric median-of-means estimator is that it can be computed efficiently by solving a convex optimization problem. See Cohen, Lee, Miller, Pachocki, and Sidford [8] for a recent result and for the history of the computational problem.

3.1. Median-of-means tournaments

The existence of a sub-Gaussian mean estimator was first proved by Lugosi and Mendelson [29]. Their estimator is an instance of *median-of-means tournaments* and may be defined as follows. Let $Z_1, \dots, Z_k$ be the sample means within each block exactly as above. For each $a \in \mathbb{R}^d$, let

$$T_a = \left\{x \in \mathbb{R}^d : \exists J \subset [k] : |J| \geq k/2 \text{ such that for all } j \in J,\ \|Z_j - x\| \leq \|Z_j - a\|\right\} \quad (3.2)$$

and define the mean estimator by

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \operatorname{radius}(T_a),$$

where $\operatorname{radius}(T_a) = \sup_{x \in T_a} \|x - a\|$. Thus, $\widehat{\mu}_n$ is chosen to minimize, over all $a \in \mathbb{R}^d$, the radius of the set T_a defined as the set of points $x \in \mathbb{R}^d$ for which $\|Z_j - x\| \leq \|Z_j - a\|$ for the majority of the blocks. If there are several minimizers, one may pick any one of them.

The set T_a may be seen as the set of points in $\mathbb{R}^d$ that are at least as close to the point cloud $\{Z_1, \dots, Z_k\}$ as the point a. The estimator $\widehat{\mu}_n$ is obtained by minimizing the radius of T_a. The sub-Gaussian performance of this estimator is established in [29]:

Let $X_1, \dots, X_n$ be independent, identically distributed random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. There exist constants $c, C > 0$ such that for any $\delta \in (0,1)$, if $k = c\lceil \log(1/\delta) \rceil$ and $n = mk$, then, with probability at least $1 - \delta$,

$$\|\widehat{\mu}_n - \mu\| \leq C\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{\max} \log(1/\delta)}{n}}\right).$$

An equivalent way of defining the median-of-means tournament estimator is

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \sup_{u \in S^{d-1}} \left(\operatorname{Median}\left\{\langle Z_j, u \rangle\right\}_{j \in [k]} - \langle a, u \rangle\right).$$

We may regard this as another notion of multivariate median of the block centers $Z_1, \dots, Z_k$. Unfortunately, unlike the geometric median, computing this median is hard in the sense that computing it (at least in its naive implementation) takes time exponential in the dimension d. However, Hopkins [15] introduced a semidefinite relaxation of the median-of-means tournament estimator that can be computed in time $O(nd + d \log(1/\delta)^c)$ for a dimension-independent constant c and, at the same time, achieves the desired sub-Gaussian guarantee under the only assumption that the covariance matrix exists. Subsequent improvements managed to decrease the running time further. For example, Cherapanamjeri, Flammarion, and

Bartlett [7] combined Hopkins' ideas with gradient-descent optimization to construct an sub-Gaussian mean estimator that is computable in time $O(nd + d\log(1/\delta)^2 + \log(1/\delta)^4)$. Based on ideas of "spectral reweighting" of Cheng, Diakonikolas, and Ge [6], Depersin and Lecué [9], and Lei, Luh, Venkat, and Zhang [24] further improve the running time. Hopkins, Li, and Zhang [17] show how spectral reweighting is essentially equivalent to the median notion introduced above. We refer to these papers for an exhaustive review of the rapidly growing literature of computational aspects of robust mean estimation.

3.2. Multivariate trimmed mean

Here we describe a quite different construction that also results in a sub-Gaussian mean estimator. The estimator, proposed and analyzed by Lugosi and Mendelson [31], is a multivariate version of the trimmed-mean estimator discussed in Section 2. The construction is as follows.

First split the data in two halves. For simplicity of the exposure, suppose we have $2n$ data points $X_1, \dots, X_n, Y_1, \dots, Y_n$. Set $\varepsilon = c\frac{\log(1/\delta)}{n}$ for an appropriate constant $c > 0$. For every $v \in S^{d-1}$, let α_v and β_v be the empirical $\varepsilon/2$ and $1 - \varepsilon/2$ quantiles based on the second half of the data $Y_1, \dots, Y_n$. Define

$$\phi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x > \beta, \\ x & \text{if } x \in [\alpha, \beta], \\ \alpha & \text{if } x < \alpha. \end{cases}$$

and for a parameter $Q > 0$, compute the univariate trimmed estimators

$$U_Q(v) = \frac{1}{n}\sum_{i=1}^{n} \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle X_i, v\rangle\big).$$

Each of these estimators is just the trimmed mean estimator of $\mathbb{E}\langle X, v\rangle = \langle \mu, v\rangle$ for a given direction v. Note that the trimming interval is widened by the global parameter Q whose role is to make sure that the univariate estimators work simultaneously. In order to convert the estimators of the projected means into a single vector, define the "slabs"

$$\Gamma(v, Q) = \big\{x \in \mathbb{R}^d : \big|\langle x, v\rangle - U_Q(v)\big| \le 2\varepsilon Q\big\}$$

and let

$$\Gamma(Q) = \bigcap_{v \in S^{d-1}} \Gamma(v, Q).$$

If $x \in \Gamma(Q)$, then the projection of x to every direction v is close to the trimmed mean estimator of $\langle \mu, v\rangle$. The main technical result of [31] is that, when

$$Q \sim \max\left(\frac{1}{\varepsilon}\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}, \sqrt{\frac{\lambda_1}{\varepsilon}}\right),$$

the set $\Gamma(Q)$ contains the mean μ, with probability $1 - \delta$. Since the diameter of $\Gamma(Q)$ is at most $4\varepsilon Q$, this guarantees the sub-Gaussian property of any element of the set $\Gamma(Q)$. The problem with such an estimator is that its construction requires knowledge of the correct value of Q that depends on the (unknown) covariance matrix Σ. This problem may be circumvented by a simple adaptive choice of Q: let i^* be the smallest integer such that $\bigcap_{i \geq i^*} \Gamma(2^i) \neq \emptyset$. Then define $\widehat{\mu}_n$ to be any point in the set

$$\bigcap_{i \in \mathbb{Z}: i \geq i^*} \Gamma(2^i) .$$

This choice is sufficient to guarantee the sub-Gaussian property of the estimator.

Remark. In some situations the Euclidean norm is not necessarily the most adequate way of measuring the accuracy of a mean estimator. Hence, it is natural to ask the following: given a norm $\|\cdot\|$, a confidence parameter $\delta \in (0,1)$, and an i.i.d. sample of cardinality n, what is the best possible accuracy ε for which there exists a mean estimator $\widehat{\mu}_n$ for which

$$\|\widehat{\mu}_n - \mu\| \leq \varepsilon \quad \text{with probability at least} 1-\delta?$$

The optimal order of magnitude of ε is now well understood even in this general setting, see Lugosi and Mendelson [28], Bahmani [2], Depersin and Lecué [10].

4. DIRECTION-DEPENDENT ACCURACY

An equivalent way of formulating the sub-Gaussian inequality (3.1) for a mean estimator $\widehat{\mu}_n$ is as follows: with probability at least $1-\delta$,

$$\forall u \in S^{d-1} : \langle \widehat{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\lambda_1 \log(1/\delta)}{n}} + \sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} \right), \tag{4.1}$$

where $\lambda_1 \geq \cdots \geq \lambda_d$ denote the eigenvalues of the covariance matrix Σ and $\operatorname{Tr}(\Sigma) = \sum_{i=1}^d \lambda_i$. We refer to the two terms on the right-hand side as the *weak* and *strong* terms. The strong term corresponds to a global component, while the weak term controls fluctuations in the worst direction, leading to the weak term which involves λ_1.

If one wanted to estimate the projection $\langle \mu, u \rangle$ in a fixed direction $u \in S^{d-1}$ by an estimator $\widehat{\nu}_n(u)$, as discussed in Section 2, the best accuracy one could hope for would be

$$\left| \widehat{\nu}_n(u) - \langle \mu, u \rangle \right| \leq C \sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}},$$

where $\sigma^2(u) = \operatorname{Var}(\langle X, u \rangle)$. Now it is natural to ask whether one can improve the inequality of (4.1) in a direction-sensitive way. In particular, a natural question is if the weak term on the right-hand side of (4.1) can be improved to $\sqrt{\sigma^2(u) \log(1/\delta)/n}$ and if it can, what price one has to pay in the strong term for such an improvement. This problem was studied by Lugosi and Mendelson [30] and in this section we recall the main results of that paper.

Once again, we turn to the canonical case of Gaussian vectors to obtain guidance about what kind of properties one can hope for. One can show (see [30]) that if the X_i are independent Gaussian vectors, then the empirical mean $\overline{\mu}_n$ satisfies that, with probability at least $1-\delta$,

$$\forall u \in S^{d-1} : \langle \overline{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} \right),$$

where C is a numerical constant. Thus, in the Gaussian case one can indeed obtain a weak term that scales optimally, without giving up anything in the strong term. In fact, the bound can be slightly improved to

$$\forall u \in S^{d-1} : \langle \overline{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\sum_{i>k_1} \lambda_i}{n}} \right)$$

where $k_1 = c \log(1/\delta)$, for some constant c. This bound is, in fact, the best one can hope for in the following sense:

Proposition 1. *Let $\overline{\mu}_n = (1/n) \sum_{i=1}^n X_i$ where the X_i are independent Gaussian vectors with mean μ and covariance matrix Σ. Suppose that there exists a constant C such that, for all δ, n, μ, and Σ, with probability at least $1 - \delta$,*

$$\forall u \in S^{d-1} : \langle \overline{\mu}_n - \mu, u \rangle \leq C \sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + S. \tag{4.2}$$

Then there exists a constant C' depending on C only, such that the "strong term" S has to satisfy

$$S \geq C' \sqrt{\frac{\sum_{i>k_0} \lambda_i}{n}}$$

where $k_0 = 1 + (2C + \sqrt{2})^2 \log(1/\delta)$.

The observation above shows that even in the well-behaved example of a Gaussian distribution, the strong term needs to be at least of the order

$$\sqrt{\frac{\sum_{i>k} \lambda_i}{n}}$$

where k is proportional to $\log(1/\delta)$.

The main result of [30] is that under an additional assumption on the distribution of X, one can construct an estimator that, up to the optimal strong term, preforms in every direction as if it were an optimal estimator of the one-dimensional marginal:

Let $X_1, \ldots, X_n$ be i.i.d. random vectors, taking values in $\mathbb{R}^d$, with mean μ and covariance matrix Σ whose eigenvalues are $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d \geq 0$. Suppose that there exists $q > 2$ and a constant κ such that, for all $u \in S^{d-1}$,

$$\left(\mathbb{E} \left| \langle X - \mu, u \rangle \right|^q \right)^{1/q} \leq \kappa \left(\mathbb{E} \langle X - \mu, u \rangle^2 \right)^{1/2}. \tag{4.3}$$

Then for every $\delta \in (0,1)$ there exists a mean estimator $\widehat{\mu}_n$ and constants $0 < c, c'$, $C < \infty$ (depending on κ and q only) such that, if $\delta \geq e^{-c'n}$, then, with probability, at least $1 - \delta$,

$$\forall u \in S^{d-1} : \langle \widehat{\mu}_n - \mu, u \rangle \leq C \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\sum_{i=c\log(1/\delta)}^d \lambda_i}{n}} \right). \tag{4.4}$$

Mean estimators with sub-Gaussian performance of the type (3.1) exist without assuming anything more than the existence of the covariance matrix. However, to achieve the improved direction-dependent performance formulated above, we need to assume that moments of order q exist for some $q > 2$. Moreover, we assume that the L_q norm of each one-dimensional marginal is related to the L_2-norm in a uniform manner, as described by (4.3). We call this a *norm-equivalence* condition. This condition is used repeatedly in a crucial way in the construction of the estimator. It is an intriguing question whether such a condition is necessary or if there exists a mean estimator satisfying an inequality of the type (4.4) under the only assumption of finite second moments. The mean estimator and the constants in the performance bound depend on the values κ and q of the norm-equivalence condition.

Next we describe the construction of the mean estimator. It is a quite complex variation of the trimmed mean estimator described in the previous section. In the form defined here, it is hopeless to have an algorithm that computes it efficiently, that is, in time polynomial in the sample size, the dimension, and $\log(1/\delta)$. It is an open question how far computationally efficient mean estimators can reach in terms of their statistical performance. In particular, it would be interesting to understand whether there is a true (i.e., rigorously provable) conflict between statistical accuracy and computational efficiency in the mean estimation problem. We note that in the related problem of robust mean estimation under adversarial contamination, such conflicts indeed seem to exist, see Hopkins and Li [16].

In the first step of the construction of the estimator, we divide the sample $X_1, \dots, X_n$ into n/m blocks of size m and compute, for each block

$$Y_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} X_{m(j-1)+i}.$$

Here m is chosen to be a constant depending on q and κ, the constants appearing in the norm equivalence condition. The purpose of this "smoothing" is to ensure that the Y_j satisfy certain "small-ball" properties.

Next, for each direction $u \in S^{d-1}$, we compute the trimmed-mean estimators

$$\widehat{\nu}_n(u) = \frac{1}{\sqrt{m}} \frac{1}{n/m - 2\theta n/m} \sum_{j \in [n/m] \setminus J_+(u) \cup J_-(u)} Y_j,$$

where the sets $J_+(u)$ and $J_-(u)$ correspond to the indices of the $\theta n/m$ smallest and $\theta n/m$ largest values of $\langle Y_j, u\rangle$ and $\theta \in (0, 1/2)$ is another constant that depends on q and κ only.

Now one can prove that the directional mean estimates $\widehat{\nu}_n(u)$ work as desired, simultaneously for all $u \in S^{d-1}$. More precisely, there exist constants $c, C' > 0$ depending on κ and q such that, with probability at least $1 - \delta$, for all $u \in S^{d-1}$,

$$\left|\widehat{\nu}_n(u) - \langle \mu, u\rangle\right| \le C' \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{n}} + \sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d} \lambda_i}{n}} \right).$$

Once we have the "directional" mean estimators $\widehat{\nu}_n(u)$ with the desired property, similarly to the multivariate trimmed-mean estimator discussed in Section 3 above, we need to find a vector $\widehat{\mu}_n$ such that $\langle \widehat{\mu}_n, u\rangle$ is close to $\widehat{\nu}_n(u)$ for all $u \in S^{d-1}$ (at the appropriate direction-dependent scale).

To this end, similarly to the case of the trimmed-mean estimator, we define "slabs." In order to define slabs of the correct width, we need to estimate the directional variances $\sigma^2(u)$. This is the problem of *covariance estimation* that has received quite a lot of attention, see Catoni [5], Giulini [13], Koltchinskii and Lounici [22], Lounici [26], Mendelson [35], Mendelson and Zhivotovksiy [36], Minsker [38], Minsker and Wei [40] for a sample of the relevant literature.

For our purposes, we only need to accurately estimate the variances $\sigma^2(u)$ in those directions $u \in S^{d-1}$ in which the variance is "not too small," meaning that it is above a certain critical level. Below the critical level, all we need is that the estimator detects that the variance is small. More precisely, we construct an estimator $\psi_n(u)$, such that, on an event of probability at least $1 - e^{-cn}$,

$$\begin{aligned} &\frac{1}{4}\sigma^2(u) \le \psi_n(u) \le 2\sigma^2(u) && \forall u \in S^{d-1} \text{ such that } \sigma^2(u) \ge r^2, \\ &\psi_n(u) \le Cr^2 && \text{otherwise.} \end{aligned}$$

Here c and C are constants depending on κ and q only and

$$r = \sqrt{\frac{c_0}{n} \sum_{i \ge c_0 n} \lambda_i}$$

for another constant $c_0 > 0$ depending on κ and q.

Once such a covariance estimator $\psi_n(u)$ is constructed, for a parameter $\rho > 0$, we may define the slabs

$$E_{u,\rho} = \left\{ v \in \mathbb{R}^d : \left| \widehat{v}_n(u) - \langle v, u \rangle \right| \le \rho + 2C' \sqrt{\frac{\psi_n(u) \log(1/\delta)}{n}} \right\}$$

and let

$$S_\rho = \bigcap_{u \in S^{d-1}} E_{u,\rho}.$$

Since $\rho > 0$, the set S_ρ is compact, and therefore the set

$$S = \bigcap_{\rho > 0 : S_\rho \neq \emptyset} S_\rho$$

is not empty. We may now define the mean estimator as any element $\widehat{\mu}_n \in S$. This estimator satisfies the announced property.

It remains to define the variance estimator $\psi_n(u)$. To this end, first we define

$$\tilde{X}_i = \frac{X_i - X_i'}{2}, \quad i \in [n]$$

(defined on a sample of size $2n$ that is independent of that used to construct the directional mean estimators $\widehat{v}_n(u)$) to obtain a sample of centered vectors with the same covariance as X.

Next we divide this sample into n/m equal blocks, where m is an appropriately chosen constant (depending on κ and q). For each block, we compute

$$Z_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \tilde{X}_{m(j-1)+i}.$$

The purpose of this step is to guarantee a certain "small-ball" property of the distribution, similarly to the definition of $\widehat{\nu}_n(u)$. Once again, $\psi_n(u)$ is a trimmed-mean estimator. More precisely, for every $u \in S^{d-1}$, if we denote by $J_+(u)$ the set of indices of the $\theta n/m$ largest values of $\langle Z_j, u\rangle$, we define

$$\psi_n(u) = \frac{1}{n/m} \sum_{j \in [n/m] \setminus J_+(u)} \langle Z_j, u\rangle^2.$$

The proof of the desired properties of both the directional mean estimator $\widehat{\nu}_n(u)$ and directional variance estimator $\psi_n(u)$ relies on novel bounds for the ratio of empirical and true probabilities that hold uniformly over certain classes of random variables. The main technical machinery that leads to the necessary directional control requires bounds for *ratios* of empirical and true probabilities that hold uniformly in a class of functions. Informally, one needs to control

$$\sup_{\{f \in \mathcal{F}, \|f\|_{L_2} \geq r\}} \sup_{t : \mathbb{P}\{f(X) > t\} \geq \Delta} \left| \frac{n^{-1} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) > t}}{\mathbb{P}\{f(X) > t\}} - 1 \right|$$

for appropriate values of r and Δ.

In other words, in [30] it is shown that, under minimal assumptions on the class $\mathcal{F}$, the empirical frequencies of level sets of every $f \in \mathcal{F}$ are close, in a multiplicative sense, to their true probabilities, as long as $\|f\|_{L_2} = \sqrt{\mathbb{E} f(X)^2}$ and $\mathbb{P}\{f(X) > t\}$ are large enough. Estimates of this flavor had been derived before, but only in a limited scope. Examples include the classical inequalities of Vapnik–Chervonenkis in VC theory, dealing with small classes of binary-valued functions (see also, Giné and Koltchinskii [12] for some results for real-valued classes). Existing ratio estimates are often based on the restrictive assumption that the collection of level sets, say of the form $\{\{x : f(x) > t\} : f \in \mathcal{F},\ t \geq t_0\}$, is small in the VC sense.

The method developed in [30] is based on a completely different argument that builds on the so-called *small-ball method* pioneered by Mendelson [32–34].

5. CONCLUSION

The problem of estimating the mean of a random vector has received a lot of recent attention both in mathematical statistics and in theoretical computer science. Understanding the possibilities and limitations of general mean estimation is an intriguing problem and the computational aspects enrich the area further with many nontrivial and exciting questions. In spite of the significant progress, many interesting questions remain to be explored. The lessons learnt from this prototypical statistical problem are expected to infuse other areas of statistics and machine learning with valuable ideas.

ACKNOWLEDGMENTS

Shahar Mendelson was my coauthor and main driving force behind a series of papers that form the basis of this article. Thanks Shahar for all the fun!

FUNDING

This work was supported by the Spanish Ministry of Economy and Competitiveness, Grant PGC2018-101643-B-I00 and by "Google Focused Award Algorithms and Learning for AI".

REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy, The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** (2002), 137–147.

[2] S. Bahmani, Nearly optimal robust mean estimation via empirical characteristic function. *Bernoulli* **27** (2021), no. 3, 2139–2158.

[3] P. Bickel, On some robust estimates of location. *Ann. Math. Stat.* **36** (1965), 847–858.

[4] O. Catoni, Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** (2012), no. 4, 1148–1185.

[5] O. Catoni, Pac-Bayesian bounds for the Gram matrix and least squares regression with a random design. 2016, arXiv:1603.05229.

[6] Y. Cheng, I. Diakonikolas, and R. Ge, High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the thirtieth annual ACM–SIAM symposium on discrete algorithms*, pp. 2755–2771, SIAM, 2019.

[7] Y. Cherapanamjeri, N. Flammarion, and P. Bartlett, Fast mean estimation with sub-Gaussian rates. 2019, arXiv:1902.01998.

[8] M. Cohen, Y. Lee, G. Miller, J. Pachocki, and A. Sidford, Geometric median in nearly linear time. In *Proceedings of the 48th annual ACM SIGACT symposium on theory of computing*, pp. 9–21, ACM, 2016.

[9] J. Depersin and G. Lecué, Robust subgaussian estimation of a mean vector in nearly linear time. 2019, arXiv:1906.03058.

[10] J. Depersin and G. Lecué, Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. 2021, arXiv:2102.00995.

[11] L. Devroye, M. Lerasle, G. Lugosi, and R. Oliveira, Sub-Gausssian mean estimators. *Ann. Statist.* **44** (2016), 2695–2725.

[12] E. Giné and V. Koltchinskii, Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** (2006), no. 3, 1143–1216.

[13] I. Giulini, Robust dimension-free Gram operator estimates. *Bernoulli* **24** (2018), 3864–3923.

[14] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust statistics: the approach based on influence functions*. Wiley Ser. Probab. Stat. 196, John Wiley & Sons, 1986.

[15] S. Hopkins, Sub-Gaussian mean estimation in polynomial time. *Ann. Statist.* **48** (2020), no. 2, 1193–1213.

[16] S. B. Hopkins and J. Li, How hard is robust mean estimation? In *Conference on learning theory*, pp. 1649–1682, PMLR, 2019.

[17] S. B. Hopkins, J. Li, and F. Zhang, Robust and heavy-tailed mean estimation made simple, via regret minimization. 2020, arXiv:2007.15839.

[18] D. Hsu and S. Sabato, Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** (2016), 1–40.

[19] P. Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35** (1964), no. 1, 73–101.

[20] P. Huber and E. Ronchetti, *Robust statistics*. Wiley, New York, 2009.

[21] M. Jerrum, L. Valiant, and V. Vazirani, Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** (1986), 186–188.

[22] V. Koltchinskii and K. Lounici, Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** (2017), no. 1, 110–133.

[23] J. C. Lee and P. Valiant, Optimal sub-gaussian mean estimation in R. 2020, arXiv:2011.08384.

[24] Z. Lei, K. Luh, P. Venkat, and F. Zhang, A fast spectral algorithm for mean estimation with sub-Gaussian rates. In *Conference on learning theory*, pp. 2598–2612, PMLR, 2020.

[25] M. Lerasle and R. I. Oliveira, Robust empirical mean estimators. 2012, arXiv:1112.3914.

[26] K. Lounici, High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** (2014), no. 3, 1029–1058.

[27] G. Lugosi and S. Mendelson, Mean estimation and regression under heavy-tailed distributions—a survey. *Found. Comput. Math.* **19** (2019), no. 5, 1145–1190.

[28] G. Lugosi and S. Mendelson, Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields* **175** (2019), 957–973.

[29] G. Lugosi and S. Mendelson, Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** (2019), 783–794.

[30] G. Lugosi and S. Mendelson, Multivariate mean estimation with direction-dependent accuracy. 2020, arXiv:2010.11921.

[31] G. Lugosi and S. Mendelson, Robust multivariate mean estimation: the optimality of trimmed mean. *Ann. Statist.* **49** (2021), 393–410.

[32] S. Mendelson, Learning without concentration. *J. ACM* **62** (2015), 21.

[33] S. Mendelson, An optimal unrestricted learning procedure. 2017, arXiv:1707.05342.

[34] S. Mendelson, Learning without concentration for general loss functions. *Probab. Theory Related Fields* **171** (2018), no. 1–2, 459–502.

[35] S. Mendelson, Approximating the covariance ellipsoid. *Commun. Contemp. Math.* **22** (2020), no. 08, 1950089.

[36] S. Mendelson and N. Zhivotovskiy, Robust covariance estimation under L_4–L_2 norm equivalence. 2018, arXiv:1809.10462.

[37] S. Minsker, Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** (2015), 2308–2335.

[38] S. Minsker, Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** (2018), 2871–2903.

[39] S. Minsker and M. Ndaoud, Robust and efficient mean estimation: approach based on the properties of self-normalized sums. 2020, arXiv:2006.01986.

[40] S. Minsker and X. Wei, Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli* **26** (2020), no. 1, 694–727.

[41] A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[42] R. I. Oliveira and P. Orenstein, *The sub-Gaussian property of trimmed means estimators*. Tech. rep., IMPA, 2019.

[43] S. Stigler, The asymptotic distribution of the trimmed mean. *Ann. Statist.* **1** (1973), 472–477.

[44] J. Tukey, Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, pp. 523–531, 1975.

[45] J. Tukey and D. McLaughlin, Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorization 1. *Sankhya, Ser. A* **25** (1963), 331–352.

GÁBOR LUGOSI

Department of Economics and Business, Pompeu Fabra University, ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain, and Barcelona School of Economics, Barcelona, Spain, gabor.lugosi@gmail.com