Electronic Journal of Statistics
Vol. 11 (2017) 440–451
ISSN: 1935-7524
DOI: 10.1214/17-EJS1228

On the estimation of the mean of a random vector

Emilien Joly*

MODAL'X
Université Paris Ouest
Nanterre, France;
e-mail: emilien.joly@u-paris10.fr

Gábor Lugosi†

ICREA,
Department of Economics, Pompeu Fabra University,
Barcelona Graduate School of Economics
Barcelona, Spain;
e-mail: gabor.lugosi@upf.edu

and

Roberto Imbuzeiro Oliveira‡

IMPA,
Rio de Janeiro, RJ, Brazil;
e-mail: rimfo@impa.br

Abstract: We study the problem of estimating the mean of a multivariate distribution based on independent samples. The main result is the proof of existence of an estimator with a non-asymptotic sub-Gaussian performance for all distributions satisfying some mild moment assumptions.

MSC 2010 subject classifications: Primary 62F10, 62F35; secondary 62H11.
Keywords and phrases: Point estimation, robustness and adaptive procedures, directional data, spatial statistics.

Received July 2016.

1. Introduction

Let X be a random vector taking values in $\mathbb{R}^d$. We assume throughout the paper that the mean vector $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = (X-\mu)(X-\mu)^T$

*Supported by the French Agence Nationale de la Recherche (ANR), under grant ANR-13-BS01-0005 (project SPADRO).
†Supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU.
‡Support from CNPq, Brazil via *Ciência sem Fronteiras* grant # 401572/2014-5. Supported by a *Bolsa de Produtividade em Pesquisa* from CNPq, Brazil. Supported by FAPESP Center for Neuromathematics (grant# 2013/ 07699-0, FAPESP - S. Paulo Research Foundation).

exist. Given n independent, identically distributed samples $X_1, \ldots, X_n$ drawn from the distribution of X, one wishes to estimate the mean vector.

A natural and popular choice is the sample mean $(1/n)\sum_{i=1}^n X_i$ that is known to have a near-optimal behavior whenever the distribution is sufficiently light tailed. However, whenever heavy tails are a concern, the sample mean is to be avoided as it may have a sub-optimal performance. While the one-dimensional case (i.e., $d = 1$) is quite well understood (see [3], [5]), various aspects of the multidimensional problem are still to be revealed. This paper aims at contributing to the understanding of the multi-dimensional case.

Before stating the main results, we briefly survey properties of some mean estimators of real-valued random variables. Some of these techniques serve as basic building blocks for the estimators we propose for the vector-valued case.

1.1. Estimating the mean of a real-valued random variable

When $d = 1$, the simplest and most popular mean estimator is the sample mean $\overline{\mu}_n = (1/n)\sum_{i=1}^n X_i$. The sample mean is unbiased and the central limit theorem guarantees an asymptotically Gaussian distribution. However, unless the distribution of X has a light (e.g., sub-Gaussian) tail, there are no non-asymptotic sub-Gaussian performance guarantees for $\overline{\mu}_n$. We refer the reader to Catoni [3] for details. However, perhaps surprisingly, there exist estimators of μ with much better concentration properties, see Catoni [3] and Devroye, Lerasle, Lugosi, and Oliveira [5].

A conceptually simple and quite powerful estimator is the so-called *median-of-means* estimator that has been proposed, in different forms, in various papers, see Nemirovsky and Yudin [15], Hsu [8], Jerrum, Valiant, and Vazirani [11], Alon, Matias, and Szegedy [1]. The median-of-means estimator is defined as follows. Given a positive integer b and $x_1, \ldots, x_b \in \mathbb{R}$, let $q_{1/2}$ denote the median of these numbers, that is,

$$q_{1/2}(x_1, \ldots, x_b) = x_i, \text{ where } \#\{k \in [b] : x_k \le x_i\} \ge \frac{b}{2} \text{ and}$$
$$\#\{k \in [b] : x_k \ge x_i\} \ge \frac{b}{2}.$$

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

For any fixed $\delta \in [e^{1-n/2}, 1)$, first choose $b = \lceil \ln(1/\delta) \rceil$ and note that $b \le n/2$ holds. Next, partition $[n] = \{1, \ldots, n\}$ into b blocks $B_1, \ldots, B_b$, each of size $|B_i| \ge \lfloor n/b \rfloor \ge 2$. Given $X_1, \ldots, X_n$, we compute the sample mean in each block

$$Y_i = \frac{1}{|B_i|} \sum_{j \in B_i} X_j$$

and define the median-of-means estimator by $\widehat{\mu}_n^{(\delta)} = q_{1/2}(Y_1, \ldots, Y_B)$. One can show (see, e.g., Hsu [8]) that for any $n \ge 4$,

$$\mathbb{P}\left\{ |\widehat{\mu}_n^{(\delta)} - \mu| > 2e\sqrt{2\mathrm{Var}(X)}\sqrt{\frac{(1 + \ln(1/\delta))}{n}} \right\} \le \delta \ , \tag{1}$$

where $\mathrm{Var}(X)$ denotes the variance of X.

Note that the median-of-means estimator $\widehat{\mu}_n^{(\delta)}$ does not require any knowledge of the variance of X. However, it depends on the desired confidence level δ and the partition $B_1, \dots, B_b$. Any partition satisfying $\forall i,\ |B_i| \geq \lfloor n/b \rfloor$ is valid in order to get (1). Hence, we do not keep the dependence on the partition $B_1, \dots, B_b$ in the notation $\widehat{\mu}_n^{(\delta)}$. Devroye, Lerasle, Lugosi, and Oliveira [5] introduce estimators that work for a large range of confidence levels under some mild assumptions. Catoni [3] introduces estimators of quite different flavor and gets a non-asymptotic result of the same form as (1). Bubeck, Cesa-Bianchi and Lugosi [2] apply these estimators in the context of bandit problems.

1.2. Estimating the mean of random vectors

Consider now the multi-dimensional case when $d > 1$. The sample mean $\overline{\mu}_n = (1/n)\sum_{i=1}^n X_i$ is still an obvious choice for estimating the mean vector μ.

If X has a multivariate normal distribution with mean vector μ and covariance matrix Σ, then $\overline{\mu}_n$ is also multivariate normal with mean μ and covariance matrix $(1/n)\Sigma$ and therefore, for $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}}\ , \tag{2}$$

where $\mathrm{Tr}(\Sigma)$ and $\lambda_{\max}$ denote the trace and largest eigenvalue of the covariance matrix, respectively (Hanson and Wright [7]). For non-Gaussian and possibly heavy-tailed distributions, one cannot expect such a sub-Gaussian behavior of the sample mean. The main goal of this paper is to investigate under what conditions it is possible to define mean estimators that reproduce a (non-asymptotic) sub-Gaussian performance similar to (2).

Lerasle and Oliveira [12], Hsu and Sabato [9], and Minsker [14] extend the median-of-means estimator to more general spaces. In particular, Minsker's results imply that for each $\delta \in (0,1)$ there exists a mean estimator $\widetilde{\mu}_n^{(\delta)}$ and a universal constant C such that, with probability at least $1-\delta$,

$$\|\widetilde{\mu}_n^{(\delta)} - \mu\| \leq C\sqrt{\frac{\mathrm{Tr}(\Sigma)\log(1/\delta)}{n}}\ . \tag{3}$$

While this bound is quite remarkable–note that no assumption other than the existence of the covariance matrix is made–, it does not quite achieve a sub-Gaussian performance bound that resembles (2). An instructive example is when all eigenvalues are identical and equal to $\lambda_{\max}$. If the dimension d is large, (2) is of the order of $\sqrt{(\lambda_{\max}/n)(d+\log(\delta^{-1}))}$ while (3) gives the order $\sqrt{(\lambda_{\max}/n)(d\log(\delta^{-1}))}$. The main result of this paper is the construction of a mean estimator that, under some mild moment assumptions, achieves a sub-Gaussian performance bound in the sense of (2). More precisely, we prove the following.

Theorem 1. *For all $d > 1$ and $\delta \in (0,1)$ there exists a mean estimator $\widehat{\mu}_n^{(\delta)}$ and a universal constant C such that if $X_1, \dots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ such that there exists a constant $K > 0$ such that, for all $v \in \mathbb{R}^d$ with $\|v\| = 1$,*

$$\mathbb{E}\left[\left((X-\mu)^T v\right)^4\right] \le K(v^T \Sigma v)^2 \ ,$$

then for all $n \ge CK \log d\,(d + \log(1/\delta))$,

$$\|\widehat{\mu}_n^{(\delta)} - \mu\| \le C\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{\lambda_{max} \log(\delta^{-1} \log d)}{n}}\right) \ .$$

The theorem guarantees the existence of a mean estimator whose performance matches the sub-Gaussian bound (2), up to the additional term of the order of $\sqrt{(1/n)\lambda_{\max} \log\log d}$ for all distributions satisfying the fourth-moment assumption given above. The additional term is clearly of minor importance. (For example, it is dominated by the first term whenever $\mathrm{Tr}(\Sigma) > \lambda_{\max} \log\log d$.) With the estimator we construct, this term is inevitable. On the other hand, the inequality of the theorem only holds for sample sizes that are at least a constant times $d \log d$. This feature is not desirable for truly high-dimensional problems, especially taking into account that Minsker's bound is "dimension-free".

The fourth-moment assumption can be interpreted as a boundedness assumption of the kurtosis of $(X-\mu)^T v$. The same assumption has be used in Catoni [4] and Giulini [6] for the robust estimation of the Gram matrix. The fourth-moment assumption may be weakened to an analogous "$(2+\varepsilon)$-th moment assumption" that we do not detail for the clarity of the exposition.

We prove the theorem by constructing an estimator in several steps. First we construct an estimator that performs well for "spherical" distributions (i.e., for distributions whose covariance matrix has a trace comparable to $d\lambda_{\max}$). This estimator is described in Section 2. In the second step, we decompose the space–in a data-dependent way–into the orthogonal sum of $O(\log d)$ subspaces such that all but one subspaces are such that the projection of X to the subspace has a spherical distribution. The last subspace is such that the projection has a covariance matrix with a small trace. In each subspace we apply the first estimator and combine them to obtain the final estimator $\widehat{\mu}_n^{(\delta)}$. The proof below provides an explicit value of the constant C, though no attempt has been made to optimize its value.

The constructed estimator is computationally so demanding that even for moderate values of d it is hopeless to compute it in reasonable time. In this sense, Theorem 1 should be regarded as an existence result. It is an interesting and important challenge to construct estimators with similar statistical performance that can be computed in polynomial time (as a function of n and d). Note that the estimator of Minsker cited above may be computed by solving a convex optimization problem, making it computationally feasible, see also Hsu and Sabato [9] for further computational considerations.

Before turning to the construction of the estimator, we note that in the classical literature of the problem, one traditionally measures the quality of a mean estimator $\widehat{\mu}_n$ by a notion of a *risk* such as the mean squared error $\mathbb{E}\|\widehat{\mu}_n - \mu\|^2$. In that case heavy tails are less of an issue as the mean squared error of the sample mean $\overline{\mu}_n$ equals $\mathrm{Tr}(\Sigma)/n$ as long as X has a finite second moment. Our focus on sub-Gaussian performance is motivated by numerous applications in statistical learning when one needs to estimate the means of many random variables simultaneously. In such cases, guaranteeing a small failure probability is a must. As it is well known from the celebrated Stein's "paradox," even for Gaussian distributions, the empirical mean can be outperformed by "shrinkage" estimators in terms of the mean squared error [10]. It is an interesting question for future research whether ideas of shrinkage may be used to improve estimators in our formulation as well.

2. An estimator for spherical distributions

In this section we construct an estimator that works well whenever the distribution of X is sufficiently spherical in the sense that a positive fraction of the eigenvalues of the covariance matrix is of the same order as $\lambda_{\max}$. More precisely, for $c \geq 1$, we call a distribution *c-spherical* if $d\lambda_{\max} \leq c\mathrm{Tr}(\Sigma)$.

For each $\delta \in (0,1)$ and unit vector $w \in S^{d-1}$ (where $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$), we may define $m_n^{(\delta)}(w)$ as the median-of-means estimate (as defined in Section 1.1) of $w^T\mu = \mathbb{E}w^TX$ based on the i.i.d. sample $w^TX_1, \ldots, w^TX_n$.

Let $N_{1/2} \subset S^{d-1}$ be a minimal 1/2-cover, that is, a set of smallest cardinality that has the property that for all $u \in S^{d-1}$ there exists $w \in N_{1/2}$ with $\|u - x\| \leq 1/2$. It is well known (see, e.g., [13, Lemma 13.1.1]) that $|N_{1/2}| \leq 8^d$.

Noting that $\mathrm{Var}(w^TX) \leq \lambda_{\max}$, by (1) and the union bound, we have that, with probability at least $1 - \delta$,

$$\sup_{w \in N_{1/2}} \left| m_n^{(\delta/8^d)}(w) - w^T\mu \right| \leq 2e\sqrt{2\lambda_{\max}\frac{\ln(e8^d/\delta)}{n}} \ .$$

In other words, if, for $\lambda > 0$, we define the empirical polytope

$$P_{\delta,\lambda} = \left\{ x \in \mathbb{R}^d : \sup_{w \in N_{1/2}} \left| m_n^{(\delta/8^d)}(w) - w^Tx \right| \leq 2e\sqrt{2\lambda\frac{\ln(e8^d/\delta)}{n}} \right\} \ ,$$

then with probability at least $1 - \delta$, $\mu \in P_{\delta,\lambda_{\max}}$. In particular, on this event, $P_{\delta,\lambda_{\max}}$ is nonempty. Suppose that an upper bound of the largest eigenvalue of the covariance matrix $\lambda \geq \lambda_{\max}$ is available. Then we may define the mean estimator

$$\widehat{\mu}_{n,\lambda}^{(\delta)} = \begin{cases} \text{any element } y \in P_{\delta,\lambda} & \text{if } P_{\delta,\lambda} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \ .$$

Now suppose that $\mu \in P_{\delta,\lambda}$ and let $y \in P_{\delta,\lambda}$ be arbitrary. Define $u = (y - \mu)/\|y - \mu\| \in S^{d-1}$, and let $w \in N_{1/2}$ be such that $\|w - u\| \leq 1/2$. (Such a w

exists by definition of $N_{1/2}$.) Then

$$\begin{aligned}\|y-\mu\| &= u^T(y-\mu)\\ &= (u-w)^T(y-\mu)+w^T(y-\mu) \le (1/2)\|y-\mu\| + 4e\sqrt{2\lambda\frac{\ln(e8^d/\delta)}{n}}\ ,\end{aligned}$$

where we used Cauchy-Schwarz and the fact that $y, \mu \in P_{\delta,\lambda}$. Rearranging, we obtain that, on the event that $\mu \in P_{\delta,\lambda}$,

$$\left\|\widehat{\mu}_{n,\lambda}^{(\delta)} - \mu\right\| \le 8e\sqrt{2\lambda\frac{d\ln 8 + \ln(e/\delta)}{n}}\ ,$$

provided that $\lambda \ge \lambda_{\max}$. Summarizing, we have proved the following.

Proposition 1. *Let $\lambda > 0$ and $\delta \in (0,1)$. For any distribution with mean μ and covariance matrix Σ such that $\lambda_{max} = \|\Sigma\| \le \lambda$, the estimator $\widehat{\mu}_{n,\lambda}^{(\delta)}$ defined above satisfies, with probability at least $1-\delta$,*

$$\left\|\widehat{\mu}_{n,\lambda}^{(\delta)} - \mu\right\| \le 8e\sqrt{2\lambda\frac{d\ln 8 + \ln(e/\delta)}{n}}\ .$$

In particular, if the distribution is c-spherical and $\lambda \le 2\lambda_{max}$, then

$$\left\|\widehat{\mu}_{n,\lambda}^{(\delta)} - \mu\right\| \le 16e\sqrt{\frac{c\mathrm{Tr}(\Sigma)\ln 8 + \lambda_{max}\ln(e/\delta)}{n}}\ .$$

The bound we obtained has the same sub-Gaussian form as (2), up to a multiplicative constant, whenever the distribution is c-spherical. To make the estimator fully data-dependent, we need to find an estimate $\widehat{\lambda}$ that falls in the interval $[\lambda_{\max}, 2\lambda_{\max}]$, with high probability. This may be achieved by splitting the sample in two parts of equal size (assuming n is even), estimating $\lambda_{\max}$ using samples from one part and computing the mean estimate defined above using the other part. In the next section we describe such a method as a part of a more general procedure.

3. Empirical eigendecomposition

In the previous section we presented a mean estimate that works well for "spherical" distributions. We will use this estimator as a building block in the construction of an estimator that has the desirable performance guarantee for distributions with any covariance matrix. In addition to finite covariances, we assume that there exists a constant $K > 0$ such that, for all $v \in \mathbb{R}^d$ with $\|v\| = 1$,

$$\mathbb{E}\left[\left((X-\mu)^T v\right)^4\right] \le K(v^T\Sigma v)^2\ . \tag{4}$$

In this section we assume that $n \ge 2(400e)^2 K \log_{3/2} d\big(d\log 25 + \log(2\log_{3/2} d) + \log(1/\delta)\big)$.

The basic idea is the following. We split the data into two equal halves. We use the first half in order to decompose the space into the sum of orthogonal subspaces such that the projection of X into each subspace is 4-spherical. Then we may estimate the projected means by the estimator of the previous section.

Next we describe how we obtain an orthogonal decomposition of the space based on n i.i.d. observations $X_1, \dots, X_n$.

Let $s = \lceil \log_{3/2} d^2 \rceil$ and $m = \lfloor n/s \rfloor$. Divide the sample into s blocks, each of size at least m. In what follows, we describe a way of sequentially decomposing $\mathbb{R}^d$ into the orthogonal sum of $s+1$ subspaces $\mathbb{R}^d = V_1 \oplus \cdots \oplus V_{s+1}$. First we construct V_1 using the first block $X_1, \dots, X_m$ of observations. Then we use the second block to build V_2, and so on, for s blocks. The key properties we need are that (a) the random vector X, projected to any of these subspaces has a 4-spherical distribution; (b) the largest eigenvalue of the covariance matrix of X, projected on V_i is at most $\lambda_{\max}(2/3)^{i-1}$.

To this end, just like in the previous section, let $N_\gamma \subset S^{d-1}$ be a minimal γ-cover of the unit sphere S^{d-1} for a sufficiently small constant $\gamma \in (0,1)$. The value $\gamma = 1/100$ is sufficient for our purposes and in the sequel we assume this value. Note that $|N_\gamma| \le (4/\gamma)^d$ (see [13, Lemma 13.1.1] for a proof of this fact).

Initially, we use the first block $X_1, \dots, X_m$. We may assume that m is even. Using these observations, for each $u \in N_\gamma$, we compute an estimate $V_m^{(\delta)}(u)$ of $u^T \Sigma u = \mathbb{E}(u^T(X-\mu))^2 = (1/2)\mathbb{E}(u^T(X-X'))^2$, where X' is an i.i.d. copy of X. We may construct the estimate by forming $m/2$ i.i.d. random variables $(1/2)(u^T(X_1 - X_{m/2+1}))^2, \dots, (1/2)(u^T(X_{m/2} - X_m))^2$ and estimate their mean by the median-of-means estimate $V_m^{(\delta)}(u)$ with parameter $\delta/(s(4/\gamma)^d)$. Then (1), together with assumption (4) implies that, with probability at least $1 - \delta/s$,

$$\sup_{u \in N_\gamma} \frac{\left| u^T \Sigma u - V_m^{(\delta)}(u) \right|}{u^T \Sigma u} \le 4e \sqrt{\frac{K \log(s(4/\gamma)^d/\delta)}{m}} \stackrel{\text{def.}}{=} \varepsilon_m \ .$$

Our assumptions on the sample size guarantee that $\varepsilon_m < 1/100$. The event that the inequality above holds is denoted by E_1 so that $\mathbb{P}\{E_1\} \ge 1 - \delta/s$.

Let $\mathcal{M}_{\delta,m}$ be the set of all symmetric positive semidefinite $d \times d$ matrices M satisfying

$$\sup_{u \in N_\gamma} \frac{\left| u^T M u - V_m^{(\delta)}(u) \right|}{u^T \Sigma u} \le \varepsilon_m \ .$$

By the argument above, $\Sigma \in \mathcal{M}_{\delta,m}$ on the event E_1. In particular, on E_1, $\mathcal{M}_{\delta,m}$ in non-empty. Define the estimated covariance matrix

$$\widehat{\Sigma}_m^{(\delta)} = \begin{cases} \text{any element of } \mathcal{M}_{\delta,m} & \text{if } \mathcal{M}_{\delta,m} \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

Since on E_1 both $\widehat{\Sigma}_m^{(\delta)}$ and Σ are in $\mathcal{M}_{\delta,m}$, on this event, we have

$$\left(u^T \Sigma u\right) \frac{1-\varepsilon_m}{1+\varepsilon_m} \le u^T \widehat{\Sigma}_m^{(\delta)} u \le \left(u^T \Sigma u\right) \frac{1+\varepsilon_m}{1-\varepsilon_m} \qquad \text{for all } u \in N_\gamma. \tag{5}$$

Now compute the spectral decomposition

$$\widehat{\Sigma}_m^{(\delta)} = \sum_{i=1}^{d} \widehat{\lambda}_i \widehat{v}_i \widehat{v}_i^T \ ,$$

where $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d \geq 0$ are the eigenvalues and $\widehat{v}_1, \ldots, \widehat{v}_d$ the corresponding orthogonal eigenvectors.

Let $u \in S^{d-1}$ be arbitrary and let v be a point in N_γ with smallest distance to u. Then

$$\begin{aligned} u^T \widehat{\Sigma}_m^{(\delta)} u &= v^T \widehat{\Sigma}_m^{(\delta)} v + 2(u-v)^T \widehat{\Sigma}_m^{(\delta)} v + (u-v)^T \widehat{\Sigma}_m^{(\delta)} (u-v) \\ &\leq v^T \widehat{\Sigma}_m^{(\delta)} v + \widehat{\lambda}_1 (2\gamma + \gamma^2) \qquad (6) \\ &\quad \text{(by Cauchy-Schwarz and using the fact that } \|u-v\| \leq \gamma) \\ &\leq (v^T \Sigma v) \frac{1+\varepsilon_m}{1-\varepsilon_m} + 3\gamma \widehat{\lambda}_1 \\ &\quad \text{(by (5))} \\ &\leq \frac{1+\varepsilon_m}{1-\varepsilon_m} \lambda_{\max} + 3\gamma \widehat{\lambda}_1 \ . \end{aligned}$$

In particular, on E_1 we have $\widehat{\lambda}_1 \leq \beta \lambda_{\max}$ where $\beta = \frac{1+\varepsilon_m}{1-\varepsilon_m} / (1 - 3\gamma) < 1.1$.

By a similar argument, we have that for any $u \in S^{d-1}$, if v is the point in N_γ with smallest distance to u, then on E_1,

$$u^T \Sigma u \leq (v^T \widehat{\Sigma}_m^{(\delta)} v) \frac{1+\varepsilon_m}{1-\varepsilon_m} + 3\gamma \lambda_{\max} \ \leq \frac{1+\varepsilon_m}{1-\varepsilon_m} \widehat{\lambda}_1 + 3\gamma \lambda_{\max} \ .$$

In particular, $\lambda_{\max} \leq \beta \widehat{\lambda}_1 \leq (4/3) \widehat{\lambda}_1$. Similarly,

$$\begin{aligned} u^T \Sigma u &\geq (v^T \widehat{\Sigma}_m^{(\delta)} v) \frac{1-\varepsilon_m}{1+\varepsilon_m} - 3\gamma \widehat{\lambda}_1 \\ &\geq \left(u^T \widehat{\Sigma}_m^{(\delta)} u - 3\gamma \widehat{\lambda}_1 \right) \frac{1-\varepsilon_m}{1+\varepsilon_m} - 3\gamma \widehat{\lambda}_1 \ . \\ &\geq \left(u^T \widehat{\Sigma}_m^{(\delta)} u \right) \frac{1-\varepsilon_m}{1+\varepsilon_m} - 6\gamma \widehat{\lambda}_1 \ . \qquad (7) \end{aligned}$$

Let $\widehat{d}_1$ be number of eigenvalues $\widehat{\lambda}_i$ that are at least $\widehat{\lambda}_1/2$ and let V_1 be the subspace of $\mathbb{R}^d$ spanned by $\widehat{v}_1, \ldots, \widehat{v}_{\widehat{d}_1}$. Denote by $\Pi_1(X)$ the orthogonal projection of the random variable X (independent of the X_i used to build V_1) onto V_1. Then for any $u \in V_1 \cap S^{d-1}$, on the event E_1, by (7),

$$u^T \Sigma u \geq \widehat{\lambda}_1 \frac{1}{2} \left(\frac{1-\varepsilon_m}{1+\varepsilon_m} - 12\gamma \right) \geq \frac{\widehat{\lambda}_1}{3}$$

and therefore

$$\mathbb{E} u^T (\Pi_1(X) - \mathbb{E}\Pi_1(X))(\Pi_1(X) - \mathbb{E}\Pi_1(X))^T u = u^T \Sigma u \in \left(\frac{\widehat{\lambda}_1}{3}, \frac{4\widehat{\lambda}_1}{3} \right) \ .$$

In particular, the ratio of the largest and smallest eigenvalues of the covariance matrix of $\Pi_1(X)$ is at most 4 and therefore the distribution of $\Pi_1(X)$ is 4-spherical.

On the other hand, on the event E_1, for any unit vector $u \in V_1^\perp \cap S^{d-1}$ in the orthogonal complement of V_1, we have $u^T \Sigma u \le 2\lambda_{\max}/3$. To see this, note that $u^T \widehat{\Sigma}_m^{(\delta)} u \le \widehat{\lambda}_1/2$ and therefore, denoting by v the point in N_γ closest to u,

$$\begin{aligned} u^T \Sigma u &= u^T \widehat{\Sigma}_m^{(\delta)} u + v^T \left(\Sigma - \widehat{\Sigma}_m^{(\delta)}\right) v + \left(v^T \widehat{\Sigma}_m^{(\delta)} v - u^T \widehat{\Sigma}_m^{(\delta)} u\right) + \left(u^T \Sigma u - v^T \Sigma v\right) \\ &\le \frac{\widehat{\lambda}_1}{2} + 2\varepsilon_m \lambda_{\max} + 3\gamma \widehat{\lambda}_1 + 3\gamma \lambda_{\max} \\ &\quad \text{(by (5), (6), and a similar argument for the last term)} \\ &\le \lambda_{\max} \left(\beta \left(\frac{1}{2} + 3\gamma\right) + 2\varepsilon_m + 3\gamma\right) \le \frac{2\lambda_{\max}}{3} \ . \end{aligned}$$

In other words, the largest eigenvalue of the covariance matrix of $\Pi_1^\perp(X)$ (the projection of X to the subspace $V_1^\perp$) is at most $(2/3)\lambda_{\max}$.

In the next step we construct the subspace $V_2 \subset V_1^\perp$. To this end, we proceed exactly as in the first step but now we replace $\mathbb{R}^d$ by $V_1^\perp$ and the sample $X_1, \ldots, X_m$ on the first block by the variables $\Pi_1^\perp(X_{m+1}), \ldots, \Pi_1^\perp(X_{2m}) \in V_1^\perp$. (Recall that $\Pi_1^\perp(X_i)$ is the projection of X_i to the subspace $V_1^\perp$). Just like in the first step, with probability at least $1 - \delta/s$ we obtain a (possibly empty) subspace V_2, orthogonal to V_1 such that $\Pi_2(X)$, the projection of X on V_2, has a 4-spherical distribution and largest eigenvalue of the covariance matrix of $\Pi_2^\perp(X)$ (the projection of X to the subspace $(V_1 \oplus V_2)^\perp$) is at most $(2/3)^2 \lambda_{\max}$.

We repeat the procedure s times and use a union bound the s events. We obtain, with probability at least $1 - \delta$, a sequence of subspaces $V_1, \ldots, V_s$, with the following properties:

(i) $V_1, \ldots, V_s$ are orthogonal subspaces.
(ii) For each $i = 1, \ldots, s$, $\Pi_i(X)$, the projection of X on V_i, has a 4-spherical distribution.
(iii) The largest eigenvalue of the covariance matrix of $\Pi_i(X)$ is at most $\lambda_1^{(i)} \le (2/3)^{i-1} \lambda_{\max}$.
(iv) The largest eigenvalue $\widehat{\lambda}_1^{(i)}$ of the estimated covariance matrix of $\Pi_i(X)$ satisfies
$$(3/4)\lambda_1^{(i)} \le \widehat{\lambda}_1^{(i)} \le 1.1 \lambda_1^{(i)} \ .$$

Note that it may happen for some $T < s$, we have $\mathbb{R}^d = V_1 \oplus \cdots \oplus V_T$. In that case we define $V_{T+1} = \cdots = V_s = \emptyset$.

4. Putting it all together

In this section we construct our final multivariate mean estimator and prove Theorem 1. To simplify notation, we assume that the sample size is $2n$. This only effects the value of the universal constant C in the statement of the theorem.

The data is split into two equal halves $(X_1, \dots, X_n)$ and $(X_{n+1}, \dots, X_{2n})$. The second half is used to construct the orthogonal spaces $V_1, \dots, V_s$ as described in the previous section. Let $\widehat{d}_1, \dots, \widehat{d}_s$ denote the dimension of these subspaces. Recall that, with probability at least $1 - \delta$, the construction is successful in the sense that the subspaces satisfy properties (i)–(iv) described at the end of the previous section. Denote this event by E. In the rest of the argument we condition on $(X_{n+1}, \dots, X_{2n})$ and assume that E occurs. All probabilities below are conditional.

If $\sum_{i=1}^{s} \widehat{d}_i < d$ (i.e., $V_1 \oplus \cdots \oplus V_s \neq \mathbb{R}^d$), then we define $V_{s+1} = (V_1 \oplus \cdots \oplus V_s)^{\perp}$ and denote by $\widehat{d}_{s+1} = d - \sum_{i=1}^{s} \widehat{d}_i$ the dimension of V_{s+1}. Let $\Pi_1, \dots, \Pi_{s+1}$ denote the projection operators on the subspaces $V_1, \dots, V_{s+1}$, respectively. For each $i = 1, \dots, s+1$, we use the vectors $\Pi_i(X_1), \dots, \Pi_i(X_n)$ to compute an estimator of the mean $\mathbb{E}\left[\Pi_i(X) | (X_{n+1}, \dots, X_{2n})\right] = \Pi_i(\mu)$.

For $i = 1, \dots, s$, we use the estimator defined in Section 2. In particular, within the $\widehat{d}_i$-dimensional space V_i, we compute $\overline{\mu}_i = \widehat{\mu}^{(\delta/(s+1))}_{n,(4/3)\widehat{\lambda}_i}$. Note that since $\widehat{\lambda}_i$ comes from an empirical estimation of Σ restricted to an empirical subspace V_i, $\overline{\mu}_i$ is an estimator constructed on the sample $X_1, \dots, X_n$. Then, by Proposition 1, with probability $1 - \delta/(s+1)$,

$$\|\overline{\mu}_i - \Pi_i(\mu)\|^2 \leq (8e)^2 \frac{(8/3)\widehat{\lambda}_1^{(i)} \left(\widehat{d}_i \ln 8 + \ln(e(2\log_{3/2} d + 1)/\delta) \right)}{n} \, .$$

In the last subspace V_{s+1}, we may use Minsker's estimator, based on $\Pi_{s+1}(X_1), \dots, \Pi_{s+1}(X_n)$ to compute an estimator $\overline{\mu}_{s+1} = \widetilde{\mu}^{(\delta/(s+1))}_n$ of $\Pi_{s+1}(\mu)$. Since the largest eigenvalue of the covariance matrix of $\Pi_{s+1}(X)$ is at most $\lambda_{\max}/d^2$, using (3), we obtain that, with probability $1 - \delta/(s+1)$,

$$\|\overline{\mu}_{s+1} - \Pi_{s+1}(\mu)\|^2 \leq C \frac{\lambda_{\max} \log((2\log_{3/2} d + 1)/\delta)}{n} \, .$$

Our final estimator is $\widehat{\mu}^{(\delta)}_n = \sum_{i=1}^{s+1} \overline{\mu}_{s+1}$. By the union bound, we have that, with probability at least $1 - \delta$,

$$\begin{aligned}
\left\| \widehat{\mu}^{(\delta)}_n - \mu \right\|^2 &= \sum_{i=1}^{s+1} \|\overline{\mu}_i - \Pi_i(\mu)\|^2 \\
&\leq (8e)^2 \frac{(8/3)\ln 8}{n} \sum_{i=1}^{s} \widehat{\lambda}_1^{(i)} \widehat{d}_i \\
&\quad + (8e)^2 (8/3) \frac{\ln(e(2\log_{3/2} d + 1)/\delta)}{n} \sum_{i=1}^{s} \widehat{\lambda}_1^{(i)} \\
&\quad + C \frac{\lambda_{\max} \log((2\log_{3/2} d + 1)/\delta)}{n}
\end{aligned}$$

First notice that, by properties (iii) and (iv) at the end of the previous section,

$$\sum_{i=1}^{s} \widehat{\lambda}_1^{(i)} \leq 1.1 \sum_{i=1}^{s} \lambda_1^{(i)} \leq 1.1 \lambda_{\max} \sum_{i=1}^{s} (2/3)^{i-1} \leq 3.3 \lambda_{\max} \, .$$

On the other hand, since

$$\mathrm{Tr}(\Sigma) = \mathbb{E}\|X-\mu\|^2 = \sum_{i=1}^{s+1} \mathbb{E}\|\Pi_i(X) - \Pi_i(\mu)\|^2$$

and for $i \leq s$ each $\Pi_i(X)$ has a 4-spherical distribution, we have that

$$\sum_{i=1}^{s} \widehat{\lambda}_1^{(i)} \widehat{d}_i \leq 1.1 \sum_{i=1}^{s} \lambda_1^{(i)} \widehat{d}_i \leq 4.4 \mathrm{Tr}(\Sigma) \ .$$

This concludes the proof of Theorem 1.

References

[1] Alon, N., Matias, Y. and Szegedy, M. (2002). The Space Complexity of Approximating the Frequency Moments. *Journal of Computer and System Sciences* **58** 137–147. MR1688610

[2] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory* **59** 7711–7717.

[3] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185. MR3052407

[4] Catoni, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*.

[5] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R. I. (2016). Sub-Gausssian mean estimators. *Annals of Statistics*. MR3576558

[6] Giulini, I. (2015). Robust dimension-free Gram operator estimates. *arXiv preprint arXiv:1511.06259*.

[7] Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Annals of Mathematical Statistics* **42** 1079–1083.

[8] Hsu, D. (2010). Robust statistics. http://www.inherentuncertainty.org/2010/12/robust-statistics.html.

[9] Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research* **17** 1–40. MR3491112

[10] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* **1** 361–379.

[11] Jerrum, M., Valiant, L. and Vazirani, V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science* **43** 186–188.

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

[13] Matoušek, J. (2002). *Lectures on discrete geometry*. Springer.

[14] Minsker, S. (2015). Geometric Median and Robust Estimation in Banach Spaces. *Bernoulli* **21** 2308–2335.
[15] Nemirovsky, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.