The Annals of Statistics
2018, Vol. 46, No. 6A, 2871–2903
https://doi.org/10.1214/17-AOS1642
© Institute of Mathematical Statistics, 2018

SUB-GAUSSIAN ESTIMATORS OF THE MEAN OF A RANDOM MATRIX WITH HEAVY-TAILED ENTRIES[1]

BY STANISLAV MINSKER

University of Southern California

Estimation of the covariance matrix has attracted a lot of attention of the statistical research community over the years, partially due to important applications such as principal component analysis. However, frequently used empirical covariance estimator, and its modifications, is very sensitive to the presence of outliers in the data. As P. Huber wrote [*Ann. Math. Stat.* **35** (1964) 73–101], "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one? As is now well known, the sample mean then may have a catastrophically bad performance...." Motivated by Tukey's question, we develop a new estimator of the (element-wise) mean of a random matrix, which includes covariance estimation problem as a special case. Assuming that the entries of a matrix possess only finite second moment, this new estimator admits sub-Gaussian or sub-exponential concentration around the unknown mean in the operator norm. We explain the key ideas behind our construction, and discuss applications to covariance estimation and matrix completion problems.

1. Introduction. Let $Y_1, \dots, Y_n \in \mathbb{C}^{d_1\times d_2}$ be a sequence of independent random matrices such that all their entries have finite second moments: $\mathbb{E}|(Y_j)_{k,l}|^2 < \infty$ for all $1 \le j \le n$, $1 \le k \le d_1$, $1 \le l \le d_2$. Let $\mathbb{E}Y_1, \dots, \mathbb{E}Y_n \in \mathbb{C}^{d_1\times d_2}$ be the expectations evaluated element-wise, meaning that $(\mathbb{E}Y_j)_{k,l} = \mathbb{E}(Y_j)_{k,l}$. The goal of this paper is to construct and study estimators of $\mathbb{E}\bar{Y} := \mathbb{E}[\frac{1}{n}\sum_{j=1}^n Y_j]$ under minimal assumptions on the distributions of $Y_1, \dots, Y_n$. In particular, we are interested in the estimators that admit tight nonasymptotic bounds and exponential deviation inequalities without imposing any additional assumptions (besides finite second moments) on $Y_1, \dots, Y_n$. For example, if $Y_j = Z_j Z_j^T$, where $Z_1, \dots, Z_n \in \mathbb{R}^d$ are i.i.d. copies of a random vector Z such that $\mathbb{E}Z = 0$, $\mathbb{E}[ZZ^T] = \Sigma$ and $\mathbb{E}\|Z\|_2^4 < \infty$, formulated problem is reduced to covariance estimation (here and in what follows, $\|\cdot\|_2$ and $\langle\cdot,\cdot\rangle$ stand for the usual Euclidean norm and Euclidean dot product, resp.).

Received September 2016; revised August 2017.

[1] Supported in part by NSF Grant DMS-1712956.

MSC2010 subject classifications. Primary 60B20, 62G35; secondary 62H12.

Key words and phrases. Random matrix, heavy tails, concentration inequality, covariance estimation, matrix completion.

Techniques developed in this paper have direct connection to several problems in high-dimensional statistics and statistical learning theory. In the past decade, these fields have seen numerous breakthroughs in structural estimation, concerned with a task of recovering a high-dimensional parameter that belongs to a set with "simple" structure from a small number of measurements. Examples include sparse linear regression, low-rank matrix recovery and structured covariance estimation. However, theoretical recovery guarantees for popular techniques (e.g., ℓ_1 and nuclear norm minimization) usually require strong assumptions on the underlying probability distribution, such as sub-Gaussian or bounded noise. What happens with the performance of the algorithms when these conditions are violated, which is the case for many real data sets modeled by heavy-tailed distributions? Can the assumptions be weakened without sacrificing the quality of theoretical guarantees? We look at examples where the answer is positive, and describe modifications of existing techniques that allow to achieve the improvements.

1.1. *Overview of the previous work.* Let us begin by briefly discussing a scalar version of the problem investigated in this paper. Assume that $X_1, \dots, X_n \in \mathbb{R}$ are i.i.d. copies of X, where $\mathbb{E}X^2 < \infty$. One of the fundamental problems in statistics is to construct the confidence interval for the unknown mean $\mathbb{E}X$ based on a given sample. A surprising fact (dating back to [38] where the "median of means" estimator was introduced, along with [3] and [22]) is that it is possible to construct a nonasymptotic confidence intervals $\hat{I}_n(\delta)$ with coverage probability $1 - \delta$ [meaning that $\Pr(\mathbb{E}X \in \hat{I}_n(\delta)) \geq 1 - \delta$ for given n and δ] and "nearly optimal" length $|\hat{I}_n(\delta)| \leq L\sqrt{\mathrm{Var}(X)}\sqrt{\frac{\log(e/\delta)}{n}}$, where $L > 0$ is an absolute constant. An in-depth study of this and closely related questions was performed in [11, 14] based on two different approaches. Note that the center of any such confidence interval is a point estimator $\hat{\mu} := \hat{\mu}(X_1, \dots, X_n, \delta)$ that satisfies $\Pr(|\hat{\mu} - \mathbb{E}X| \geq L\sqrt{\mathrm{Var}(X)}\sqrt{\frac{\log(e/\delta)}{n}}) \leq \delta$. Because the only assumption on X is the existence of a second moment, it is natural to call such an estimator "robust":[2] it admits strong deviation bounds even for the heavy-tailed distributions that can be used to model outliers in the data. Ideas behind these results have also been extended to empirical risk minimization methods [5, 30] which cover a wide range of statistical applications. Let us emphasize that the aforementioned estimators do not require any assumptions on the "shape" of the distribution, such as unimodality or elliptical symmetry.

Generalizations of univariate results to the case of random vectors and random matrices are not straightforward since element-wise deviation inequalities do not always translate into desired bounds. In some cases, element-wise bounds yield

[2] For the classical treatment of robust estimators based on the notion of a breakdown point, we refer the reader to [20].

inequalities for the "wrong" norm: for example, estimating each entry of the covariance matrix results in a deviation inequality for the Frobenius norm, while we are frequently interested in the bounds for the operator norm that can be much smaller. An approach which often yields "dimension-free" bounds was proposed in [18] and [35] (using generalizations of the median in higher dimensions); however, to the best of our knowledge, results of these papers are still not sufficient to obtain deviation guarantees in the operator norm that we are mainly interested in. Under more restrictive assumptions on the sequence of random matrices $Y_1, \dots, Y_n$ (such as $\|Y_j\| \le M$ almost surely for some fixed $M > 0$, $j = 1, \dots, n$, where $\|\cdot\|$ stands for the operator norm), behavior of the sample mean $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$ has been analyzed with the help of matrix concentration inequalities [1, 39, 41].

A closely related covariance matrix estimation problem has been extensively studied in the past decades. A comprehensive review is beyond the scope of this Introduction, so we will just mention few classical results and more recent work related to the current line of research. Statistical properties of the sample covariance matrix for Gaussian and sub-Gaussian observations have been investigated in detail (see [7, 8, 25, 26, 44] and references therein); under weaker moment assumptions, sample covariance estimator has been studied in [40]. Some popular robust estimators of scatter are discussed in [21], including the Minimum Covariance Determinant (MCD) estimator and the Minimum Volume Ellipsoid estimator (MVE). However, rigorous results for these estimators are available only for elliptically symmetric distributions; see [6] for results on MCD and [13] for results on MVE. Popular Maronna's [34] and Tyler's [43, 45] M-estimators of scatter also admit theoretical guarantees for the family of elliptically symmetric distributions, but we are unaware of any results extending beyond this case.

Recent papers of O. Catoni [12] and I. Guilini [17], Fan et al. [15] are closest in spirit to our work. For instance, in [12] the author constructs a robust estimator of the Gram matrix of a random vector $Z \in \mathbb{R}^d$ (as well as its covariance matrix) via estimating the quadratic form $\mathbb{E}\langle Z, u\rangle^2$ uniformly over $\|u\|_2 = 1$, and obtains error bounds for the operator norm. The latter (univariate) estimators for the quadratic form are based on the fruitful ideas originating in [11]. However, results of these works cannot be straightforwardly extended beyond covariance estimation, and are obtained under more stringent (compared to the present paper) assumptions on the underlying distribution (such as a known upper bound on the kurtosis of $\langle Z, u\rangle^2$ for any u of norm 1). In [15], authors obtain error bounds for norms other than the operator norm which is the main focus of the present paper.

Finally, let us mention that the problem of robust matrix recovery (that is discussed as an example below) has also received attention recently: for instance, the work [9, 24] investigates robust matrix completion under the "low rank + sparse" model. In [16], authors study low-rank matrix recovery under the assumption that the additive noise has only $(2 + \varepsilon)$ moments, and obtain strong results via truncation argument. We propose a different approach based on general techniques developed in this paper and achieve similar results for the matrix completion problem while requiring only the finite variance of the noise.

1.2. *Organization of the paper.* Section 2 contains definitions, notation and background material. Our main results are introduced in Section 3. After presenting core results, we discuss applications to covariance estimation and low-rank matrix completion in Section 4, and illustrate the role of various quantities involved in the general bounds through these examples. Sections 5 and 6 discuss adaptation to unknown parameters that appear in our construction, and contain longer proofs.

The Appendix contains proofs of several technical lemmas, while other details and statements not included in the main text can be found in the Supplementary Material [36].

2. Preliminaries.

In this section, we introduce the main notation and recall several useful facts from linear algebra, matrix analysis and probability theory that we rely on in the subsequent exposition.

2.1. *Definitions and notation.*

Given $A \in \mathbb{C}^{d_1 \times d_2}$, let $A^* \in \mathbb{C}^{d_2 \times d_1}$ be the Hermitian adjoint of A. If A is self-adjoint, we will write $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ for the largest and smallest eigenvalues of A. Next, we will introduce the matrix norms used in the paper.

Everywhere below, $\|\cdot\|$ stands for the operator norm $\|A\| := \sqrt{\lambda_{\max}(A^*A)}$. If $d_1 = d_2 = d$, we denote by $\operatorname{tr} A$ the trace of A. Next, for $A \in \mathbb{C}^{d_1 \times d_2}$, the nuclear norm $\|\cdot\|_1$ is defined as $\|A\|_1 = \operatorname{tr}(\sqrt{A^*A})$, where $\sqrt{A^*A}$ is a nonnegative definite matrix such that $(\sqrt{A^*A})^2 = A^*A$. The Frobenius (or Hilbert–Schmidt) norm is $\|A\|_{\mathrm{F}} = \sqrt{\operatorname{tr}(A^*A)}$, and the associated inner product is $\langle A_1, A_2 \rangle = \operatorname{tr}(A_1^* A_2)$. Finally, set $\|A\|_{\max} := \sup_{i,j} |a_{i,j}|$. For $Y \in \mathbb{C}^d$, $\|Y\|_2$ stands for the usual Euclidean norm of Y.

Given two self-adjoint matrices A and B, we will write $A \succeq B$ (or $A \succ B$) iff $A - B$ is nonnegative (or positive) definite.

Given a sequence $Y_1, \dots, Y_n$ of random matrices, $\mathbb{E}_j[\cdot]$ will stand for the conditional expectation $\mathbb{E}[\cdot | Y_1, \dots, Y_j]$.

Finally, for $a, b \in \mathbb{R}$, set $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$.

2.2. *Tools from linear algebra.*

In this section, we collect several facts from linear algebra, matrix analysis and probability theory that are frequently used in our arguments.

DEFINITION 2.1. Given a real-valued function f defined on an interval $\mathbb{T} \subseteq \mathbb{R}$ and a self-adjoint $A \in \mathbb{C}^{d \times d}$ with the eigenvalue decomposition $A = U \Lambda U^*$ such that $\lambda_j(A) \in \mathbb{T}$, $j = 1, \dots, d$, define $f(A)$ as $f(A) = U f(\Lambda) U^*$, where

$$f(\Lambda) = f\left(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}\right) = \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{pmatrix}.$$

Additionally, we will often use the following facts.

FACT 2.1. Let $A \in \mathbb{C}^{d\times d}$ be a self-adjoint matrix, and f_1, f_2 be two real-valued functions such that $f_1(\lambda_j) \ge f_2(\lambda_j)$ for $j = 1, \dots, d$. Then $f_1(A) \succeq f_2(A)$.

FACT 2.2. Let $A, B \in \mathbb{C}^{d\times d}$ be two self-adjoint matrices such that $A \succeq B$. Then $\lambda_j(A) \ge \lambda_j(B)$, $j = 1, \dots, d$, where $\lambda_j(\cdot)$ stands for the jth largest eigenvalue. Moreover, $\operatorname{tr} e^A \ge \operatorname{tr} e^B$.

FACT 2.3. The matrix logarithm is operator monotone: if $A \succ 0$, $B \succ 0$ and $A \succeq B$, then $\log(A) \succeq \log(B)$.

PROOF. See [4]. □

FACT 2.4. Let $A \in \mathbb{C}^{d\times d}$ be a self-adjoint matrix. Then $I + A + \frac{A^2}{2} \succ 0$. Moreover,

$$-\log\left(I + A + \frac{A^2}{2}\right) \preceq \log\left(I - A + \frac{A^2}{2}\right).$$

PROOF. In view of the definition of a matrix function, the first claim follows from scalar inequality $1 + t + t^2/2 > 0$ for $t \in \mathbb{R}$. Similarly, the second relation follows from the inequality $-\log(1 + t + t^2/2) \le \log(1 - t + t^2/2)$ for $t \in \mathbb{R}$. □

FACT 2.5 (Lieb's concavity theorem). Given a fixed self-adjoint matrix H, the function

$$A \mapsto \operatorname{tr}\exp\bigl(H + \log(A)\bigr)$$

is concave on the cone of positive definite matrices.

PROOF. See [31] and [42].[3] □

FACT 2.6. Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a convex function. Then $A \mapsto \operatorname{tr} f(A)$ is convex on the set of self-adjoint matrices. In particular, for any self-adjoint matrices A, B,

$$\operatorname{tr} f\left(\frac{A+B}{2}\right) \le \frac{1}{2}\operatorname{tr} f(A) + \frac{1}{2}\operatorname{tr} f(B).$$

[3] Let us mention that Lieb's theorem is one of the key tools for proving matrix concentration inequalities, and its power in this context was first demonstrated by J. Tropp [41].

PROOF. This is a consequence of Peierls inequality; see Theorem 2.9 in [10] and the comments following it. □

Finally, we introduce the Hermitian dilation which allows to reduce many problems involving general rectangular matrices to the case of Hermitian operators. Given the rectangular matrix $A \in \mathbb{C}^{d_1 \times d_2}$, the Hermitian dilation $\mathcal{H} : \mathbb{C}^{d_1 \times d_2} \mapsto \mathbb{C}^{(d_1+d_2)\times(d_1+d_2)}$ is defined as

$$\mathcal{H}(A) = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}. \tag{2.1}$$

Since $\mathcal{H}(A)^2 = \left(\begin{smallmatrix} AA^* & 0 \\ 0 & A^*A \end{smallmatrix}\right)$, it is easy to see that $\|\mathcal{H}(A)\| = \|A\|$. Another tool useful in dealing with rectangular matrices is the following lemma.

LEMMA 2.1. *Let $S \in \mathbb{C}^{d_1 \times d_1}$, $T \in \mathbb{C}^{d_2 \times d_2}$ be self-adjoint matrices, and $A \in \mathbb{C}^{d_1 \times d_2}$. Then*

$$\left\| \begin{pmatrix} S & A \\ A^* & T \end{pmatrix} \right\| \geq \left\| \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \right\|.$$

PROOF. See Section A.1 in the Appendix. □

3. Main results. Our construction has its roots in the technique proposed by O. Catoni [11] for estimating the univariate mean. Let us briefly recall the main ideas of Catoni's approach. Assume that $\xi, \xi_1, \dots, \xi_n$ is a sequence of i.i.d. random variables such that $\mathbb{E}\xi = \mu$ and $\mathrm{Var}(\xi) \leq v^2$. Catoni's estimator is defined as follows: let $\psi(x) : \mathbb{R} \mapsto \mathbb{R}$ be a nondecreasing function such that for all $x \in \mathbb{R}$,

$$-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2). \tag{3.1}$$

See Remark 1 below for examples of such functions. Given $\theta > 0$, let $\hat{\mu}_\theta$ be such that

$$\sum_{j=1}^{n} \psi(\theta(\xi_j - \hat{\mu}_\theta)) = 0 \tag{3.2}$$

(clearly, $\hat{\mu}_\theta$ always exists due to monotonicity). Set $\eta = v\sqrt{\frac{2t}{n(1-2t/n)}}$ and $\theta_* = \sqrt{\frac{2t}{n(v^2+\eta^2)}}$. Assuming that $n > 2t$, it is shown in [11] that $|\hat{\mu}_{\theta_*} - \mu| \leq \eta$ with probability $\geq 1 - 2e^{-t}$.

We proceed by presenting a multivariate extension of the estimator $\hat{\mu}_\theta$. We will first formulate main results for the self-adjoint matrices, and will later deduce the general case of rectangular matrices as a corollary. Let $Y_1, \dots, Y_n \in \mathbb{C}^{d \times d}$ be a sequence of independent self-adjoint random matrices such that

$\sigma_n^2 := \| \sum_{j=1}^n \mathbb{E}Y_j^2 \| < \infty$. Let Ψ be such that $\Psi'(x) = \psi(x)$ for all $x \in \mathbb{R}$, and set

$$\widehat{T}_\theta^* = \underset{S\in\mathbb{C}^{d\times d}, S=S^*}{\operatorname{argmin}} \left[\operatorname{tr} \sum_{j=1}^n \Psi\big(\theta(Y_j - S)\big) \right], \tag{3.3}$$

where $\theta > 0$ is an appropriate constant. It follows from Fact 2.6 that $\widehat{T}_\theta^*$ exists, moreover, it is unique if $\psi(x)$ is strictly increasing. It is also not hard to see that (3.3) is equivalent to

$$\sum_{j=1}^n \psi\big(\theta(Y_j - \widehat{T}_\theta^*)\big) = 0_{d\times d}. \tag{3.4}$$

Indeed, if $F_\psi(S) := \operatorname{tr} \sum_{j=1}^n \Psi(\theta(Y_j - S))$, then (3.4) simply states that the gradient of F_ψ evaluated at $\widehat{T}_\theta^*$ is equal to zero; see Lemma A.1 in the Appendix for more details.

To understand the properties of the estimator defined via (3.3) and (3.4), we will first consider another estimator $\widehat{T}_\theta^{(0)}$ that shares many important properties with $\widehat{T}_\theta^*$ but is easier to analyze.

The "preliminary estimator" $\widehat{T}_\theta^{(0)}$ is constructed as follows: given $\theta > 0$ and a function ψ satisfying (3.1), set $X_j := \psi(\theta Y_j)$, $j = 1, \dots, n$ and

$$\widehat{T}_\theta^{(0)} := \frac{1}{n\theta} \sum_{j=1}^n X_j. \tag{3.5}$$

In other words, $\widehat{T}_\theta^{(0)}$ is an average of "ψ-truncated" observations. Since $X_j \simeq \theta Y_j$ for small θ and a smooth function ψ, we expect that $\widehat{T}_\theta^{(0)}$ is close to $\frac{1}{n}\sum_{j=1}^n \mathbb{E}Y_j$. In the following sections, we will make this intuition more precise. In particular, we will establish the following (so far informally stated) results.

THEOREM. 1. *Assume that the observations $Y_1, \dots, Y_n$ are i.i.d. copies of $Y \in \mathbb{C}^{d\times d}$ and the parameter θ is chosen properly. Then*

$$\Pr\left(\|\widehat{T}_\theta^{(0)} - \mathbb{E}Y\| \ge \sigma\sqrt{\frac{t}{n}} \right) \le 2d \exp\left(-\frac{t}{2}\right),$$

where $\sigma^2 := \sigma_n^2/n = \|\mathbb{E}Y^2\|$.

2. *Assume that n is large enough and θ is chosen properly. Then the estimator $\widehat{T}_\theta^*$ defined via (3.4) satisfies the inequality*

$$\Pr\left(\|\widehat{T}_\theta^* - \mathbb{E}Y\| \ge C_1\sigma_0\sqrt{\frac{t}{n}} \right) \le C_2 d \exp\left(-\frac{t}{2}\right),$$

where $C_1, C_2 > 0$ are absolute constants and $\sigma_0^2 := \|\mathbb{E}(Y - \mathbb{E}Y)^2\|$.

Note that the "variance term" $\|\mathbb{E}Y^2\|$ appearing in the first part of the bound above is akin to the second moment, while in the second bound it is replaced by $\sigma_0^2 = \|\mathbb{E}(Y - \mathbb{E}Y)^2\|$; presence of the term $\|\mathbb{E}Y^2\|$ can be explained by the fact that the estimator $\widehat{T}_\theta^{(0)}$ is obtained via bias-producing truncation. We remark that in some applications, such as matrix completion discussed in Section 4, even the estimator $\widehat{T}_\theta^{(0)}$ with "suboptimal" variance term suffices to obtain good bounds.

REMARK 1. Most of our results do not depend on the concrete choice of the function ψ. One possibility is

$$\psi_1(x) = \begin{cases} \log\left(1 + x + \dfrac{x^2}{2}\right), & x \geq 0, \\ -\log\left(1 - x + \dfrac{x^2}{2}\right), & x < 0. \end{cases} \tag{3.6}$$

Another example is

$$\psi_2(x) = \begin{cases} 1/2, & x > 1, \\ x - \operatorname{sign}(x) \cdot \dfrac{x^2}{2}, & x \in [-1, 1], \\ -1/2, & x < -1. \end{cases} \tag{3.7}$$

Since the latter function is bounded, it can provide additional advantages (such as robustness) in applications. However, note that $\psi_2(x)$ does not satisfy (3.1); instead, it satisfies a slightly weaker inequality

$$-\log(1 - x + x^2) \leq \psi_2(x) \leq \log(1 + x + x^2),$$

hence all subsequent results hold for ψ_2 as well, albeit with slightly worse constant factors. We also note that both ψ_1 and ψ_2 are operator Lipschitz functions; see Lemma A.3 for details.

3.1. *Bounds for the moment generating function.* In this section, we will establish deviation inequalities for the estimator $\widehat{T}_\theta^{(0)} = \frac{1}{n\theta}\sum_{j=1}^n \psi(\theta Y_j)$. The lemma below is the cornerstone of our results. As before, given $\theta > 0$, let $X_j = \psi(\theta Y_j)$.

LEMMA 3.1. *The following inequalities hold*:

$$\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^n (X_j - \theta\mathbb{E}Y_j)\right) \leq \operatorname{tr}\exp\left(\frac{\theta^2}{2}\sum_{j=1}^n \mathbb{E}Y_j^2\right), \tag{3.8}$$

$$\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^n (\theta\mathbb{E}Y_j - X_j)\right) \leq \operatorname{tr}\exp\left(\frac{\theta^2}{2}\sum_{j=1}^n \mathbb{E}Y_j^2\right). \tag{3.9}$$

PROOF. Note that

$$
\begin{aligned}
&\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^{n}(X_j-\theta\mathbb{E}Y_j)\right)\\
&\quad=\mathbb{E}\mathbb{E}_{n-1}\operatorname{tr}\exp\left(\left[\sum_{j=1}^{n-1}(X_j-\theta\mathbb{E}Y_j)-\theta\mathbb{E}Y_n\right]+\psi(\theta Y_n)\right)\\
&\quad\le\mathbb{E}\mathbb{E}_{n-1}\operatorname{tr}\exp\left(\left[\sum_{j=1}^{n-1}(X_j-\theta\mathbb{E}Y_j)-\theta\mathbb{E}Y_n\right]+\log\bigl(I+\theta Y_n+\theta^2Y_n^2/2\bigr)\right)\\
&\quad\le\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^{n-1}(X_j-\theta\mathbb{E}Y_j)+\log\bigl(I+\theta\mathbb{E}Y_n+\theta^2\mathbb{E}Y_n^2/2\bigr)-\theta\mathbb{E}Y_n\right),
\end{aligned}
$$

where the first inequality follows from the semidefinite relation $\psi(\theta Y_n)\preceq\log(I+\theta Y_n+\frac{\theta^2}{2}Y_n^2)$ and Fact 2.2, and the second inequality follows from Lieb's concavity theorem (Fact 2.5) with $H=\sum_{j=1}^{n-1}(X_j-\theta\mathbb{E}Y_j)-\theta\mathbb{E}Y_n$ and Jensen's inequality for conditional expectation. We also note that $I+\theta\mathbb{E}Y_n+\theta^2\mathbb{E}Y_n^2/2\succ 0$ since $I+\theta Y_n+\theta^2Y_n^2/2\succ 0$ almost surely, hence $\log(I+\theta\mathbb{E}Y_n+\theta^2\mathbb{E}Y_n^2/2)$ is well defined. Repeating the steps for $X_{n-1},\dots,X_1$, we obtain the inequality

$$
\begin{aligned}
&\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^{n}(X_j-\theta\mathbb{E}Y_j)\right)\\
&\quad\le\operatorname{tr}\exp\left(\sum_{j=1}^{n}\bigl(\log\bigl(I+\theta\mathbb{E}Y_j+\theta^2\mathbb{E}Y_j^2/2\bigr)-\theta\mathbb{E}Y_j\bigr)\right).
\end{aligned}
\tag{3.10}
$$

It remains to note that by Fact 2.1 and the inequality $\log(1+x)\le x$ (that holds $\forall x>-1$), for all $j=1,\dots,n$

$$
\log\bigl(I+\theta\mathbb{E}Y_j+\theta^2\mathbb{E}Y_j^2/2\bigr)\preceq\theta\mathbb{E}Y_j+\frac{\theta^2}{2}\mathbb{E}Y_j^2,
$$

or $\log(I+\theta\mathbb{E}Y_j+\theta^2\mathbb{E}Y_j^2/2)-\theta\mathbb{E}Y_j\preceq\frac{\theta^2}{2}\mathbb{E}Y_j^2$. The first inequality (3.8) now follows from (3.10) and Fact 2.2.

To establish the second inequality of the lemma, we use the relation $-X_j=-\psi(\theta Y_j)\preceq\log(I-\theta Y_j+\frac{\theta^2}{2}Y_j^2)$ [which follows from (3.1) and Fact 2.1] together with the Fact 2.2 to deduce that

$$
\begin{aligned}
&\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^{n}(\theta\mathbb{E}Y_j-X_j)\right)\\
&\quad\le\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^{n}\bigl(\log\bigl(I+\theta(-Y_j)+\theta^2Y_j^2/2\bigr)-\theta\mathbb{E}(-Y_j)\bigr)\right),
\end{aligned}
$$

and apply inequality (3.8) to the sequence $-Y_1, \dots, -Y_n$ with

$$X_j = \log\bigl(I + \theta(-Y_j) + \theta^2(-Y_j)^2/2\bigr), \qquad j = 1, \dots, n. \qquad \square$$

We are ready to state and prove the main result of this section.

THEOREM 3.1. *Let $Y_1, \dots, Y_n \in \mathbb{C}^{d\times d}$ be a sequence of independent self-adjoint random matrices, and $\sigma_n^2 \ge \|\sum_{j=1}^n \mathbb{E}Y_j^2\|$. Then for all $\theta > 0$*

$$\Pr\biggl(\biggl\|\sum_{j=1}^n \biggl(\frac{1}{\theta}\psi(\theta Y_j) - \mathbb{E}Y_j\biggr)\biggr\| \ge t\sqrt{n}\biggr) \le 2d\exp\biggl(-\theta t\sqrt{n} + \frac{\theta^2\sigma_n^2}{2}\biggr).$$

In particular, setting $\theta = \frac{t\sqrt{n}}{\sigma_n^2}$, we get the "sub-Gaussian" tail bound $2d\exp(-\frac{t^2}{2\sigma_n^2/n})$, for a given $t > 0$. Alternatively, setting $\theta = \frac{\sqrt{n}}{\sigma_n^2}$ (independent of t), we obtain sub-exponential concentration with tail $2d\exp(-\frac{2t-1}{2\sigma_n^2/n})$ for all $t > 1/2$.

REMARK 2. In the important special case when $Y_j, j = 1, \dots, n$ are i.i.d. copies of Y, we will often use the following equivalent form of of the bound: assume that $\sigma^2 \ge \|\mathbb{E}Y^2\|$, then replacing t by $\sigma\sqrt{s}$ and setting $\theta := \sqrt{\frac{s}{n}}\frac{1}{\sigma}$ implies that

$$\Pr\biggl(\bigl\|\widehat{T}_\theta^{(0)} - \mathbb{E}Y\bigr\| \ge \sigma\sqrt{\frac{s}{n}}\biggr) \le 2d\exp(-s/2), \tag{3.11}$$

where $\widehat{T}_\theta^{(0)}$ was defined in (3.5).

PROOF. As before, set $X_j := \psi(\theta Y_j)$, $j = 1, \dots, n$. Then

$$\begin{aligned}
&\Pr\biggl(\lambda_{\max}\biggl(\frac{1}{\theta}\sum_{j=1}^n (X_j - \theta\mathbb{E}Y_j)\biggr) \ge s\biggr)\\
&\quad = \Pr\biggl(\exp\biggl(\lambda_{\max}\biggl(\sum_{j=1}^n (X_j - \theta\mathbb{E}Y_j)\biggr)\biggr) \ge e^{\theta s}\biggr)\\
&\quad \le e^{-\theta s}\mathbb{E}\operatorname{tr}\exp\biggl(\sum_{j=1}^n (X_j - \theta\mathbb{E}Y_j)\biggr) \le e^{-\theta s}\operatorname{tr}\exp\biggl(\frac{\theta^2}{2}\sum_{j=1}^n \mathbb{E}Y_j^2\biggr)\\
&\quad \le d\exp\biggl(-\theta s + \frac{\theta^2}{2}\biggl\|\sum_{j=1}^n \mathbb{E}Y_j^2\biggr\|\biggr),
\end{aligned}$$

where we used Chebyshev's inequality, the fact that $e^{\lambda_{\max}(A)} = \lambda_{\max}(e^A)$ and the inequality $\lambda_{\max}(e^A) \le \operatorname{tr} e^A$ on the second step, the first inequality of Lemma 3.1

on the third step, and the bound $\operatorname{tr} e^A \le d e^{\|A\|}$ on the last step (here and below, $A \in \mathbb{C}^{d\times d}$ is an arbitrary self-adjoint matrix). Similarly, since $-\lambda_{\min}(A) = \lambda_{\max}(-A)$, we have

$$\begin{aligned}
&\Pr\left(\lambda_{\min}\left(\frac{1}{\theta}\sum_{j=1}^{n}(X_j - \theta\mathbb{E}Y_j)\right) \le -s\right)\\
&\quad = \Pr\left(\lambda_{\max}\left(\frac{1}{\theta}\sum_{j=1}^{n}(\theta\mathbb{E}Y_j - X_j)\right) \ge s\right)\\
&\quad \le e^{-\theta s}\mathbb{E}\operatorname{tr}\exp\left(\sum_{j=1}^{n}(\theta\mathbb{E}Y_j - X_j)\right) \le e^{-\theta s}\operatorname{tr}\exp\left(\frac{\theta^2}{2}\sum_{j=1}^{n}\mathbb{E}Y_j^2\right)\\
&\quad \le d\exp\left(-\theta s + \frac{\theta^2}{2}\left\|\sum_{j=1}^{n}\mathbb{E}Y_j^2\right\|\right),
\end{aligned}$$

where we used the second inequality of Lemma 3.1 instead. The result follows by taking $s := t\sqrt{n}$ since for a self-adjoint matrix A, $\|A\| = \max(\lambda_{\max}(A), -\lambda_{\min}(A))$. □

The main weakness of the estimator $\widehat{T}_\theta^0$ discussed above is the fact that the "variance term" $\|\sum_{j=1}^n \mathbb{E}Y_j^2\|$ appearing in the bound is akin to the second moment (the price we pay for applying bias-producing truncation) while we would like to replace it by $\|\sum_{j=1}^n \mathbb{E}(Y_j - \mathbb{E}Y_j)^2\|$. This problem will be addressed in detail in Section 6. In particular, we will show the following: assume that $Y_1, \dots, Y_n$ are i.i.d. copies of Y, $\sigma_0^2 \ge \|\mathbb{E}(Y - \mathbb{E}Y)^2\|$, $\theta_0 = \sqrt{\frac{2t}{n}}\frac{1}{\sigma_0}$, and n is large enough ($n \gtrsim d^2$). Then, with exponentially high probability with respect to s, the solution $\widehat{T}_{\theta_0}^*$ of equation (3.4) satisfies $\|\widehat{T}_{\theta_0}^* - \mathbb{E}Y\| \le C\sigma_0\sqrt{\frac{s}{n}}$ for an absolute constant $C > 0$. Another problem is the fact that one needs to know the value of $\|\sum_{j=1}^n \mathbb{E}Y_j^2\|$ (or its tight upper bound) a priori to choose the "optimal" value of parameter θ. This issue and its resolution based on adaptive estimators is discussed in Section 5. We conclude this discussion with few additional comments.

REMARK 3. 1. Sub-Gaussian guarantees provided by Theorem 3.1 hold for a given confidence parameter $t > 0$ that has to be fixed a priori: in particular, the optimal value of θ depends it. However, as it was noted in [14], this is sufficient to construct (via Lepski's method [29]) estimators that admit sub-Gaussian tails uniformly over t in a certain range. We discuss the details in Section 6 of the Supplementary Material [36].

2. Let $Y_1, \dots, Y_n \in \mathbb{C}^{d\times d}$ be i.i.d. copies of Y, and $\sigma_0^2 = \|\mathbb{E}(Y - \mathbb{E}Y)^2\|$. It is interesting to compare our estimator [in particular, bound (3.11)] to the guarantees for the sample mean $\frac{1}{n}\sum_{j=1}^n Y_j$. Under an additional restrictive boundedness assumption requiring that $\|Y\| \le M$ almost surely, the noncommutative

Bernstein's inequality (see Theorem 1.4 in [41]) implies that $\|\frac{1}{n}\sum_{j=1}^n Y_j - \mathbb{E}Y\| \le 2\sigma_0\sqrt{\frac{t}{n}} \vee \frac{4}{3}\frac{Mt}{n}$ with probability $\ge 1-2de^{-t/2}$. Hence, even under additional strong assumptions our technique allows to obtain guarantees that compare favorably to the sample mean. However, as noted in [41], in the case when $\|Y\| \le M$ almost surely, the size of $\mathbb{E}\|\frac{1}{n}\sum_{j=1}^n Y_j - \mathbb{E}Y\|$ is controlled by σ_0^2 while the scale of deviations of the random variable $|\|\frac{1}{n}\sum_{j=1}^n Y_j - \mathbb{E}Y\| - \mathbb{E}\|\frac{1}{n}\sum_{j=1}^n Y_j - \mathbb{E}Y\||$ depends on the "weak variance" parameter $\sigma_*^2 = \sup_{\|v\|_2=1}\mathbb{E}\langle(Y-\mathbb{E}Y)v, v\rangle^2 \le \sigma_0^2$. It is not clear if similar improvements are achievable in the case of heavy-tailed distributions; see Remark 6 for additional comments.

3.2. *Bounds depending on the effective dimension.* The bound obtained in Theorem 3.1 explicitly depends on the dimension d of random matrices. An example is Section 3.2.1 below shows that the dimensional factor in the right-hand side of the inequality is unavoidable in general. However, it is possible to prove a similar inequality which only includes the "effective dimension" defined as

$$\bar{d} := \frac{\operatorname{tr}(\sum_{j=1}^n \mathbb{E}Y_j^2)}{\|\sum_{j=1}^n \mathbb{E}Y_j^2\|}, \tag{3.12}$$

which can be much smaller than d if $\sum_{j=1}^n \mathbb{E}Y_j^2$ has many eigenvalues that are close to 0. The following result holds.

THEOREM 3.2. *Let $Y_1, \dots, Y_n \in \mathbb{C}^{d\times d}$ be a sequence of independent self-adjoint random matrices, and $\sigma_n^2 \ge \|\sum_{j=1}^n \mathbb{E}Y_j^2\|$. Then*

$$\Pr\left(\left\|\sum_{j=1}^n\left(\frac{1}{\theta}\psi(\theta Y_j) - \mathbb{E}Y_j\right)\right\| \ge t\sqrt{n}\right) \le 2\bar{d}\left(1+\frac{1}{\theta t\sqrt{n}}\right)\exp\left(-\theta t\sqrt{n} + \frac{\theta^2\sigma_n^2}{2}\right).$$

REMARK 4. As before, we can set $\theta = \frac{t\sqrt{n}}{\sigma_n^2}$ to get

$$\Pr\left(\left\|\sum_{j=1}^n\left(\frac{1}{\theta}\psi(\theta Y_j) - \mathbb{E}Y_j\right)\right\| \ge t\sqrt{n}\right) \le 2\bar{d}\left(1+\frac{\sigma_n^2/n}{t^2}\right)\exp\left(-\frac{t^2}{2\sigma_n^2/n}\right).$$

For the values of $t \ge \sqrt{\sigma_n^2/n}$ (when the bound becomes useful), it further simplifies to

$$\Pr\left(\left\|\sum_{j=1}^n\left(\frac{1}{\theta}\psi(\theta Y_j) - \mathbb{E}Y_j\right)\right\| \ge t\sqrt{n}\right) \le 4\bar{d}\exp\left(-\frac{t^2}{2\sigma_n^2/n}\right).$$

For the "sub-exponential regime" with $\theta = \frac{\sqrt{n}}{\sigma_n^2}$, we get that for all $t \geq \frac{1}{2} \vee \sigma_n^2/n$ simultaneously,

$$\Pr\left(\left\|\sum_{j=1}^{n}\left(\frac{1}{\theta}\psi(\theta Y_j) - \mathbb{E}Y_j\right)\right\| \geq t\sqrt{n}\right) \leq 4\bar{d}\exp\left(-\frac{2t-1}{2\sigma_n^2/n}\right).$$

PROOF. The argument is similar in spirit to the proof of Theorem 3.1. Details are included in Section 3 of the Supplementary Material [36]. □

3.2.1. *Dimensional factor in Theorem* 3.1. The example below shows that the dimensional factor in Theorem 3.1 is unavoidable in general. Assume that $\psi(x) = \psi_1(x)$ as defined in (3.6), $\theta = 1$, $n = d$, and let Y_j, $j \leq d$ be independent and such that $\psi_1(Y_j) = \gamma_j e_j e_j^T$, where γ_j, $j \leq d$ are i.i.d. random variables with density $p(x) = e^{-2|x|}$, and $\{e_1, \dots, e_d\}$ is the standard Euclidean basis. Recalling that $Y_j = \psi_1^{-1}(\gamma_j)e_j e_j^T$, it is easy to check that $\mathbb{E}Y_j = 0_{d\times d}$, and that $\|\sum_{j=1}^{d}\mathbb{E}Y_j^2\| = \mathbb{E}(\psi_1^{-1}(\gamma_1))^2 < \infty$. Theorem 3.1 implies that

$$\Pr\left(\left\|\sum_{j=1}^{d}\gamma_j e_j e_j^T\right\| \geq s\right) \leq f(d)e^{-s}$$

with $f(d) \leq Cd$ for some absolute constant C. Since $\|\sum_{j=1}^{d}\gamma_j e_j e_j^T\| = \max(|\gamma_1|, \dots, |\gamma_d|)$, it follows from Lemma 7.2 of the Supplementary Material [36] that $\Pr(\|\sum_{j=1}^{d}\gamma_j e_j e_j^T\| \geq (\frac{1}{2} - \tau)\log d) \geq c(\tau)$ for any $0 < \tau < 1/2$ and some constant $c(\tau) > 0$. This shows that the dimensional factor $f(d)$ cannot grow slower than $d^{1/2-\tau}$ for any $\tau > 0$.

3.3. *Bounds for arbitrary rectangular matrices.* In this section, we will deduce results for arbitrary matrices from the bounds for self-adjoint operators. Let $Y_1, \dots, Y_n \in \mathbb{C}^{d_1\times d_2}$ be independent, and assume that

$$\sigma_n^2 \geq \max\left(\left\|\sum_{j=1}^{n}\mathbb{E}Y_j Y_j^*\right\|, \left\|\sum_{j=1}^{n}\mathbb{E}Y_j^* Y_j\right\|\right).$$

Given $\theta > 0$, set $X_j := \psi(\theta\mathcal{H}(Y_j))$ [where $\mathcal{H}(\cdot)$ is the self-adjoint dilation, see equation (2.1)] and define $\widehat{T} \in \mathbb{C}^{(d_1+d_2)\times(d_1+d_2)}$ as

$$\widehat{T} := \widehat{T}(\theta) = \sum_{j=1}^{n}\frac{1}{\theta}X_j.$$

Let $\hat{T}_{11} \in \mathbb{C}^{d_1\times d_1}$, $\hat{T}_{22} \in \mathbb{C}^{d_2\times d_2}$, $\hat{T}_{12} \in \mathbb{C}^{d_1\times d_2}$ be such that $\widehat{T} = \begin{pmatrix}\hat{T}_{11} & \hat{T}_{12}\\ \hat{T}_{12}^* & \hat{T}_{22}\end{pmatrix}$. Since $\widehat{T}$ is "close" to $\sum_{j=1}^{n}\mathcal{H}(\mathbb{E}Y_j)$ for the proper choice of θ, it is natural to expect that $\hat{T}_{12}$ is close to $\sum_{j=1}^{n}\mathbb{E}Y_j$.

COROLLARY 3.1. *Under the assumptions stated above*,

$$\Pr\left(\left\|\hat{T}_{12}-\sum_{j=1}^{n}\mathbb{E}Y_j\right\|\ge t\sqrt{n}\right)\le 2(d_1+d_2)\exp\left(-\theta t\sqrt{n}+\frac{\theta^2\sigma_n^2}{2}\right)$$

and

$$\Pr\left(\left\|\hat{T}_{12}-\sum_{j=1}^{n}\mathbb{E}Y_j\right\|\ge t\sqrt{n}\right)\le 2\bar{d}\left(1+\frac{1}{\theta t\sqrt{n}}\right)\exp\left(-\theta t\sqrt{n}+\frac{\theta^2\sigma_n^2}{2}\right),$$

where $\bar{d}=2\frac{\operatorname{tr}(\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j)}{\|\sum_{j=1}^{n}\mathbb{E}Y_jY_j^*\|\vee\|\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j\|}$.

PROOF. Note that

$$\left\|\sum_{j=1}^{n}\mathbb{E}\mathcal{H}(Y_j)^2\right\|=\max\left(\left\|\sum_{j=1}^{n}\mathbb{E}Y_jY_j^*\right\|,\left\|\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j\right\|\right)\le\sigma_n^2.$$

Theorem 3.1 applied to self-adjoint random matrices $\mathcal{H}(Y_j)\in\mathbb{C}^{(d_1+d_2)\times(d_1+d_2)}$, $j=1,\dots,n$ implies that $\|\widehat{T}-\sum_{j=1}^{n}\mathcal{H}(\mathbb{E}Y_j)\|\le t\sqrt{n}$ with probability $\ge 1-2(d_1+d_2)\exp(-\theta t\sqrt{n}+\frac{\theta^2\sigma_n^2}{2})$. It remains to apply Lemma 2.1:

$$\begin{aligned}\left\|\widehat{T}-\sum_{j=1}^{n}\mathcal{H}(\mathbb{E}Y_j)\right\|&=\left\|\begin{pmatrix}\hat{T}_{11} & \hat{T}_{12}-\sum_{j=1}^{n}\mathbb{E}Y_j\\ \hat{T}_{12}^*-\sum_{j=1}^{n}\mathbb{E}Y_j^* & \hat{T}_{22}\end{pmatrix}\right\|\\&\ge\left\|\begin{pmatrix}0 & \hat{T}_{12}-\sum_{j=1}^{n}\mathbb{E}Y_j\\ \hat{T}_{12}^*-\sum_{j=1}^{n}\mathbb{E}Y_j^* & 0\end{pmatrix}\right\|=\left\|\hat{T}_{12}-\sum_{j=1}^{n}\mathbb{E}Y_j\right\|,\end{aligned}$$

and the first inequality follows. To obtain the second inequality, it is enough to use Theorem 3.2 instead of Theorem 3.1 and note that

$$\operatorname{tr}\left(\sum_{j=1}^{n}\mathbb{E}\mathcal{H}(Y_j)^2\right)=\operatorname{tr}\left(\sum_{j=1}^{n}\mathbb{E}Y_jY_j^*\right)+\operatorname{tr}\left(\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j\right)=2\operatorname{tr}\left(\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j\right)$$

since for any $1\le j\le n$, $\operatorname{tr}(\mathbb{E}Y_jY_j^*)=\mathbb{E}\operatorname{tr}(Y_jY_j^*)=\mathbb{E}\operatorname{tr}(Y_j^*Y_j)$. □

In a particular case when $Y\in\mathbb{R}^d$ is a random vector such that $\mathbb{E}YY^T=\Sigma$ and $Y_1,\dots,Y_n$ are its i.i.d. copies, $\max(\|\sum_{j=1}^{n}\mathbb{E}Y_jY_j^*\|,\|\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j\|)=n\operatorname{tr}\Sigma$

and $\operatorname{tr}(\sum_{j=1}^{n}\mathbb{E}Y_j^*Y_j)=n\operatorname{tr}\Sigma$, hence $\bar{d}=2$ and the estimator $\hat{T}_{12}$ admits the following bound: if we replace t by $\sqrt{s}\sqrt{\operatorname{tr}\Sigma}$ and set $\theta=\sqrt{\frac{s}{n}}\frac{1}{\sqrt{\operatorname{tr}\Sigma}}$ in the second bound of Corollary 3.1, then

$$\Pr\left(\left\|\frac{\hat{T}_{12}}{n}-\mathbb{E}Y\right\|_2\geq\sqrt{\operatorname{tr}\Sigma}\sqrt{\frac{s}{n}}\right)\leq 4(1+1/s)e^{-s/2}.$$

3.4. *Bounds under weaker moment assumptions.* In this section, we discuss the mean estimation problem under weaker moment conditions. Namely, assume that $Y_1,\dots,Y_n$ are independent self-adjoint random matrices such that $\|\mathbb{E}|Y_j|^\alpha\|<\infty$ for some $\alpha\in(1,2]$ and all $1\leq j\leq n$. Let ψ_α satisfy

$$-\log\bigl(1-x+c_\alpha|x|^\alpha\bigr)\leq\psi_\alpha(x)\leq\log\bigl(1+x+c_\alpha|x|^\alpha\bigr)$$

for all $x\in\mathbb{R}$, where $c_\alpha=\frac{\alpha-1}{\alpha}\vee\sqrt{\frac{2-\alpha}{\alpha}}$. The fact that such ψ_α exists follows from Lemma A.2 in the Appendix. For example, one can take $\psi_\alpha(x)=\log(1+x+c_\alpha|x|^\alpha)$. The following result holds.

THEOREM 3.3. *Assume that $v_n^\alpha\geq\|\sum_{j=1}^{n}\mathbb{E}|Y_j|^\alpha\|$. Then for any positive t and θ,*

$$\Pr\left(\left\|\sum_{j=1}^{n}\left(\frac{1}{\theta}\psi_\alpha(\theta Y_j)-\mathbb{E}Y_j\right)\right\|\geq t\right)\leq 2d\exp\bigl(-\theta t+c_\alpha\theta^\alpha v_n^\alpha\bigr).$$

PROOF. The argument repeats the steps of Lemma 3.1 and Theorem 3.1, the only difference being that application of Fact 2.4 is replaced by Lemma A.2. □

REMARK 5. In the special case when $Y_1,\dots,Y_n$ are i.i.d. copies of Y with $v=\|\mathbb{E}|Y|^\alpha\|^{1/\alpha}$, setting $t=vn^{1/\alpha}s^{\frac{\alpha-1}{\alpha}}$ and $\theta=(\frac{1}{\alpha c_\alpha})^{1/(\alpha-1)}(\frac{s}{n})^{1/\alpha}\frac{1}{v}$ gives the inequality

$$\Pr\left(\left\|\frac{1}{n\theta}\sum_{j=1}^{n}\psi_\alpha(\theta Y_j)-\mathbb{E}Y\right\|\geq v\left(\frac{s}{n}\right)^{\frac{\alpha-1}{\alpha}}\right)$$
$$\leq 2d\exp\left(-\frac{\alpha-1}{\alpha}\left(\frac{1}{\alpha c_\alpha}\right)^{1/(\alpha-1)}s\right).$$

Note that for $\alpha=2$, we recover (3.11).

Before we proceed with discussion or further improvements and adaptation issues, let us demonstrate applications of developed techniques to popular problems in statistics and highlight the advantages over existing results.

4. Examples.

We present two examples which highlight the potential improvements obtained via our technique in popular scenarios: estimation of the covariance matrix in Frobenius and operator norms, and low-rank matrix completion problem.

4.1. *Estimation of the covariance matrix in operator norm.*

Let $Z \in \mathbb{R}^d$ be a random vector with $\mathbb{E}Z = \mu$, $\mathbb{E}\|Z - \mu\|_2^4 < \infty$, $\Sigma = \mathbb{E}[(Z-\mu)(Z-\mu)^T]$, and let $Z_1, \dots, Z_{2n}$ be i.i.d. copies of Z. Let us first assume that $\mu = 0$, and define

$$\widetilde{\Sigma}_{2n}(\theta) = \frac{1}{2n\theta} \sum_{j=1}^{2n} \psi(\theta Z_j Z_j^T),$$

where $\psi(\cdot)$ satisfies (3.1). Let $\sigma^2 \geq \|\mathbb{E}\|Z\|_2^2 ZZ^T\|$ and $\tilde{\theta} = \sqrt{\frac{t}{n}}\frac{1}{\sigma}$. It is straightforward to deduce from Theorem 3.1 that with probability $\geq 1 - 2de^{-t}$,

$$\|\widetilde{\Sigma}_{2n}(\tilde{\theta}) - \Sigma\| \leq \sigma\sqrt{\frac{t}{n}}.$$

REMARK 6. 1. Note that for any matrix $X = \lambda UU^T$ of rank 1 (where $\|U\|_2 = 1$),

$$\psi(X) = \psi(\lambda)UU^T \qquad [\text{since } \psi(0) = 0],$$

hence $\tilde{\Sigma}_{2n}(\tilde{\theta}) = \frac{1}{2n\tilde{\theta}} \sum_{j=1}^{2n} \psi(\tilde{\theta}\|Z_j\|_2^2)\frac{Z_j Z_j^T}{\|Z_j\|_2^2}$. In particular, this expression is easy to evaluate numerically; in general, computation of the estimator (3.5) requires n singular value decompositions.

2. Parameter σ is closely related to the *effective rank* defined as $\mathrm{r}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$ [44]; clearly, it always true that $\mathrm{r}(\Sigma) \leq d$. The quantity $\sqrt{\mathrm{r}(\Sigma)}\|\Sigma\|$ has been shown to control the expected error of the sample covariance estimator in the Gaussian setting [26]. Under the additional assumption that the kurtosis of the linear forms $\langle Z, v\rangle$, $v \neq 0$, is uniformly bounded by K, it is possible to show that (see Lemma 2.3 in [37]) that $\sigma^2 \leq K\mathrm{r}(\Sigma)\|\Sigma\|^2$. On the other hand, fluctuations of the error around its expected value in the Gaussian case [26] are controlled by the "weak variance" $\sup_{v \in \mathbb{R}^d : \|v\|_2 = 1} \mathbb{E}^{1/2}\langle Z, v\rangle^4 \leq \sqrt{K}\|\Sigma\|$, while in our bounds fluctuations are controlled by the "strong variance" σ^2; this fact leaves room for improvement in our construction and proof techniques.

Of course, the initial assumption that μ is known is often unrealistic, hence we modify the estimator as follows. Given $\theta > 0$, set

$$Y_j = \frac{1}{2}(Z_{2j-1} - Z_{2j})(Z_{2j-1} - Z_{2j})^T,$$

$$\widehat{\Sigma}_{2n}(\theta) = \frac{1}{n\theta} \sum_{j=1}^{n} \psi(\theta Y_j).$$

Let $\hat{\sigma}^2 \geq \frac{1}{2}\|\mathbb{E}((Z-\mu)(Z-\mu)^T)^2 + \operatorname{tr}(\Sigma)\Sigma + 2\Sigma^2\|$, and $\hat{\theta} = \sqrt{\frac{t}{n}}\frac{1}{\hat{\sigma}}$. Our covariance estimator is then defined as $\widehat{\Sigma}_{2n} := \widehat{\Sigma}_{2n}(\hat{\theta})$. The following result can be deduced from Theorem 3.1.

COROLLARY 4.1. *With probability* $\geq 1 - 2de^{-t}$,

$$\|\widehat{\Sigma}_{2n} - \Sigma\| \leq \sqrt{2}\hat{\sigma}\sqrt{\frac{t}{n}}.$$

Before presenting the proof, let us make several additional remarks.

REMARK 7. 1. It is not hard to show that (see Corollary 7.1 of the Supplementary Material [36]) that $\|\mathbb{E}((Z-\mu)(Z-\mu)^T)^2\| \geq \operatorname{tr}(\Sigma)\|\Sigma\|$; hence it is enough to choose $\hat{\sigma}^2 \geq \|\Sigma\|^2 + \sigma_0^2 = \|\Sigma\|^2 + \|\mathbb{E}((Z-\mu)(Z-\mu)^T)^2\|$. In view of Remark 6, this expression can be further simplified under the bounded kurtosis assumption, and one can choose $\hat{\sigma}^2 \geq \|\Sigma\|^2(1 + K\mathrm{r}(\Sigma))$, where K is the uniform bound on the kurtosis of the coordinates of Z, and $\mathrm{r}(\Sigma)$ is the effective rank.

2. Construction of $\widehat{\Sigma}_{2n}(\theta)$ essentially halves the effective sample size. While the loss of a constant factor can be deemed insignificant in non-asymptotic theoretical bounds, it is undesirable in applications. A more natural version of the estimator based on a sample of size $2n$ is the U-statistic

$$\bar{\Sigma}_{2n}(\theta) = \frac{1}{\binom{2n}{2}} \sum_{1 \leq i < j \leq 2n} \frac{1}{\theta}\psi\left(\frac{\theta}{2}(Z_i - Z_j)(Z_i - Z_j)^T\right).$$

Another possibility to avoid “halving” the sample size is to center the data using a robust estimator of location, such as the spatial median or the median-of-means estimator [23, 33, 35]. Analysis of the estimators of these types is not covered in the present paper, and requires a slightly different set of technical tools to deal with dependent summands; see [37] for results in this direction.

PROOF OF COROLLARY 4.1. Note that for all $j = 1, \ldots, n$, $\mathbb{E}Y_j = \Sigma$. Since $Y_1, \ldots, Y_n$ are i.i.d. random matrices, Theorem 3.1 applies (see Remark 2), giving that

$$\Pr\left(\|\hat{\Sigma}(\hat{\theta}) - \Sigma\| \geq \hat{\sigma}\sqrt{\frac{2t}{n}}\right) \leq 2de^{-t},$$

where $\hat{\sigma}^2 \geq \|\mathbb{E}Y_1^2\|$. It is easy to check that

$$\|\mathbb{E}Y_1^2\| = \frac{1}{2}\|\mathbb{E}((Z-\mu)(Z-\mu)^T)^2 + \operatorname{tr}(\Sigma)\Sigma + 2\Sigma^2\|,$$

and the result follows. □

4.2. *Estimation of the covariance matrix in Frobenius norm.* Next, we present an estimator which achieves strong deviation guarantees in the Frobenius norm. Estimation of the covariance matrix with respect to this norm has been previously investigated in the literature, for instance, see [28], [8] and references therein; Frobenius norm is a natural choice when one wants to understand the effect of the rank of an unknown covariance matrix on the estimation error [32]. Let $\hat{S}_{2n}$ be the sample covariance estimator based on $Z_1, \dots, Z_{2n}$:

$$\hat{S}_{2n} = \frac{1}{\binom{2n}{2}} \sum_{1 \le i < j \le 2n} \frac{(Z_i - Z_j)(Z_i - Z_j)^T}{2}.$$

The following "soft thresholding" estimator has been studied in [32]; here, $\tau > 0$ is a fixed threshold parameter:

$$\hat{S}^{\tau}_{2n} = \operatorname*{argmin}_{A \in \mathbb{R}^{d \times d}} \bigl[\|A - \hat{S}_{2n}\|_{\mathrm{F}}^2 + \tau \|A\|_1\bigr]. \tag{4.1}$$

We propose to replace the sample covariance $\hat{S}_{2n}$ by $\widehat{\Sigma}_{2n}$, and consider

$$\widehat{\Sigma}^{\tau}_{2n} = \operatorname*{argmin}_{A \in \mathbb{R}^{d \times d}} \bigl[\|A - \widehat{\Sigma}_{2n}\|_{\mathrm{F}}^2 + \tau \|A\|_1\bigr]. \tag{4.2}$$

It is not hard to see (e.g., see the proof of Theorem 1 in [32]) that $\widehat{\Sigma}^{\tau}_{2n}$ can be written explicitly as

$$\widehat{\Sigma}^{\tau}_{2n} = \sum_{j=1}^{d} \max\bigl(\lambda_j(\widehat{\Sigma}_{2n}) - \tau/2, 0\bigr) v_j(\widehat{\Sigma}_{2n}) v_j(\widehat{\Sigma}_{2n})^T,$$

where $\lambda_j(\widehat{\Sigma}_{2n})$ and $v_j(\widehat{\Sigma}_{2n})$ are the eigenvalues and corresponding eigenvectors of $\widehat{\Sigma}_{2n}$. The following result holds.

THEOREM 4.1. *For any*

$$\begin{aligned} &\tau \ge 4\hat{\sigma}\sqrt{\frac{t + \log(2d)}{2n}}, \\ &\|\widehat{\Sigma}^{\tau}_{2n} - \Sigma\|_{\mathrm{F}}^2 \le \inf_{A \in \mathbb{R}^{d \times d}} \biggl[\|A - \Sigma\|_{\mathrm{F}}^2 + \frac{(1+\sqrt{2})^2}{8} \tau^2 \operatorname{rank}(A)\biggr] \end{aligned} \tag{4.3}$$

with probability $\ge 1 - e^{-t}$.

The result stated above mimics the (almost) optimal rates obtained in [32] (in the situation when no data is missing) under significantly weaker assumptions on the underlying distribution.

PROOF OF THEOREM 4.1. The proof is based on the following lemma.

LEMMA 4.1. *Inequality* (4.4) *holds on the event* $\mathcal{E} = \{\tau \ge 2\|\widehat{\Sigma}_{2n} - \Sigma\|\}$.

To verify this statement, it is enough to repeat the steps of the proof of Theorem 1 in [32], replacing each occurrence of the sample covariance $\hat{S}_{2n}$ by its robust counterpart $\widehat{\Sigma}^{\tau}_{2n}$.

The result then follows from Corollary 4.1 that $\Pr(\mathcal{E}) \ge 1 - e^{-t}$ whenever $\tau \ge 4\hat{\sigma}\sqrt{\frac{t+\log(2d)}{2n}}$. □

4.3. *Matrix completion.* Let $A_0 \in \mathbb{R}^{d_1 \times d_2}$ be an unknown matrix, and assume that we observe a random subset of its entries contaminated by noise. The goal is to estimate A_0 from a small number of such noisy measurements under an additional assumption that A_0 is likely to be of low rank (or can be well approximated by a low rank matrix). More specifically, let

$$\mathcal{X} = \{e_j(d_1)e_k^T(d_2), 1 \le j \le d_1, 1 \le k \le d_2\},$$

where $e_j(d_1)$ and $e_k(d_2)$ are the elements of the canonical bases of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Let X have uniform distribution $\Pi := \mathrm{Unif}(\mathcal{X})$ on $\mathcal{X}$, and assume that the noisy linear measurement Y has the form

$$Y = \mathrm{tr}(X^T A_0) + \xi,$$

where $\mathbb{E}(\xi|X) = 0$. Finally, assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. copies of (X, Y).

It is easy to check that $\mathbb{E}(YX) = \frac{1}{d_1 d_2} A_0$, hence the natural unbiased estimator of A_0 is

$$\widehat{A} = \frac{d_1 d_2}{n} \sum_{j=1}^{n} Y_j X_j.$$

To incorporate the structural (low-rank) assumption on A_0, the following estimator has been considered in the literature: let $\tau > 0$, and define

$$\begin{aligned}\widehat{A}^{\tau} &= \operatorname*{argmin}_{A \in \mathbb{R}^{d_1 \times d_2}} \left[\frac{1}{d_1 d_2}\|A - \widehat{A}\|_{\mathrm{F}}^2 + \tau\|A\|_1\right] \\ &= \operatorname*{argmin}_{A \in \mathbb{R}^{d_1 \times d_2}} \left[\frac{1}{d_1 d_2}\|A\|_{\mathrm{F}}^2 - \left\langle \frac{2}{n}\sum_{j=1}^{n} Y_j X_j, A\right\rangle + \tau\|A\|_1\right].\end{aligned}$$

Note that one can use the symmetric version $\widehat{A}_s \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$ of $\widehat{A}$ instead, defined as

$$\widehat{A}_s = \frac{d_1 d_2}{n} \sum_{j=1}^{n} Y_j \mathcal{H}(X_j),$$

so that $\mathbb{E}\widehat{A}_s = \mathcal{H}(A_0)$, and consider the equivalent convex minimization problem

$$\begin{aligned}\widehat{A}^\tau &= \underset{A\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}}\left[\frac{1}{d_1d_2}\|\mathcal{H}(A)-\mathcal{H}(\widehat{A}_s)\|_{\mathrm{F}}^2 + 2\tau\|A\|_1\right]\\ &= \underset{A\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}}\left[\frac{1}{d_1d_2}\|\mathcal{H}(A)\|_{\mathrm{F}}^2 - \left\langle \frac{2}{n}\sum_{j=1}^n Y_j\mathcal{H}(X_j), \mathcal{H}(A)\right\rangle + 2\tau\|A\|_1\right].\end{aligned}$$

However, strong theoretical guarantees for this estimator exist only when the "noise term" ξ_j is either bounded with probability 1, or has sub-exponential tails. We propose to replace $\widehat{A}_s$ with a robust estimator

$$\widehat{R} = \frac{d_1d_2}{n\theta}\sum_{j=1}^n \psi\big(\theta Y_j\mathcal{H}(X_j)\big),$$

where $\psi(\cdot)$ satisfies (3.1) and

$$\theta := \theta(t,n,A_0) = \frac{1}{\|A_0\|_{\max}\vee\sqrt{\operatorname{Var}(\xi)}}\sqrt{\frac{(t+\log(2(d_1+d_2)))(d_1\wedge d_2)}{n}}.$$

The reasoning behind this choice of θ is explained below. Consider

$$\widehat{R}^\tau = \underset{A\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}}\left[\frac{1}{d_1d_2}\|\mathcal{H}(A)\|_{\mathrm{F}}^2 - \left\langle\frac{2}{d_1d_2}\widehat{R}, \mathcal{H}(A)\right\rangle + 2\tau\|A\|_1\right].$$

Finally, set

$$M = \widehat{R} - \mathbb{E}\big(Y\mathcal{H}(X)\big).$$

The following result holds.

THEOREM 4.2. *Assume that ξ_j is independent of X_j, $j=1,\dots,n$, and that* $\operatorname{Var}(\xi)<\infty$. *For any*

$$\tau \ge 4\big(\|A_0\|_{\max}\vee\sqrt{\operatorname{Var}(\xi)}\big)\sqrt{\frac{t+\log(2(d_1+d_2))}{n(d_1\wedge d_2)}},$$

$$\frac{1}{d_1d_2}\|\widehat{R}^\tau - A_0\|_F^2 \le \inf_{A\in\mathbb{R}^{d_1\times d_2}}\left[\frac{1}{d_1d_2}\|A-A_0\|_{\mathrm{F}}^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 d_1d_2\tau^2\operatorname{rank}(A)\right]$$

with probability $\ge 1-e^{-t}$.

Note that we only assume that $\operatorname{Var}(\xi)<\infty$, while in [16], a similar result is obtained under a slightly stronger assumption requiring that $\mathbb{E}|\xi|^{2+\varepsilon}<\infty$ for some $\varepsilon>0$.

PROOF. Define $\mathbb{A}\subseteq\mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$ to be the image of $\mathbb{R}^{d_1\times d_2}$ under $\mathcal{H}(\cdot)$:

$$\mathbb{A} = \big\{B\in\mathbb{R}^{(d_1+d_2)\times(d_1+d_2)} : B=\mathcal{H}(A) \text{ for some } A\in\mathbb{R}^{d_1\times d_2}\big\}.$$

We begin with the following inequality.

LEMMA 4.2. *Assume that $\tau \ge 2\|M\|$. Then*

$$\frac{1}{d_1 d_2}\|\mathcal{H}(\widehat{R}^\tau) - \mathcal{H}(A_0)\|_{\mathrm{F}}^2 \le \inf_{B\in\mathbb{A}}\left[\frac{1}{d_1 d_2}\|B - \mathcal{H}(A_0)\|_{\mathrm{F}}^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 d_1 d_2 \tau^2 \operatorname{rank}(B)\right].$$

PROOF. By the definition of $\widehat{R}^\tau$, we see that

$$\mathcal{H}(\widehat{R}^\tau) = \operatorname*{argmin}_{B\in\mathbb{A}}\left[\frac{1}{d_1 d_2}\|B\|_{\mathrm{F}}^2 - \left\langle \frac{2}{d_1 d_2}\widehat{R}, B\right\rangle + \tau\|B\|_1\right].$$

If we replace $\frac{1}{d_1 d_2}\widehat{R}$ by $\frac{1}{d_1 d_2}\widehat{A}_s = \frac{1}{n}\sum_{j=1}^n Y_j \mathcal{H}(X_j)$, the result follows from Theorem 1 in [27] immediately. To obtain the current statement, it is enough to repeat the argument of Theorem 1 in [27], replacing each occurrence of the matrix $\frac{1}{d_1 d_2}\widehat{A}_s$ by $\frac{1}{d_1 d_2}\widehat{R}$. □

To complete the proof, we will estimate each side of the inequality of Lemma 4.2. First, it is obvious from the definition of the Frobenius norm that

$$\frac{1}{d_1 d_2}\|\mathcal{H}(\widehat{R}^\tau) - \mathcal{H}(A_0)\|_{\mathrm{F}}^2 = \frac{2}{d_1 d_2}\|\widehat{R}^\tau - A_0\|_{\mathrm{F}}^2. \tag{4.4}$$

Next, since $\operatorname{rank}(\mathcal{H}(A)) = 2\operatorname{rank}(A)$,

$$\begin{aligned}
&\inf_{B\in\mathbb{A}}\left[\frac{1}{d_1 d_2}\|B - \mathcal{H}(A_0)\|_{\mathrm{F}}^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 d_1 d_2 \tau^2 \operatorname{rank}(B)\right] \\
&\quad = 2\inf_{A\in\mathbb{R}^{d_1\times d_2}}\left[\frac{1}{d_1 d_2}\|A - A_0\|_{\mathrm{F}}^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 d_1 d_2 \tau^2 \operatorname{rank}(A)\right].
\end{aligned} \tag{4.5}$$

It remains to estimate the probability of the event $\mathcal{E} = \{\tau \ge 2\|M\|\}$. Let

$$\sigma^2 := \max(\|\mathbb{E}[Y^2 X X^T]\|, \|\mathbb{E}[Y^2 X^T X]\|).$$

LEMMA 4.3. *Assume that ξ_j is independent of X_j, $j = 1, \dots, n$. Then*

$$\sigma^2 \le (\operatorname{Var}(\xi) \vee \|A_0\|_{\max}^2)\frac{2}{d_1 \wedge d_2}.$$

PROOF. Note that $\mathbb{E}[Y^2 X X^T] = \mathbb{E}[\xi^2 X X^T] + \mathbb{E}[(\operatorname{tr}(X^T A_0))^2 X X^T]$. Moreover, $|\operatorname{tr}(X^T A_0)| \le \max_{i,j}|(A_0)_{i,j}| = \|A_0\|_{\max}$, and $\|\mathbb{E} X X^T = \frac{1}{d_1}\|$, hence

$$\|\mathbb{E}[Y^2 X X^T]\| \le \operatorname{Var}(\xi)\frac{1}{d_1} + \|A_0\|_{\max}^2 \frac{1}{d_1}.$$

Similarly,

$$\|\mathbb{E}[Y^2 X^T X]\| \le \operatorname{Var}(\xi)\frac{1}{d_2} + \|A_0\|_{\max}^2 \frac{1}{d_2}. \qquad \square$$

Applying Theorem 3.1 (see Remark 2) with

$$\begin{aligned}\theta &= \sqrt{\frac{2(t+\log(2(d_1+d_2)))}{n}}\frac{1}{((\operatorname{Var}(\xi)\vee\|A_0\|_{\max}^2)\frac{2}{d_1\wedge d_2})^{1/2}}\\ &= \frac{1}{\|A_0\|_{\max}\vee\sqrt{\operatorname{Var}(\xi)}}\sqrt{\frac{(t+\log(2(d_1+d_2)))(d_1\wedge d_2)}{n}},\end{aligned}$$

we see that

$$\|M\| \le 2\big(\|A_0\|_{\max}\vee\sqrt{\operatorname{Var}(\xi)}\big)\sqrt{\frac{t+\log(2(d_1+d_2))}{n(d_1\wedge d_2)}}$$

with probability $\ge 1-e^{-t}$. The final result now follows from the combination of this inequality with (4.4), (4.5) and Lemma 4.2. □

5. Optimal choice of θ and adaptation to the unknown second moment.

To make results of Theorem 3.1 useful, one has to set the value for the parameter θ which in turn depends on the (usually unknown) norm $\sigma_n^2 = \|\sum_{j=1}^n \mathbb{E}Y_j^2\|$. To address this problem, we develop a simple adaptive solution based on Lepski's method.

Lepski's method [29] is a powerful general technique that allows to adapt to the unknown structure of the problem, for example, bandwidth selection in nonparametric estimation, or an unknown second moment in our case. Let $Y_1,\dots,Y_n \in \mathbb{C}^{d\times d}$ be independent self-adjoint random matrices with $\sigma_n^2 = \|\sum_{j=1}^n \mathbb{E}Y_j^2\|$, and assume that $\sigma_{\min},\sigma_{\max}$ are such that

$$\sigma_{\min} \le \frac{\sigma_n}{\sqrt{n}} \le \sigma_{\max}.$$

Parameters $\sigma_{\min}$ and $\sigma_{\max}$ are "crude" preliminary bounds that can differ from $\sigma_n/\sqrt{n}$ by several orders of magnitude. Let $\sigma_j = \sigma_{\min}2^j$ and

$$\mathcal{J} = \{j \in \mathbb{Z} : \sigma_{\min} \le \sigma_j < 2\sigma_{\max}\}$$

be a set of cardinality $|\mathcal{J}| \le 1 + \log_2(\sigma_{\max}/\sigma_{\min})$, and for each $j \in \mathcal{J}$ set $\theta_j = \theta(j,t) = \sqrt{\frac{2t}{n}}\frac{1}{\sigma_j}$. Define

$$T_{n,j} = \frac{1}{n\theta_j}\sum_{i=1}^n \psi(\theta_j Y_i),$$

where $\psi(\cdot)$ satisfies (3.1). Finally, set

$$j_* := \min\left\{ j \in \mathcal{J} : \forall k > j \text{ s.t. } k \in \mathcal{J}, \|T_{n,k} - T_{n,j}\| \le 2\sigma_k \sqrt{\frac{2t}{n}} \right\} \tag{5.1}$$

and $T_n^* := T_{n,j_*}$.

The next result shows that adaptation is possible at the cost of an additional multiplicative constant factor 6 in the deviation bound.

THEOREM 5.1. *The following inequality holds for any* $t > 0$:

$$\Pr\left(\|T_n^* - \mathbb{E}Y\| \ge 6(\sigma_n/\sqrt{n})\sqrt{\frac{2t}{n}} \right) \le 2d \log_2\left(\frac{2\sigma_{\max}}{\sigma_{\min}} \right) e^{-t}.$$

PROOF. Let $\bar{j} = \min\{ j \in \mathcal{J} : \sigma_j \ge \frac{\sigma_n}{\sqrt{n}}\}$ (hence $\sigma_{\bar{j}} \le 2\frac{\sigma_n}{\sqrt{n}}$). First, we will show that $j_* \le \bar{j}$ with high probability. Indeed,

$$\begin{aligned}
\Pr(j_* > \bar{j}) &\le \Pr\left(\bigcup_{k \in \mathcal{J} : k > \bar{j}} \left\{ \|T_{n,k} - T_{n,\bar{j}}\| > 2\sigma_k \sqrt{\frac{2t}{n}} \right\} \right) \\
&\le \Pr\left(\|T_{n,\bar{j}} - \mathbb{E}Y\| > \sigma_{\bar{j}} \sqrt{\frac{2t}{n}} \right) + \sum_{k \in \mathcal{J} : k > \bar{j}} \Pr\left(\|T_{n,k} - \mathbb{E}Y\| > \sigma_k \sqrt{\frac{2t}{n}} \right) \\
&\le 2de^{-t} + 2d \log_2\left(\frac{\sigma_{\max}}{\sigma_{\min}} \right) e^{-t},
\end{aligned}$$

where we used Theorem 3.1 to bound each of the probabilities in the sum. The display above implies that the event

$$\mathcal{B} = \bigcap_{k \in \mathcal{J} : k \ge \bar{j}} \left\{ \|T_{n,k} - \mathbb{E}Y\| \le \sigma_k \sqrt{\frac{2t}{n}} \right\}$$

of probability $\ge 1 - 2d \log_2(\frac{2\sigma_{\max}}{\sigma_{\min}})e^{-t}$ is contained in $\mathcal{E} = \{j_* \le \bar{j}\}$. Hence, on $\mathcal{B}$ we have that

$$\begin{aligned}
\|T_n^* - \mathbb{E}Y\| &\le \|T_n^* - T_{n,\bar{j}}\| + \|T_{n,\bar{j}} - \mathbb{E}Y\| \le 2\sigma_{\bar{j}} \sqrt{\frac{2t}{n}} + \sigma_{\bar{j}} \sqrt{\frac{2t}{n}} \\
&\le 4\frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}} + 2\frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}} = 6\frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}},
\end{aligned}$$

and result follows. □

REMARK 8. It follows from the proof that constant factor 6 in Theorem 5.1 can be reduced to $3 + \varepsilon$ for any $\varepsilon > 0$ by considering the “finer grid”, that is, replacing $\mathcal{J}$ by $\{j \in \mathbb{Z} : \sigma_{\min} \le \kappa^j \sigma_{\min} < \kappa\sigma_{\max}\}$ for some $1 < \kappa < 2$, at the cost of replacing $\log_2(\frac{2\sigma_{\max}}{\sigma_{\min}})$ by $\log_2(\frac{\kappa\sigma_{\max}}{\sigma_{\min}})/\log_2 \kappa$.

6. From bounds depending on $\|\mathbb{E}Y^2\|$ to bounds depending on $\|\mathbb{E}(Y - \mathbb{E}Y)^2\|$.

Assume that $Y_1, \dots, Y_n$ are i.i.d. copies of $Y \in \mathbb{C}^{d\times d}$. In this section, we build upon previously established bounds to provide performance guarantees for the estimator defined via (3.3), (3.4). To this end, we study a version of the steepest descent scheme for the problem (3.3) initialized at the point $\widehat{T}_\theta^{(0)}$, namely, $\hat{T}_0 := \widehat{T}_{\theta_0}^{(0)}$ and

$$\hat{T}_k = \hat{T}_{k-1} + \frac{1}{n\theta_k}\sum_{j=1}^n \psi\big(\theta_k(Y_j - \hat{T}_{k-1})\big), \qquad k \ge 1$$

for an appropriate choice of $\theta_k, k \ge 0$. Note that for any nonrandom self-adjoint matrix S and $\theta_S = \sqrt{\frac{s}{n}}\frac{1}{\|\mathbb{E}(Y-S)^2\|^{1/2}}$, Theorem 3.1 implies that

$$\Pr\left(\|T_n(S) - \mathbb{E}Y\| \ge \|\mathbb{E}(Y - S)^2\|^{1/2}\sqrt{\frac{s}{n}}\right) \le 2d\exp(-s/2),$$

where $T_n(S) = S + \frac{1}{n\theta_S}\sum_{j=1}^n \psi(\theta_S(Y_j - S))$. Hence, if we use random S which is "not too far" from $\mathbb{E}Y$ with high probability, we expect that the deviation guarantees will still hold with the "variance parameter" close to $\|\mathbb{E}(Y - \mathbb{E}Y)^2\|$.

Everywhere in this section, we will assume that one has access to some known (possibly very crude) bounds for $\sigma^2 = \|\mathbb{E}Y^2\|$ and $\sigma_0^2 = \|\mathbb{E}(Y - \mathbb{E}Y)^2\|$.

ASSUMPTION 1. Let $\sigma_{\min}, \sigma_{0,\min}$ and $\sigma_{\max}, \sigma_{0,\max}$ be known constants such that

$$\sigma_{\min} \le \sigma \le \sigma_{\max} \quad \text{and} \quad \sigma_{0,\min} \le \sigma_0 \le \sigma_{0,\max}.$$

6.1. *Two-step estimation based on sample splitting.*

We will first discuss the simplest (but not the most efficient) approach based on splitting the sample $Y_1, \dots, Y_n$ into two disjoint subsets G_1 and G_2 of cardinality $\ge \lfloor n/2 \rfloor$ each, and performing one step of the steepest descent. The main advantage of this approach is the fact that it requires very mild assumptions. The idea is to apply Lepski's method (as discussed in Section 5) twice: on the first step, we obtain an estimator $\hat{T}_0$ based on subsample G_1, and on the second step we apply Lepski's method again to the subsample $\{Y_j - \hat{T}_0 : 1 \le j \le n, Y_j \in G_2\}$.

Here is the more detailed description: set $\sigma_j = 2^j\sigma_{\min}$,

$$\mathcal{J}_1 = \{j \in \mathbb{Z} : \sigma_{\min} \le \sigma_j < 2\sigma_{\max}\}$$

and $\sigma_{0,j} = 2^j\sigma_{0,\min}$,

$$\mathcal{J}_2 = \left\{j \in \mathbb{Z} : \sigma_{0,\min} \le \sigma_{0,j} < 2\left(\sigma_{0,\max} + 12\sigma_{\max}\sqrt{\frac{t}{n}}\right)\right\},$$

and let $\hat{T}_0$ be the "Lepski-type" adaptive estimator based on the subsample G_1 defined as

$$\hat{T}_0 = T_{|G_1|, j_1^*}(0; G_1),$$

where

$$T_{|G_1|,j}(S; G_1) = \frac{1}{|G_1|\theta_j} \sum_{i=1}^{|G_1|} \psi\big(\theta_j(Y_i - S)\big),$$

$\theta_j = \sqrt{\frac{2t}{n/2}}\frac{1}{\sigma_j}$, $\psi(\cdot)$ satisfies (3.1) and

$$j_1^* := \min\Bigg\{ j \in \mathcal{J}_1 : \forall k \in \mathcal{J}_1 \text{ s.t. } k > j, \\ \big\| T_{|G_1|,k}(0; G_1) - T_{|G_1|,j}(0; G_1) \big\| \le 2\sigma_k \sqrt{\frac{2t}{|G_1|}} \Bigg\}$$

$\hat{T}_1$ is then defined as follows:

$$\hat{T}_1 = \hat{T}_0 + T_{|G_2|, j_2^*}(\hat{T}_0; G_2),$$

where

$$T_{|G_2|,j}(S; G_2) = \frac{1}{|G_2|\theta_{0,j}} \sum_{i=|G_1|+1}^{n} \psi\big(\theta_{0,j}(Y_i - S)\big), \qquad \theta_{0,j} = \sqrt{\frac{2t}{n/2}}\frac{1}{\sigma_{0,j}}$$

and

$$j_2^* := \min\Bigg\{ j \in \mathcal{J}_2 : \forall k \in \mathcal{J}_2 \text{ s.t. } k > j, \\ \big\| T_{|G_2|,k}(\hat{T}_0; G_2) - T_{|G_2|,j}(\hat{T}_0; G_2) \big\| \le 2\sigma_{0,k} \sqrt{\frac{2t}{|G_2|}} \Bigg\}.$$

THEOREM 6.1. *With probability at least*

$$1 - 2d\left(2 + \log_2\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right) + \log_2\left(\frac{\sigma_{0,\max} + 12\sigma_{\max}\sqrt{t/n}}{\sigma_{0,\min}}\right)\right)e^{-t},$$

the following inequality holds:

$$\|\hat{T}_1 - \mathbb{E}Y\| \le 12\left(\sigma_0 + 12\sigma\sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}.$$

PROOF. See Section 4 in the Supplementary Material [36]. □

The main feature of this result is the variance term $\sigma_0 + 12\sigma\sqrt{\frac{t}{n}}$ that can be much smaller compared to σ as long as $t \ll n$.

6.2. *Results for the estimator $\widehat{T}_\theta^*$ defined via equation* (3.4). We will next show how to design an estimator with deviations controlled by the "correct" variance term without sample splitting (however, subject to the condition that the sample size is sufficiently large). In what follows, we will make an additional assumption about the function ψ.

ASSUMPTION 2. Function $\psi(\cdot)$ satisfies (3.1) and is operator Lipschitz, meaning that $\|\psi(A)-\psi(B)\| \le L\|A-B\|$ for all self-adjoint $A, B \in \mathbb{C}^{d\times d}$, with Lipschitz constant L *independent* of the dimension d.

For example, we may take $\psi=\psi_1$ or $\psi=\psi_2$ (see Lemma A.3 for details). As before, let $t>0$ be fixed, set $\sigma_{0,j}=2^j\sigma_{0,\min}$,

$$\mathcal{J}=\{j\in\mathbb{Z}:\sigma_{0,\min}\le\sigma_{0,j}<2\sigma_{0,\max}\},$$

$$\theta=\sqrt{\frac{2t}{n}}\frac{1}{\sigma_{\max}} \quad \text{and} \quad \theta_j=\sqrt{\frac{2t}{n}}\frac{1}{\sigma_{0,j}} \qquad \text{for } j\in\mathcal{J}.$$

For all $j\in\mathcal{J}$, define $\delta_j^{(0)}=\sigma_{\max}\sqrt{\frac{2t}{n}}$ and

$$\delta_j^{(k)}=\frac{12}{5}\sigma_{0,j}\sqrt{\frac{2t}{n}}+6^{-k}\left(\sigma_{\max}\sqrt{\frac{2t}{n}}-\frac{12}{5}\sigma_{0,j}\sqrt{\frac{2t}{n}}\right) \tag{6.1}$$

for $k\ge1$. Next, for each $j\in\mathcal{J}$, we define

$$T_{n,j}^{(0)}:=T_n^{(0)}=\frac{1}{n\theta}\sum_{i=1}^n\psi(\theta Y_i), \tag{6.2}$$

(independent of j),[4] and

$$T_{n,j}^{(k)}:=T_{n,j}^{(k-1)}+\frac{1}{n\theta_j}\sum_{i=1}^n\psi\big(\theta_j\big(Y_i-T_{n,j}^{(k-1)}\big)\big)$$

for $k\ge1$. Finally, we apply Lepski's method to the collection of estimators $\{T_{n,j}^{(k)}: j\in\mathcal{J}\}$. To this end, define $\hat{T}_k:=T_{n,j_k^*}^{(k)}$, where

$$j_k^*=\min\big\{j\in\mathcal{J}:\forall l\in\mathcal{J}\text{ s.t. } l>j, \|T_{n,l}^{(k)}-T_{n,j}^{(k)}\|\le2\delta_l^{(k)}\big\}.$$

Note that the estimator $\hat{T}_k$ is completely data-dependent. We are ready to state the main result of this section.

[4] Particular choice of $T_n^{(0)}$ does not matter as long as $\|T_n^{(0)}-\mathbb{E}Y\|$ is small with high probability.

THEOREM 6.2. *Let*

$$\tau = 1.1K\sqrt{\frac{d^2+Lt}{n}} + \sqrt{\frac{2t}{n}}\frac{1}{2\sigma_0},$$

where $K > 0$ *is an absolute constant, and assume that* $\tau \le 1/6$. *Moreover, assume that*

$$\left(\frac{24}{5}\sigma_{0,\max} \vee \sigma_{\max}\right)\sqrt{\frac{2t}{n}} \le 1. \tag{6.3}$$

Then for all $k \ge 0$ *simultaneously,*

$$\|\hat{T}_k - \mathbb{E}Y\| \le 3\left[(1-6^{-k})\frac{24}{5}\sigma_0\sqrt{\frac{2t}{n}} + 6^{-k}\sigma_{\max}\sqrt{\frac{2t}{n}}\right]$$

with probability $\ge 1 - 8d(1+2\log_2(\frac{12\sigma_{\max}}{5\sigma_{0,\min}}))\log_2(\frac{2\sigma_{0,\max}}{\sigma_{0,\min}})e^{-t}$.

PROOF. See Section 5 in the Supplementary Material [36]. □

The next corollary easily follows from the preceding result. Let $\mathcal{A}$ be the event of probability

$$\Pr(\mathcal{A}) \ge 1 - 8d\left(1 + 2\log_2\left(\frac{12\sigma_{\max}}{5\sigma_{0,\min}}\right)\right)\log_2\left(\frac{2\sigma_{0,\max}}{\sigma_{0,\min}}\right)e^{-t}$$

defined in Theorem 6.2. Since by the properties of the steepest descent scheme $T_{n,j}^{(k)}$ converges to the solution (denoted $\widehat{T}_{\theta_j}^*$) of the problem (3.3), we can easily deduce the following inequality.

COROLLARY 6.1. *Let* $\{\widehat{T}_{\theta_j}^*\}_{j\in\mathcal{J}}$ *satisfy the equations*

$$\frac{1}{n\theta_j}\sum_{i=1}^n \psi(\theta_j(Y_i - \widehat{T}_{\theta_j}^*)) = 0_{d\times d}, \qquad j \in \mathcal{J}.$$

Then on event $\mathcal{A}$, $\|\widehat{T}_{\theta_j}^* - \mathbb{E}Y\| \le \lim_{k\to\infty}\delta_j^{(k)} = \frac{12}{5}\sigma_{0,j}\sqrt{\frac{2t}{n}}$.

One can further apply Lepski's method (see Section 5) to the collection $\{\widehat{T}_{\theta_j}^*\}_{j\in\mathcal{J}}$ to obtain a completely data-dependent estimator $\widehat{T}^*$ that satisfies

$$\|\widehat{T}^* - \mathbb{E}Y\| \le \frac{72}{5}\sigma_0\sqrt{\frac{2t}{n}}$$

with high probability (in particular, on event $\mathcal{A}$).

7. Numerical simulation results. Numerical simulation was performed for covariance estimation problem. Data was simulated as follows: let $U = (U^{(1)}, \dots, U^{(100)})^T \in \mathbb{R}^{100}$ be a vector with i.i.d. coordinates such that $U^{(j)} \overset{\mathrm{d}}{=} \frac{1}{\sqrt{2c(q)}}(\xi_{j,1} - \xi_{j,2})$, where $\xi_{j,1}$ and $\xi_{j,2}$, $j = 1, \dots, 100$, are independent random variables with probability density function

$$p_\xi(t; q) = \frac{q}{(1+t)^{1+q}} I\{t \ge 0\}$$

(which belongs to the Pareto family), $c(q) = \mathrm{Var}(\xi) = \frac{q}{(q-1)^2(q-2)}$ and $q = 4.01$; in particular, $\mathrm{Var}(U^{(j)}) = 1$. Finally, let $Z = \sqrt{\Sigma} U$, where Σ is a diagonal matrix with $\Sigma_{11} = 10$, $\Sigma_{22} = 5$, $\Sigma_{33} = 1$ and $\Sigma_{jj} = \frac{1}{97}$, $j \ge 4$, in particular, $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ^T = \Sigma$.

The goal of numerical experiment was to evaluate the quality of estimation of the covariance matrix Σ as well as its first eigenvector e_1 corresponding to $\lambda_1 = 10$. We tested two scenarios with sample sizes equal n to 100 and 1000. In both cases, we generated $Z_1, \dots, Z_n$, i.i.d. copies of Z and centered the data via the spatial (or geometric) median defined as

$$\widehat{M}_n = \underset{y \in \mathbb{R}^{100}}{\operatorname{argmin}} \sum_{j=1}^{100} \|y - Z_j\|_2.$$

We compared two estimators, $\widehat{S}_n$ and $\widehat{\Sigma}_n$ constructed as follows: set $Z_j^0 := Z_j - \widehat{M}_n$ for brevity, and

$$\widehat{S}_n = \frac{1}{n} \sum_{j=1}^{n} Z_j^0 {Z_j^0}^T,$$

which is the analogue of sample covariance with "robust centering".

Next, $\widehat{\Sigma}_n$ was constructed using a version of Lepski's method described in Section 5. We provide details for completeness: set

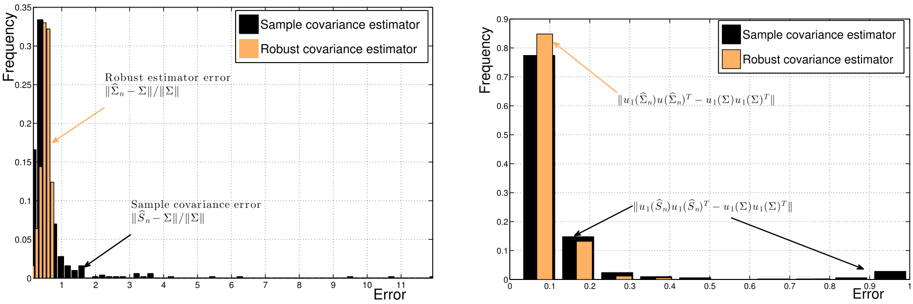
$$\sigma_{\max} := 2\sqrt{\left\| \frac{1}{n} \sum_{j=1}^{n} \|Z_j^0\|_2^2 Z_j^0 {Z_j^0}^T \right\|}, \qquad \sigma_{\min} = \frac{\sigma_{\max}}{100},$$

$$\mathcal{J} = \{ j \in \mathbb{Z} : \sigma_{\min} < 1.3^j \le \sigma_{\max} \},$$

and let $\psi(\cdot)$ be the function defined in (3.6). Let $t = \log 10$, and for $j \in \mathcal{J}$, set $\theta_j = \sqrt{\frac{2t}{n}} \frac{1}{1.3^j}$ and $\hat{\Sigma}_{n,j} = \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi(\theta_j Z_i^0 {Z_i^0}^T)$. Finally, define

$$j_* := \min\left\{ j \in \mathcal{J} : \forall k > j, \|\hat{\Sigma}_{n,k} - \hat{\Sigma}_{n,j}\| \le 1.3^k \sqrt{\frac{t}{n}} \right\}$$

(note that we modified some constants compared to the "theoretical" version), and finally set $\widehat{\Sigma}_n := \hat{\Sigma}_{n,j_*}$.



(a) Covariance matrix estimation error (b) First principal component estimation error

FIG. 1. *Sample size $n = 100$, dimension $d = 100$.*

Quality of covariance estimation was evaluated via comparing $\frac{\|\widehat{S}_n - \Sigma\|}{\|\Sigma\|}$ with $\frac{\|\widehat{\Sigma}_n - \Sigma\|}{\|\Sigma\|}$ over 500 runs of simulations. We also compared errors of estimation of projectors onto the first principal component,

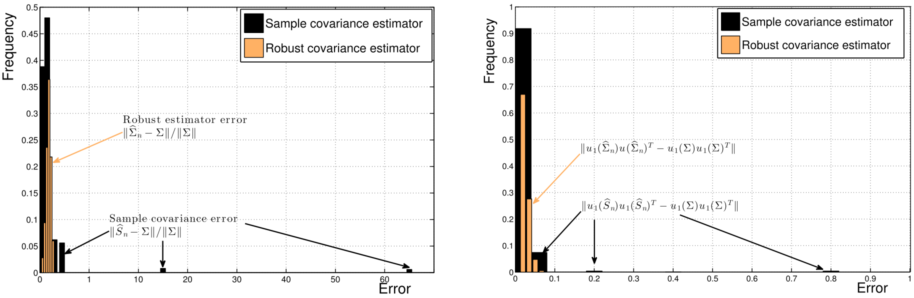
$$\|u_1(\widehat{S}_n)u_1(\widehat{S}_n)^T - u_1(\Sigma)u_1(\Sigma)^T\| \quad \text{and} \quad \|u_1(\widehat{\Sigma}_n)u_1(\widehat{\Sigma}_n)^T - u_1(\Sigma)u_1(\Sigma)^T\|,$$

where $u_1(\cdot)$ denotes the eigenvector corresponding to the largest eigenvalue of a matrix. Histograms illustrating performance of both estimators are presented in Figure 1(a) and (b) (for the sample size $n = 100$), and in Figure 2(a) and (b) (for the sample size $n = 1000$). It is clear from the graphs that in all scenarios, $\widehat{\Sigma}_n$ performs significantly better than $\widehat{S}_n$.

APPENDIX: SUPPLEMENTARY RESULTS

LEMMA A.1. *Let $F : \mathbb{R} \mapsto \mathbb{R}$ be a continuously differentiable function, and $S \in \mathbb{C}^{d\times d}$ be a self-adjoint matrix. Then the gradient of $G(S) := \operatorname{tr} F(S)$ is*

$$\nabla G(S) = F'(S),$$



(a) Covariance matrix estimation error (b) First principal component estimation error

FIG. 2. *Sample size $n = 1000$, dimension $d = 100$.*

where F' is the derivative of F and $F'(S) : \mathbb{C}^{d\times d} \mapsto \mathbb{C}^{d\times d}$ is the matrix function in the sense of Definition 2.1.

PROOF. We will first check the claim assuming that F is a polynomial of the form $F(x) = x^k, k \in \mathbb{N}$. Let $H = H^*$ be a self-adjoint operator, and consider the directional derivative $dG(S; H)$ of G in direction H:

$$\begin{aligned} dG(S; H) &= \lim_{t\to 0} \frac{1}{t} \operatorname{tr}\big((S + tH)^k - S^k\big) = \sum_{j=1}^{k} \operatorname{tr}\big(S^{j-1} H S^{k-j}\big) \\ &= \operatorname{tr}\big(kS^{k-1}H\big) = \big\langle F'(S), H\big\rangle, \end{aligned}$$

hence the claim holds for monomials. By linearity, it also holds for arbitrary polynomials. It remains to extend the claim to arbitrary continuously differentiable function via a standard approximation argument (for instance, see [4], Chapter 5, Section 3). □

LEMMA A.2. *Let $1 < \alpha \le 2$ and $c_\alpha = \frac{\alpha-1}{\alpha} \vee \sqrt{\frac{2-\alpha}{\alpha}}$. Then $1 + y + c_\alpha|y|^\alpha > 0$ and*

$$-\log\big(1 + y + c_\alpha|y|^\alpha\big) \le \log\big(1 - y + c_\alpha|y|^\alpha\big) \qquad \textit{for all } y \in \mathbb{R}.$$

PROOF. To check the first claim, it is enough to note that $f(y) = 1 + y + c_\alpha|y|^\alpha$ is convex and its minimum is attained for $y_m = -(\frac{1}{\alpha c_\alpha})^{1/(\alpha-1)}$. It is easy to check that $f(y_m) = 1 - y_m + \frac{y_m}{\alpha}$, which implies that $f(y_m) > 0 \iff c_\alpha > \frac{\alpha-1}{\alpha^2}$ which always holds since $c_\alpha \ge \frac{\alpha-1}{\alpha}$ and $\alpha > 1$.

For the second part, it is enough to show that $(1+c_\alpha|y|^\alpha+y)(1+c_\alpha|y|^\alpha-y) \ge 1$ for all $y \in \mathbb{R}$, which is equivalent to claiming that $c_\alpha^2 y^{2\alpha} + 2c_\alpha y^\alpha \ge y^2$, $y \ge 0$. Note that for any $\tau \in (-1, 1)$, $p, q > 0$ such that $1/p + 1/q = 1$, and $y \ge 0$,

$$y^2 = y^{1-\tau} y^{1+\tau} \le \frac{y^{p(1-\tau)}}{p} + \frac{y^{q(1+\tau)}}{q}.$$

Choosing $p := \frac{\alpha}{2(\alpha-1)}$, $q := \frac{\alpha}{2-\alpha}$, we get $y^2 \le \frac{2(\alpha-1)}{\alpha} y^\alpha + \frac{2-\alpha}{\alpha} y^{2\alpha}$ which is further bounded above by $2c_\alpha y^\alpha + c_\alpha^2 y^{2\alpha}$ for $c_\alpha = \frac{\alpha-1}{\alpha} \vee \sqrt{\frac{2-\alpha}{\alpha}}$. □

LEMMA A.3. *Functions $\psi_1(x)$ and $\psi_2(x)$ defined in Remark 1 are operator Lipschitz, with Lipschitz constants independent of the dimension.*

PROOF. The Lipshitz property of $\psi_1(x)$ follows from Theorem 1.6.1 in [2]. The result for $\psi_2(x)$ follows from Theorem 1.1.1 in the same paper. □

A.1. Proof of Lemma 2.1. For a self-adjoint matrices R, Q, $\|R\| \geq \|Q\|$ iff $\|R^2\| \geq \|Q^2\|$. Clearly,

$$\begin{pmatrix} S & A \\ A^* & T \end{pmatrix}^2 = \begin{pmatrix} S^2 + AA^* & SA + AT \\ A^*S + TA^* & T^2 + A^*A \end{pmatrix}.$$

It implies that $\|(\begin{smallmatrix} S & A \\ A^* & T \end{smallmatrix})^2\| \geq \|S^2 + AA^*\| \geq \|AA^*\|$ and $\|(\begin{smallmatrix} S & A \\ A^* & T \end{smallmatrix})^2\| \geq \|T^2 + A^*A\| \geq \|A^*A\|$. Since $(\begin{smallmatrix} 0 & A \\ A^* & 0 \end{smallmatrix})^2 = (\begin{smallmatrix} AA^* & 0 \\ 0 & A^*A \end{smallmatrix})$, we obtain

$$\left\|\begin{pmatrix} S & A \\ A^* & T \end{pmatrix}^2\right\| \geq \left\|\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}^2\right\|,$$

and the result follows.

Acknowledgements. I want to thank L. Goldstein, A. Juditsky, A. Nemirovski, as well as the anonymous referees and the Associate Editor for their insightful suggestions that helped to improve the quality of presentation.

SUPPLEMENTARY MATERIAL

Supplementary material for the paper: Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries (DOI: 10.1214/17-AOS1642SUPP; .pdf). The supplement contains technical details and proofs not included in the main text of the paper.

REFERENCES

[1] AHLSWEDE, R. and WINTER, A. (2002). Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory* **48** 569–579. MR1889969

[2] ALEKSANDROV, A. B. and PELLER, V. V. (2016). Operator Lipschitz functions. *Russian Math. Surveys* **71** 605.

[3] ALON, N., MATIAS, Y. and SZEGEDY, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing* 20–29. ACM, New York.

[4] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[5] BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602

[6] BUTLER, R. W., DAVIES, P. L. and JHUN, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.* **21** 1385–1400. MR1241271

[7] CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59.

[8] CAI, T. T., ZHANG, C. H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[9] CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000

[10] CARLEN, E. (2010). Trace inequalities and quantum entropy: An introductory course. Available at http://www.mathphys.org/AZschool/material/AZ09-carlen.pdf.

[11] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

[12] CATONI, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. Preprint. Available at arXiv:1603.05229.

[13] DAVIES, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* MR1193314 1828–1843.

[14] DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2015). Sub-Gaussian mean estimators. Preprint. Available at arXiv:1509.05845.

[15] FAN, J., WANG, W. and ZHONG, Y. (2016). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. Preprint. Available at arXiv:1603.03516.

[16] FAN, J., WANG, W. and ZHU, Z. (2016). Robust low-rank matrix recovery. Preprint. Available at arXiv:1603.08315.

[17] GIULINI, I. (2015). PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. Preprint. Available at arXiv:1511.06263.

[18] HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

[19] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101.

[20] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ.

[21] HUBERT, M., ROUSSEEUW, P. J. and VAN AELST, S. (2008). High-breakdown robust multivariate methods. *Statist. Sci.* **23** 92–119. MR2431867

[22] JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188.

[23] JOLY, E., LUGOSI, G. and OLIVEIRA, R. I. (2017). On the estimation of the mean of a random vector. *Electron. J. Stat.* **11** 440–451. MR3619312

[24] KLOPP, O., LOUNICI, K. and TSYBAKOV, A. B. (2017). Robust matrix completion. *Probab. Theory Related Fields* **169** 523–564. MR3704775

[25] KOLTCHINSKII, V. and LOUNICI, K. (2016). New asymptotic results in principal component analysis. Preprint. Available at arXiv:1601.01457.

[26] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133.

[27] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[28] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278.

[29] LEPSKI, O. (1992). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

[30] LERASLE, M. and OLIVEIRA, R. I. (2011). Robust empirical mean estimators. Preprint. Available at arXiv:1112.3914.

[31] LIEB, E. H. (1973). Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.* **11** 267–288. MR0332080

[32] LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058.

[33] LUGOSI, G. and MENDELSON, S. (2017). Sub-Gaussian estimators of the mean of a random vector. Preprint. Available at arXiv:1702.00482.

[34] MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. MR0388656

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

[36] MINSKER, S. (2018). Supplement to “Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries.” DOI:10.1214/17-AOS1642SUPP.

[37] MINSKER, S. and WEI, X. (2017). Estimation of the covariance structure of heavy-tailed distributions. Preprint. Available at arXiv:1708.00502.

[38] NEMIROVSKI, A. and YUDIN, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York.

[39] OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Preprint. Available at arXiv:0911.0600.

[40] SRIVASTAVA, N. and VERSHYNIN, R. (2013). Covariance estimation for distributions with $2 + \varepsilon$ moments. *Ann. Probab.* **41** 3081–3111. MR3127875

[41] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[42] TROPP, J. A. (2015). An introduction to matrix concentration inequalities. Preprint. Available at arXiv:1501.01571.

[43] TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734

[44] VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027.

[45] ZHANG, T., CHENG, X. and SINGER, A. (2016). Marčenko–Pastur law for Tyler's M-estimator. *J. Multivariate Anal.* **149** 114–123. MR3507318

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: minsker@usc.edu