Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries

Stanislav Minsker*
e-mail: minsker@usc.edu

Abstract: 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 outliers in the data. As P. Huber wrote [Hub64], "...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 this 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 will explain the key ideas behind our construction, as well as applications to covariance estimation and matrix completion problems.

1. Introduction.

Let $Y_1, \ldots, Y_n \in \mathbb{C}^{d_1 \times d_2}$ 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 $1 \leq j \leq n$, $1 \leq k \leq d_1$, $1 \leq l \leq d_2$. Let $EY_1, \ldots, EY_n \in \mathbb{C}^{d_1 \times d_2}$ be the expectations evaluated element-wise, meaning that $(EY_j)_{k,l} = \mathbb{E}(Y_j)_{k,l}$. The goal of this paper is to construct and study estimators of $\overline{EY} := \mathbb{E}\left[\frac{1}{n} \sum_{j=1}^{n} Y_j\right]$ under minimal assumptions on the distributions of $Y_1, \ldots, Y_n$. In particular, we are interested in the estimators that admit tight non-asymptotic bounds and exponential deviation inequalities without imposing any additional assumptions (besides finite second moments) on $Y_1, \ldots, Y_n$. For example, if $Y_j = Z_jZ_j^T$, where $Z_1, \ldots, Z_n \in \mathbb{R}^d$ 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.

Techniques developed in this paper have direct applications to some 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., Lasso 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 with many real data sets? Can the assumptions be weakened without sacrificing the quality of theoretical guarantees? We look at examples

*Department of Mathematics, University of Southern California
where the answer is positive, and describe modifications of existing techniques necessary 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 $X_1, \ldots, X_n \in \mathbb{R}$ 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 unknown mean $\mathbb{E}X$ based on a given sample.

A surprising fact (dating back to [NY83] where the “median of means” estimator was introduced, along with [AMS96]) is that it’s possible to construct a non-asymptotic confidence intervals $I_n(\delta)$ with coverage probability $1 - \delta$ (meaning that $\Pr(\mathbb{E}X \in I_n(\delta)) \geq 1 - \delta$ for given $n$ and $\delta$) and “nearly optimal” length $|I_n(\delta)| \leq L \sqrt{\text{Var}(X)} \sqrt{\frac{\log(e/\delta)}{n}}$, where $L > 0$ is an absolute constant. An in-depth study of this and closely related questions was performed in [Cat12, DLLO15] based on two different approaches. Note that the center of any such confidence interval is a point estimator $\hat{\mu} := \hat{\mu}(X_1, \ldots, X_n, \delta)$ that satisfies $\Pr(|\hat{\mu} - \mathbb{E}X| \geq L \sqrt{\text{Var}(X)} \sqrt{\frac{\log(e/\delta)}{n}}) \leq \delta$. Because the only assumption on $X$ is the existence of a second moment, it is natural to call such an estimator “robust” \(^1\); it admits strong deviation bounds even for 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 [LO11, BJL15] which covers a wide range of statistical applications. Let us emphasize that these results 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 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 [HS16] and [Min15] (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 “sub-Gaussian - type” deviation guarantees in the operator norm that we are mainly interested in (see (3.7) below for an example of such an estimator). Under more restrictive assumptions on the sequence of random matrices $Y_1, \ldots, Y_n$ (such as $\|Y_j\| \leq M$ almost surely, $j = 1, \ldots, n$), 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 [AW02, Oli09, Tro12a].

A closely related covariance 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 [KL14, KL16, Ver10, CZZ10, CRZ16] and references therein; under weaker moment assumptions, sample covariance estimator has been studied in [SV13]. Some popular robust estimators of scatter are discussed in [HRVA08], including the Minimum Covariance

\(^1\)For the classical treatment of robust estimators based on the notion of a breakdown point, we refer the reader to [HR09].
Determinant (MCD) estimator and the Minimum Volume Ellipsoid estimator (MVE). However, rigorous results for these estimators are available only for elliptically symmetric distributions; see [BDJ93] for results on MCD and [Dav92] for results on MVE. Popular Maronna’s [Mar76] and Tyler’s [Tyl87, ZCS16] 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 [Cat16] and I. Guilini [Giu15] are closest in spirit to our work. For instance, in [Cat16] 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 [Cat12]. 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 uniform upper bound on the kurtosis of $\langle Z, u \rangle^2$).

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, in [CLMW11, KLT14] authors investigate robust matrix completion under the “low rank + sparse” model. In [FWZ16], authors study low-rank matrix recovery under the assumption that the additive noise has $(2+\varepsilon)$ moments, and obtain strong results via truncation argument. We propose a different approach based on general techniques developed in this paper and achieve related results for matrix completion 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. Sections 5 and 6 discuss adaptation to unknown parameters that appear in our construction, as well as improvements of the initial estimator via iterative procedure. As a by-product of our results, we develop a variant of PAC-Bayesian uniform deviation bound in the noncommutative framework; it is discussed in section 5.2.1.

Appendices A and B contain technical results and complements to the main text, and appendix C includes numerical simulation results that illustrate the advantages of proposed techniques.

2. Preliminaries.

In this section, we introduce 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 $\lambda_{\text{max}}(A)$ and $\lambda_{\text{min}}(A)$ 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 $\|A\| := \sqrt{\lambda_{\text{max}}(A^*A)}$. If $d_1 = d_2 = d$, we denote by 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
\[ \|A\{1} = \text{tr} (\sqrt{A^*A}) \text{, where } \sqrt{A^*A} \text{ is a nonnegative definite matrix such that } (\sqrt{A^*A})^2 = A^*A. \] The Frobenius (or Hilbert-Schmidt) norm is \[ \|A\|_F = \sqrt{\text{tr}(A^*A)} \text{, and the associated inner product is } \langle A_1, A_2 \rangle = \text{tr}(A_1^*A_2). \] Finally, set \[ \|A\|_{\text{max}} := \sup_{i,j} |a_{ij}|. \] 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 \( Y_1, \ldots, Y_n \) of random matrices, \( \mathbb{E}_j [\cdot] \) will stand for the conditional expectation \( \mathbb{E}[\cdot | Y_1, \ldots, Y_j] \).

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 and probability theory.

In this section, we collect the 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 \( \lambda_j(A) \in \mathbb{T}, \ j = 1, \ldots, d \), define \( f(A) \) as \( f(A) = Uf(\Lambda)U^* \), where

\[
  f(\Lambda) = f \begin{pmatrix} 
    \lambda_1 \\
    \vdots \\
    \lambda_d 
  \end{pmatrix} = \begin{pmatrix} 
    f(\lambda_1) \\
    \vdots \\
    f(\lambda_d) 
  \end{pmatrix}.
\]

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 \( f_1(\lambda_j) \geq f_2(\lambda_j) \) for \( j = 1, \ldots, d \). 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 \( \lambda_j(A) \geq \lambda_j(B) \), \( j = 1, \ldots, d \), where \( \lambda_j(\cdot) \) stands for the \( j \)-th largest eigenvalue. Moreover, \( \text{tr} e^A \geq \text{tr} e^B \).

**Fact 2.3.** Matrix logarithm is operator monotone: if \( A \succeq 0 \), \( B \succeq 0 \) and \( A \succeq B \), then \( \log(A) \succeq \log(B) \).

**Proof.** See [Bha97].

**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 \( -\log(1 + t + t^2/2) \leq \log(1 - t + t^2/2) \) for \( t \in \mathbb{R} \).

**Fact 2.5** (Lieb’s concavity theorem). Given a fixed self-adjoint matrix \( H \), the function

\[
  A \mapsto \text{tr} \exp(H + \log(A))
\]

is concave on the cone of positive definite matrices.
Proof. See [Lie73] and [Tro15].

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

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

Proof. This is a consequence of Peierls inequality, see Theorem 2.9 in [Car10] 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 $H : \mathbb{C}^{d_1 \times d_2} \mapsto \mathbb{C}^{(d_1+d_2) \times (d_1+d_2)}$ is defined as

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

(2.1)

Since $H(A)^2 = \begin{pmatrix} AA^* & 0 \\
0 & A^*A \end{pmatrix},$ it is easy to see that $\|H(A)\| = \|A\|.$ 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.

The following lemma says that a scalar Lipschitz function remains Lipschitz on the space of Hermitian matrices:

**Lemma 2.2.** Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a Lipschitz-continuous function with Lipschitz constant $L,$ and let $A, B$ be two Hermitian matrices. Then

$$
\|f(A) - f(B)\|_F \leq L \|A - B\|_F.
$$

Proof. Lemma VII.5.5 in [Bha97] states that for all matrices $C$ and normal matrices $A, B,$

$$
\|f(A)C - Cf(B)\|_F \leq L \|AC - CB\|_F.
$$

Let $C = I$ to get the result.

The remaining tools presented below are needed only in the technical proofs of section 6.2, and are not crucial for the rest of the paper.

**Lemma 2.3** (Matrix Hoeffding inequality). Let $Z_1, \ldots, Z_n \in \mathbb{C}^{d \times d}$ be a sequence of independent self-adjoint random matrices such that for all $1 \leq k \leq n,$

$$
\mathbb{E}Z_k = 0 \text{ and } \|Z_k\| \leq M_k \text{ almost surely.}
$$

Then $\left\| \sum_{j=1}^n Z_j \right\| \leq t$ with probability $\geq 1 - 2d \exp \left( -\frac{t^2}{8 \sum_{j=1}^n M_j^2} \right).$
Proof. See Theorem 1.3 in [Tro12b].

We conclude this section by recalling the notion of Talagrand’s generic chaining complexity (see [Tal14]) and several related results. Given a metric space \((T, \rho)\), let \(\{\Delta_n\}\) be a nested sequence of partitions of \(T\) such that \(\text{card } \Delta_0 = 1\) and \(\text{card } \Delta_n \leq 2^{2^n}\). For \(s \in T\), let \(\Delta_n(s)\) be the unique subset of \(\Delta_n\) containing \(s\). The generic chaining complexity \(\gamma_2(T, \rho)\) is defined as

\[
\gamma_2(T, \rho) := \inf_{\{\Delta_n\}} \sup_{s \in T} \sum_{n \geq 0} 2^{\frac{n}{2}} D(\Delta_n(s))
\]

where the infimum is taken over all admissible sequences of partitions and \(D(A) := D(A, \rho)\) stands for the diameter of a set \(A\).

Dudley’s entropy integral bound (see [Tal14]) states that

\[
\gamma_2(T, \rho) \leq \frac{1}{2\sqrt{2} - 1} \int_0^{D(T)} \sqrt{\log N(T, \rho, \varepsilon/4)} d\varepsilon.
\]

(2.2)

We will say that \(\mathbb{C}^{d \times d}\)-valued stochastic process \(\{X(t), \ t \in T\}\) has sub-Gaussian increments with respect to the metric \(\rho\) if for all \(t_1, t_2 \in T\),

\[
\Pr (\|X_{t_1} - X_{t_2}\| \geq s\rho(t_1, t_2)) \leq 2de^{-s^2/2},
\]

where \(\|\cdot\|\) is the operator norm.

**Lemma 2.4.** Let \((T, \rho)\) be a metric space and let \(\mathbb{C}^{d \times d}\)-valued stochastic process \(\{X(t), \ t \in T\}\) have sub-Gaussian increments with respect to \(\rho\). There exists an absolute constant \(C > 0\) such that for any \(t_0 \in T\) and any \(s \geq 1\),

\[
\sup_{t \in T} \|X_t - X_{t_0}\| \leq C (\gamma_2(T, \rho) + \sqrt{s}D(T))
\]

with probability \(\geq 1 - 2de^{-s}\).

**Proof.** See Theorem 3.2 in [Dir13].

3. Main results.

Our construction has its roots in the technique proposed by O. Catoni [Cat12] for estimation of the univariate mean. Let us briefly recall the main ideas of Catoni’s approach. Assume that \(\xi, \xi_1, \ldots, \xi_n\) is a sequence of i.i.d. random variables such that \(\mathbb{E} \xi = \mu\) and \(\text{Var}(\xi) \leq \nu^2\). Catoni’s estimator is defined as follows: let \(\psi(x) : \mathbb{R} \mapsto \mathbb{R}\) be a non-decreasing function such that for all \(x \in \mathbb{R}\)

\[
- \log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2).
\]

(3.1)

See remark 2 below for examples of such functions. Given \(\theta > 0\), let \(\hat{\mu}_\theta\) be such that

\[
\sum_{j=1}^n \psi (\theta (\xi_j - \hat{\mu}_\theta)) = 0
\]

(3.2)
(clearly, $\hat{\mu}_\theta$ always exists). 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 [Cat12] that $|\hat{\mu}_\theta - \mu| \leq \eta$ with probability $\geq 1 - 2e^{-t}$.

While direct extension of this technique to the case of random matrices does not seem to be straightforward, the following approach succeeds in a more general framework. We will first describe its univariate version: assume that $E\xi^2 \leq \kappa^2$, let $\tilde{\theta} = \sqrt{\frac{2t}{n\kappa^2}}$, and define

$$\tilde{\mu} = \frac{1}{n\tilde{\theta}} \sum_{j=1}^n \psi (\tilde{\theta} \xi_j).$$

It is then not hard to show (see our more general results below) that $|\tilde{\mu} - \mu| \leq \kappa \sqrt{\frac{2t}{n}}$ with probability $\geq 1 - 2e^{-\tilde{\theta}}$ for any $\psi$ satisfying (3.1).

**Remark 1.** The assumption that $\psi$ is non-decreasing can be dropped in the definition of $\tilde{\mu}$ (although it is useful in the definition of Catoni’s estimator $\hat{\mu}_\theta$), and we will only require $\psi$ to satisfy inequality (3.1) in the rest of the paper.

The disadvantage of $\tilde{\mu}$ when compared to $\hat{\mu}_\theta$, is that the variance $v^2$ is replaced by the second moment $\kappa^2$. However, this problem can be alleviated via an iterative procedure described in section 6, which also establishes the connections to the matrix version of equation (3.2).

We proceed by presenting an extension of the estimator $\tilde{\mu}$ to the noncommutative case. We first state the results for self-adjoint matrices and will later deduce the general case of rectangular matrices as a corollary. Let $Y_1, \ldots, Y_n \in \mathbb{C}^{d \times d}$ be a sequence of independent self-adjoint random matrices such that $\sigma^2_n := \left\| \sum_{j=1}^n E Y_j^2 \right\| < \infty$. Given $\theta > 0$, and a function $\psi$ satisfying (3.1), set

$$X_j := \psi (\theta Y_j), \quad j = 1, \ldots, n.$$

If $\theta$ is small and $\psi$ is smooth, then $X_j \simeq \theta Y_j$, so it is natural to estimate $\frac{1}{n} \sum_{j=1}^n E Y_j$ by $\frac{1}{n\theta} \sum_{j=1}^n X_j$ for an appropriate choice of $\theta$. In the following sections (see Theorem 3.1), we make this intuition formal by presenting the deviation bounds for $\left\| \frac{1}{n\theta} \sum_{j=1}^n (\psi (\theta Y_j) - \theta E Y_j) \right\|$.

**Remark 2.** Most of our results do not depend on the concrete choice of the function $\psi$ satisfying (3.1). One possible choice is $\psi(x) = \log \left( 1 + x + \frac{x^2}{2} \right)$. Another useful choice is

$$\psi(x) = \begin{cases} \log(2), & x \geq 1, \\ -\log \left( 1 - x + \frac{x^2}{2} \right), & x \in [0,1), \\ \log \left( 1 + x + \frac{x^2}{2} \right), & x \in (-1,0), \\ -\log(2), & x \leq -1. \end{cases}$$

(3.3)

Since the latter function is bounded, it can provide additional advantages (such as robustness) in applications. Simple truncation $\psi_\tau(x) = (|x| \wedge \tau) \text{sign}(x)$ does not satisfy (3.1) for any $\tau$. However,

$$-\log \left( 1 - x + x^2 \right) \leq (|x| \wedge 1) \text{sign}(x) \leq \log \left( 1 + x + x^2 \right),$$

hence all of our results extend to truncations, albeit with slightly worse constant factors.
3.1. Bounds for the moment generating function

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:

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n} (X_j - \theta \mathbb{E} Y_j) \right) \leq \exp \left( \frac{\theta^2}{2} \sum_{j=1}^{n} \mathbb{E} Y_j^2 \right), \quad (3.4)
\]

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n} (\theta \mathbb{E} Y_j - X_j) \right) \leq \exp \left( \frac{\theta^2}{2} \sum_{j=1}^{n} \mathbb{E} Y_j^2 \right). \quad (3.5)
\]

**Proof.** Note that

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n} (X_j - \theta \mathbb{E} Y_j) \right) =
\]

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n-1} (X_j - \theta \mathbb{E} Y_j) \right) + \psi(\theta Y_n) \leq
\]

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n-1} (X_j - \theta \mathbb{E} Y_j) \right) + \log \left( I + \theta \mathbb{E} Y_n + \frac{\theta^2}{2} \mathbb{E} Y_n^2 \right) \leq
\]

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n-1} (X_j - \theta \mathbb{E} Y_j) + \log \left( I + \theta \mathbb{E} Y_n + \frac{\theta^2}{2} \mathbb{E} Y_n^2 \right) - \theta \mathbb{E} Y_n \right),
\]

where the first inequality follows from the semidefinite relation \( \psi(\theta Y_n) \leq \log \left( I + \theta \mathbb{E} Y_n + \frac{\theta^2}{2} \mathbb{E} Y_n^2 \right) \) 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 \( I + \theta \mathbb{E} Y_n + \frac{\theta^2}{2} \mathbb{E} Y_n^2 \geq 0 \) since \( I + \theta \mathbb{E} Y_n + \frac{\theta^2}{2} \mathbb{E} Y_n^2 \geq 0 \) almost surely, hence \( \log \left( I + \theta \mathbb{E} Y_n + \frac{\theta^2}{2} \mathbb{E} Y_n^2 \right) \) is well-defined. Repeating the steps for \( X_{n-1}, \ldots, X_1 \), we obtain the inequality

\[
\mathbb{E} \exp \left( \sum_{j=1}^{n} (X_j - \theta \mathbb{E} Y_j) \right) \leq \exp \left( \sum_{j=1}^{n} \log \left( I + \theta \mathbb{E} Y_j + \frac{\theta^2}{2} \mathbb{E} Y_j^2 \right) - \theta \mathbb{E} Y_j \right) \quad (3.6)
\]

It remains to note that by fact 2.1 and the inequality \( \log(1+x) \leq x \) (\( \forall x > -1 \)), for all \( j = 1, \ldots, n \)

\[
\log \left( I + \theta \mathbb{E} Y_j + \frac{\theta^2}{2} \mathbb{E} Y_j^2 \right) \leq \theta \mathbb{E} Y_j + \frac{\theta^2}{2} \mathbb{E} Y_j^2,
\]
or \( \log \left( I + \theta EY_j + \theta^2 EY_j^2 / 2 \right) - \theta EY_j \leq \frac{\theta^2}{2} EY_j^2 \). The first inequality (3.4) now follows from (3.6) and fact 2.2.

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

\[
\mathbb{E} \operatorname{tr} \exp \left( \sum_{j=1}^{n} (\theta EY_j - X_j) \right) \leq \mathbb{E} \operatorname{tr} \exp \left( \sum_{j=1}^{n} \left( \log \left( I + \theta (-Y_j) + \theta^2 Y_j^2 / 2 \right) - \theta \mathbb{E}(-Y_j) \right) \right),
\]

and apply inequality (3.4) to the sequence \(-Y_1, \ldots, -Y_n\) with \(X_j = \log \left( I + \theta (-Y_j) + \theta^2 (-Y_j)^2 / 2 \right)\), \(j = 1, \ldots, n\).

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

**Theorem 3.1.** Let \(Y_1, \ldots, Y_n \in \mathbb{C}^{d \times d}\) be a sequence of independent self-adjoint random matrices, and \(\sigma_n^2 \geq \left\| \sum_{j=1}^{n} EY_j^2 \right\|\). Then

\[
\Pr \left( \left\| \sum_{j=1}^{n} \left( \frac{1}{\theta} \psi(\theta Y_j) - EY_j \right) \right\| \geq t\sqrt{n} \right) \leq 2d \exp \left( -\theta t \sqrt{n} + \frac{\theta^2 \sigma_n^2}{2} \right),
\]

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

**Remark 3.** In the important special case when \(Y_j, j = 1, \ldots, n\) are i.i.d. copies of \(Y\), we will often use the following equivalent form of the bound: assume that \(\sigma^2 \geq \|EY^2\|\), then replacing \(t\) by \(s \sqrt{n} / \sigma\) and setting \(\theta = \sqrt{\frac{1}{n}\sigma}\) implies that

\[
\Pr \left( \left\| \frac{1}{n\theta} \sum_{j=1}^{n} \psi(\theta Y_j) - EY \right\| \geq \sigma \sqrt{\frac{s}{n}} \right) \leq 2d \exp \left( -s / 2 \right). \tag{3.7}
\]

**Proof.** As before, set \(X_j := \psi(\theta Y_j), j = 1, \ldots, n\). Then

\[
\Pr \left( \lambda_{\max} \left( \frac{1}{\theta} \sum_{j=1}^{n} (X_j - \theta EY_j) \right) \geq s \right) = \Pr \left( \exp \left( \lambda_{\max} \left( \sum_{j=1}^{n} (X_j - \theta EY_j) \right) \right) \geq e^s \right)
\]

\[
\leq e^{-\theta s} \mathbb{E} \operatorname{tr} \exp \left( \sum_{j=1}^{n} (X_j - \theta EY_j) \right) \leq e^{-\theta s} \operatorname{tr} \exp \left( \frac{\theta^2}{2} \sum_{j=1}^{n} EY_j^2 \right)
\]

\[
\leq d \exp \left( -\theta s + \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} EY_j^2 \right\| \right),
\]
where we used Chebyshev’s inequality, the fact that $e^{\lambda_{\max}(A)} = \lambda_{\max}(e^A)$ and the inequality $\lambda_{\max}(e^A) \leq \text{tr } e^A$ on the second step, the first inequality of Lemma 3.1 on the third step, and the bound $\text{tr } e^A \leq d e^{\|A\|}$ 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

$$
\Pr \left( \lambda_{\min} \left( \frac{1}{\theta} \sum_{j=1}^{n} (X_j - \theta E Y_j) \right) \leq -s \right) = \Pr \left( \lambda_{\max} \left( \frac{1}{\theta} \sum_{j=1}^{n} (\theta E Y_j - X_j) \right) \geq s \right)
$$

$$
\leq e^{-\theta s} \text{tr } \exp \left( \sum_{j=1}^{n} (\theta E Y_j - X_j) \right) \leq e^{-\theta s} \text{tr } \exp \left( \frac{\theta^2}{2} \sum_{j=1}^{n} E Y_j^2 \right)
$$

$$
\leq d \exp \left( -\theta s + \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} E Y_j^2 \right\| \right),
$$

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 discussed above is the fact that the “variance term” $\left\| \sum_{j=1}^{n} E Y_j^2 \right\|$ is akin to the second moment while we would like to replace it by $\left\| \sum_{j=1}^{n} E (Y_j - E Y_j)^2 \right\|$. This problem will be addressed in detail in section 6. In particular, results of that section imply the following: assume that $Y_1, \ldots, Y_n$ are i.i.d. copies of $Y$, $\sigma_0^2 \geq E(Y - EY)^2$, $\theta_0 = \sqrt{\frac{2t}{n}} \frac{1}{\sigma_0}$, and $n$ is large enough ($n \geq d^2$). Then, with exponentially high (in $t$) probability, there exists a solution $S$ of the equation

$$
\sum_{j=1}^{n} \psi(\theta_0 (Y_j - S)) = 0, \ S \in K = \{S : \left\| S - EY \right\| \leq 1 \},
$$

moreover, any such solution must satisfy $\left\| \tilde{S} - EY \right\| \leq C \sigma_0 \sqrt{\frac{2t}{n}}$ for an absolute constant $C > 0$. However, in some applications (such as matrix completion, see section 4) even the estimator with “suboptimal” variance term suffices to get good bounds.

Another problem is the fact that one needs to know the value of $\left\| \sum_{j=1}^{n} E Y_j^2 \right\|$ (or its tight upper bound) a priori to choose the “optimal” value of $\theta$. This issue and its resolution based on adaptive estimators is discussed in section 5. We conclude this section with few additional comments.

Remark 4. 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 $\theta$ depends on it. However, as it was noted in [DLLO15], this is sufficient to construct (via Lepski’s method [Lep92]) estimators that admit sub-Gaussian tails uniformly over $t$ in a certain range. We discuss the details in appendix B.

2. Let $Y_1, \ldots, Y_n \in \mathbb{C}^{d \times d}$ be i.i.d. copies of $Y$, and $\sigma_0^2 = \left\| E(Y - EY)^2 \right\|$. It is interesting to compare our estimator (in particular, bound (3.7)) to the guarantees for the sample mean $\frac{1}{n} \sum_{j=1}^{n} Y_j$. Under an additional restrictive boundedness assumption requiring that $\|Y\| \leq M$ almost surely, the noncommutative Bernstein’s inequality (see Theorem 1.4 in [Tro12b]) implies that $\left\| \frac{1}{n} \sum_{j=1}^{n} Y_j - EY \right\| \leq 2\sigma_0 \sqrt{\frac{t}{n}} \vee \frac{1}{3} \frac{M t}{n}$ with probability $\geq 1 - 2de^{-t/2}$. Hence, even
under additional strong assumptions our technique allows to obtain guarantees that compare favorably to the sample mean.

3.2. Bounds depending on the effective dimension.

The bound obtained in Theorem 3.1 explicitly depends on the dimension $d$ of random matrices. It is possible to replace it by the “effective dimension” defined as

$$d := \frac{\text{tr} \left( \sum_{j=1}^{n} EY_j^2 \right)}{\left\| \sum_{j=1}^{n} EY_j^2 \right\|}$$

(3.8)

which can be much smaller than $d$ if $\sum_{j=1}^{n} EY_j^2$ has many small eigenvalues. The following result holds:

Theorem 3.2. Let $Y_1, \ldots, Y_n \in \mathbb{C}^{d \times d}$ be a sequence of independent self-adjoint random matrices, and $\sigma_n^2 \geq \left\| \sum_{j=1}^{n} EY_j^2 \right\|$. Then

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

Remark 5. As before, we can set $\theta = \frac{t\sqrt{n}}{\sigma_n}$ to get

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

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

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

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

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

Proof. See section A.2.

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 $Y_1, \ldots, Y_n \in \mathbb{C}^{d_1 \times d_2}$ be independent, and assume that

$$\sigma_n^2 \geq \max \left( \left\| \sum_{j=1}^{n} EY_j^* Y_j \right\|, \left\| \sum_{j=1}^{n} EY_j Y_j^* \right\| \right).$$
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 $\hat{T} \in \mathbb{C}^{(d_1 + d_2) \times (d_1 + d_2)}$ as

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

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 $\hat{T} = \begin{pmatrix} \hat{T}_{11} & \hat{T}_{12} \\ \hat{T}_{12}^* & \hat{T}_{22} \end{pmatrix}$. Since $\hat{T}$ is "close" to $\sum_{j=1}^{n} \mathcal{H}(\mathbb{E}Y_j)$ for the proper choice of $\theta$, 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,

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

and

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

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

**Proof.** Note that $\left\| \sum_{j=1}^{n} \mathbb{E} \mathcal{H}(Y_j)^2 \right\| = \max \left( \left\| \sum_{j=1}^{n} \mathbb{E}Y_j Y_j^* \right\|, \left\| \sum_{j=1}^{n} \mathbb{E}Y_j^* Y_j \right\| \right) \leq \sigma_n^2$. Theorem 3.1 applied to self-adjoint random matrices $\mathcal{H}(Y_j) \in \mathbb{C}^{(d_1 + d_2) \times (d_1 + d_2)}$, $j = 1, \ldots, n$ implies that

$$\left\| \hat{T} - \sum_{j=1}^{n} \mathcal{H}(\mathbb{E}Y_j) \right\| \leq t \sqrt{n}$$

with probability $\geq 1 - 2(d_1 + d_2) \exp \left( -\frac{\theta^2 \sigma_n^2}{2} \right)$. It remains to apply Lemma 2.1:

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

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

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

since for any $1 \leq j \leq n$, $\text{tr} \left( \mathbb{E}Y_j Y_j^* \right) = \text{E} \text{tr} (Y_j Y_j^*) = \text{E} \text{tr} (Y_j^* Y_j)$.
In a particular case when \( Y \in \mathbb{R}^d \) is a random vector such that \( \mathbb{E} Y Y^T = \Sigma \) and \( Y_1, \ldots, Y_n \) are its i.i.d. copies, max \( \left( \left\| \sum_{j=1}^n \mathbb{E} Y_j Y_j^T \right\|, \left\| \sum_{j=1}^n \mathbb{E} Y_j^* Y_j \right\| \right) = n \text{tr} \Sigma \) and \( \text{tr} \left( \sum_{j=1}^n \mathbb{E} Y_j^* Y_j \right) = n \text{tr} \Sigma \), hence \( \bar{d} = 2 \) and the estimator \( \hat{T}_{12} \) admits the following bound: if we replace \( t \) by \( \sqrt{s} \sqrt{\text{tr} \Sigma} \) and set \( \theta = \sqrt{\frac{1}{n \sqrt{\text{tr} \Sigma}}} \) in the second bound of corollary 3.1, then

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

3.4. Bounds under weaker moment assumptions and further remarks.

In this section, we discuss the mean estimation problem under weaker moment conditions. Namely, assume that \( Y_1, \ldots, Y_n \) are independent self-adjoint random matrices such that \( \left\| \mathbb{E}|Y_j|^\alpha \right\| < \infty \) for all \( 1 \leq j \leq n \). Let \( \psi_\alpha \) satisfy, for all \( x \in \mathbb{R} \) and some \( \alpha \in (1, 2] \),

\[
- \log(1 - x + c_\alpha |x|^{\alpha}) \leq \psi_\alpha(x) \leq \log(1 + x + c_\alpha |x|^{\alpha}),
\]

where \( c_\alpha = \frac{\alpha - 1}{\alpha} \vee \sqrt{\frac{2 - \alpha}{\alpha}} \). The fact that such \( \psi_\alpha \) 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 \( v_\alpha^\alpha \geq \left\| \sum_{j=1}^n \mathbb{E}|Y_j|^\alpha \right\| \). Then for any positive \( t \) and \( \theta \),

\[
\mathbb{P} \left( \left\| \sum_{j=1}^n \left( \frac{1}{\theta} \psi_\alpha(\theta Y_j) - \mathbb{E} Y_j \right) \right\| \geq t \right) \leq 2d \exp \left( -\theta t + c_\alpha \theta^\alpha v_\alpha^\alpha \right).
\]

**Proof.** The argument repeats the steps of lemma 3.1 and theorem 3.1, the only difference being that fact 2.4 is replaced by lemma A.2. \( \square \)

**Remark 6.** In the special case when \( Y_1, \ldots, Y_n \) are i.i.d. copies of \( Y \) with \( v = \left\| \mathbb{E}|Y|^\alpha \right\|^{1/\alpha} \), setting \( t = vn^{1/\alpha} (\frac{a - 1}{\alpha})^{1/(\alpha - 1)} (\frac{a}{n})^{1/\alpha} \frac{1}{v} \) gives the inequality

\[
\mathbb{P} \left( \left\| \frac{1}{n \theta} \sum_{j=1}^n \psi_\alpha(\theta Y_j) - \mathbb{E} Y \right\| \geq v \left( \frac{a}{n} \right)^{\frac{a - 1}{\alpha}} \right) \leq 2d \exp \left( -\frac{a - 1}{\alpha} \left( \frac{1}{ac_\alpha} \right)^{1/(\alpha - 1)} \right).
\]

Note that for \( \alpha = 2 \), we recover (3.7).

Before we proceed with discussion or further improvements and adaptation issues, let us present few 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.

Let \(Z \in \mathbb{R}^d\) be a random vector with \(\mathbb{E}[Z] = \mu, \mathbb{E}\|Z - \mu\|^4_2 < \infty, \Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^T]\), and let \(Z_1, \ldots, Z_{2n}\) be i.i.d. copies of \(Z\). Let us first assume that \(\mu = 0\), and define
\[
\widehat{\Sigma}_{2n}(\theta) = \frac{1}{2n\theta} \sum_{j=1}^{2n} \psi(\theta Z_j Z_j^T),
\]
where \(\psi(\cdot)\) satisfies (3.1). Let \(\sigma^2 \geq \mathbb{E}\|Z\|^2_2 \mathbb{E}(Z^2)\) and \(\hat{\theta} = \sqrt{\frac{t}{n}} \frac{1}{\sigma}\). It is straightforward to deduce from Theorem 3.1 that with probability \(\geq 1 - 2de^{-t}\),
\[
\mathbb{E}\|\widehat{\Sigma}_{2n}(\hat{\theta}) - \Sigma\| \leq \sigma \sqrt{\frac{t}{n}}.
\]

**Remark 7.** Note that for any matrix \(X = \lambda U U^T\) of rank 1 (where \(\|U\|_2 = 1\)),
\[
\psi(X) = \psi(\lambda) U U^T \quad \text{(since \(\psi(0) = 0\))},
\]
hence \(\widehat{\Sigma}_{2n}(\hat{\theta}) = \frac{1}{2n\theta} \sum_{j=1}^{2n} \psi(\hat{\theta} \|Z_j\|_2 Z_j Z_j^T) \|Z_j\|_2^2\). In particular, this expression is easy to evaluate numerically.

Of course, our initial assumption that \(\mu\) is known is often unrealistic, hence we modify the estimator as follows. Given \(\theta > 0\), define
\[
Y_j = \frac{1}{2} (Z_{2j-1} - Z_{2j}) (Z_{2j-1} - Z_{2j})^T,
\]
\[
\widehat{\Sigma}_{2n}(\theta) = \frac{1}{n\theta} \sum_{j=1}^{n} \psi(\theta Y_j).
\]
Let \(\hat{\sigma}^2 \geq \frac{1}{2} \mathbb{E}((Z - \mu)(Z - \mu)^T)^2 + \text{tr}(\Sigma) + 2\Sigma^2\), and \(\hat{\theta} = \sqrt{\frac{t}{n}} \frac{1}{\hat{\sigma}}\). Our covariance estimator is defined as \(\widehat{\Sigma}_{2n} := \widehat{\Sigma}_{2n}(\hat{\theta})\).

**Remark 8.** Construction of \(\widehat{\Sigma}_{2n}(\theta)\) essentially halves the effective sample size. While the loss of a constant factor can be deemed insignificant in theoretical bounds, it is undesirable in applications. A more “statistically natural” version of the estimator based on a sample of size \(2n\) is the U-statistic
\[
\Sigma_{2n}(\theta) = \frac{1}{(2n)^2} \sum_{1 \leq i < j \leq 2n} \frac{1}{\theta} \psi\left(\frac{\theta}{2} (Z_i - Z_j) (Z_i - Z_j)^T\right).
\]
Analysis of the estimators of this type is not covered in the paper, and more work in this direction is required. Another possibility to avoid “halving” the sample size is to center the data using a robust estimator of location (such as the spatial median).

The following result can be deduced from Theorem 3.1:

**Corollary 4.1.** With probability \(\geq 1 - 2de^{-t}\),
\[
\mathbb{E}\|\widehat{\Sigma}_{2n} - \Sigma\| \leq 2\hat{\sigma} \sqrt{\frac{t}{n}}.
\]
Proof. Note that for all \( j = 1, \ldots, n \), \( EY_j = \Sigma \). Since \( Y_1, \ldots, Y_n \) are i.i.d. random matrices, Theorem 3.1 applies (see remark 3), giving that

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

where \( \hat{\sigma}^2 \geq \|EY_1^2\| \). It is easy to check that

\[
\|EY_1^2\| = \frac{1}{2} \left\| E \left( (Z - \mu)(Z - \mu)^T \right)^2 + \text{tr} (\Sigma) \Sigma + 2\Sigma^2 \right\|
\]

and result follows.

Next, we present an estimator which achieves strong deviation guarantees in the Frobenius norm (as well as in the operator norm). Let \( \hat{\Sigma}_{2n} \) be the sample covariance estimator based on \( Z_1, \ldots, Z_{2n} \):

\[
\hat{\Sigma}_{2n} = \frac{1}{(\frac{2n}{2})} \sum_{1 \leq i < j \leq 2n} \frac{(Z_i - Z_j)(Z_i - Z_j)^T}{2}.
\]

The following “soft thresholding” estimator has been studied in [Lou14] (here, \( \tau > 0 \) is a fixed “threshold parameter"):

\[
\hat{\Sigma}^\tau_{2n} = \arg\min_{A \in \mathbb{R}^{d \times d}} \left[ \|A - \hat{\Sigma}_{2n}\|_F^2 + \tau \|A\|_1 \right].
\] (4.1)

We propose to replace the sample covariance \( \hat{\Sigma}_{2n} \) by \( \hat{\Sigma}_{2n} \), and consider

\[
\hat{\Sigma}^\tau_{2n} = \arg\min_{A \in \mathbb{R}^{d \times d}} \left[ \|A - \hat{\Sigma}_{2n}\|_F^2 + \tau \|A\|_1 \right].
\] (4.2)

It is not hard to see (e.g., see the proof of Theorem 1 in [Lou14]) that \( \hat{\Sigma}^\tau_{2n} \) can be written explicitly as

\[
\hat{\Sigma}^\tau_{2n} = \sum_{j=1}^d \max \left( \lambda_j(\hat{\Sigma}_{2n}) - \tau/2, 0 \right) v_j(\hat{\Sigma}_{2n})v_j(\hat{\Sigma}_{2n})^T,
\]

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

**Theorem 4.1.** For any

\[
\tau \geq 4\sigma \sqrt{\frac{t + \log(2d)}{2n}}
\]

\[
\|\hat{\Sigma}^\tau_{2n} - \Sigma\|_F^2 \leq \inf_{A \in \mathbb{R}^{d \times d}} \left[ \|A - \Sigma\|_F^2 + \frac{(1 + \sqrt{2})^2}{8} \tau^2 \text{rank}(A) \right],
\] (4.3)

with probability \( \geq 1 - e^{-t} \).

Proof. The proof is based on the following lemma:
Lemma 4.1. Inequality (4.3) holds on the event $E = \{ \tau \geq 2 \left\| \hat{\Sigma}_{2n} - \Sigma \right\| \}$.

To verify this statement, it is enough to repeat the steps of the proof of Theorem 1 in [Lou14], replacing each occurrence of the sample covariance $\hat{\Sigma}_{2n}$ by the “robust” estimator $\hat{\Sigma}^\tau_{2n}$. Finally, it follows from corollary 4.1 that $\Pr(E) \geq 1 - e^{-t}$ whenever $\tau \geq 4\sigma \sqrt{t + \log(2d)}/\sqrt{2n}$.

4.2. 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

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

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 $\Pi := \text{Unif}(X)$ on $X$, and assume that the noisy linear measurement $Y$ has the form

$$Y = \text{tr} (X^TA_0) + \xi,$$

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

It is easy to check that $\mathbb{E}(YX) = \frac{1}{d_1d_2} A_0$, hence the natural unbiased estimator of $A_0$ is

$$\hat{A} = \frac{d_1d_2}{n} \sum_{j=1}^n Y_j X_j.$$

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

$$\hat{A}^\tau = \arg\min_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1d_2} ||A - \hat{A}||_2^2 + \tau ||A||_1 \right]$$

$$= \arg\min_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1d_2} ||A||_F^2 - \left\langle \frac{2}{n} \sum_{j=1}^n Y_j X_j, A \right\rangle + \tau ||A||_1 \right].$$

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

$$\hat{A}_s = \frac{d_1d_2}{n} \sum_{j=1}^n Y_j \mathcal{H}(X_j),$$

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

$$\hat{A}^\tau = \arg\min_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1d_2} ||\mathcal{H}(A) - \mathcal{H}(\hat{A}_s)||_F^2 + 2\tau ||A||_1 \right]$$

$$= \arg\min_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1d_2} ||\mathcal{H}(A)||_F^2 - \left\langle \frac{2}{n} \sum_{j=1}^n Y_j \mathcal{H}(X_j), \mathcal{H}(A) \right\rangle + 2\tau ||A||_1 \right].$$
However, strong theoretical guarantees for this estimator exist only when the “noise term” $\xi_j$ is either bounded with probability 1, or has sub-exponential tails. We propose to replace $\hat{A}_s$ with a “robust” estimator

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

where $\psi(\cdot)$ satisfies (3.1) and $\theta := \theta(t, n, A_0) = \frac{1}{\|A_0\|_{\max} \sqrt{\text{Var}(\xi)}} \sqrt{\frac{t + \log(2(d_1 + d_2))}{n(d_1 \wedge d_2)}}$ (the reasoning behind this choice of $\theta$ is given below). Consider

$$\hat{R}^\tau = \arg\min_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1 d_2} \|\mathcal{H}(A)\|_F^2 - \left( \frac{2}{d_1 d_2} \hat{R}, \mathcal{H}(A) \right) + 2\|A\|_1 \right].$$

Finally, set

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

The following result holds:

**Theorem 4.2.** Assume that $\xi_j$ is independent of $X_j$, $j = 1, \ldots, n$, and that $\text{Var}(\xi) < \infty$. For any

$$\tau \geq 4 \left( \|A_0\|_{\max} \vee \sqrt{\text{Var}(\xi)} \right) \sqrt{\frac{t + \log(2(d_1 + d_2))}{n(d_1 \wedge d_2)}},$$

$$\frac{1}{d_1 d_2} \left\| \hat{R}^\tau - A_0 \right\|_F^2 \leq \inf_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1 d_2} \|A - A_0\|_F^2 + \left( \frac{1 + \sqrt{2}}{2} \right)^2 d_1 d_2 \tau^2 \text{rank}(A) \right].$$

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

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

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

We begin with the following inequality:

**Lemma 4.2.** Assume that $\tau \geq 2\|M\|$. Then

$$\frac{1}{d_1 d_2} \left\| \mathcal{H}(\hat{R}^\tau) - \mathcal{H}(A_0) \right\|_F^2 \leq \inf_{B \in A} \left[ \frac{1}{d_1 d_2} \|B - \mathcal{H}(A_0)\|_F^2 + \left( \frac{1 + \sqrt{2}}{2} \right)^2 d_1 d_2 \tau^2 \text{rank}(B) \right].$$

**Proof.** By the definition of $\hat{R}^\tau$, we see that

$$\mathcal{H}(\hat{R}^\tau) = \arg\min_{B \in A} \left[ \frac{1}{d_1 d_2} \|B\|_F^2 - \left( \frac{2}{d_1 d_2} \hat{R}, B \right) + \tau\|B\|_1 \right].$$

If we replace $\frac{1}{d_1 d_2} \hat{R}$ by $\frac{1}{d_1 d_2} \hat{A}_s = \frac{1}{b} \sum_{j=1}^n Y_j \mathcal{H}(X_j)$, the result follows from Theorem 1 in [KLT11] immediately. To obtain the current statement, it is enough to repeat the argument of Theorem 1 in [KLT11], replacing each occurrence of the matrix $\frac{1}{m_1 m_2} \hat{A}_s$ by $\frac{1}{m_1 m_2} \hat{R}$. \hfill $\square$

---

3It is also possible to use truncation $\psi(x) = \text{sign}(x)(|x| \wedge 1)$, however, resulting bounds have slightly worse constants; see remark 2.
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

$$
\frac{1}{d_1 d_2} \| \mathcal{H}(\hat{R}^\tau) - \mathcal{H}(A_0) \|_F^2 = \frac{2}{d_1 d_2} \| \hat{R}^\tau - A_0 \|_F^2.
$$

(4.4)

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

$$
\inf_{B \in A} \left[ \frac{1}{d_1 d_2} \| B - \mathcal{H}(A_0) \|_F^2 + \left( \frac{1 + \sqrt{2}}{2} \right)^2 d_1 d_2 \tau^2 \text{rank}(B) \right]

= 2 \inf_{A \in \mathbb{R}^{d_1 \times d_2}} \left[ \frac{1}{d_1 d_2} \| A - A_0 \|_F^2 + \left( \frac{1 + \sqrt{2}}{2} \right)^2 d_1 d_2 \tau^2 \text{rank}(A) \right].
$$

(4.5)

It remains to estimate the probability of the event \( E = \{ \tau \geq 2 \| M \| \} \). Let

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

**Lemma 4.3.** Assume that \( \xi_j \) is independent of \( X_j \), \( j = 1, \ldots, n \). Then

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

**Proof.** Note that \( \mathbb{E} [ Y^2 X X^T ] = \mathbb{E} [ \xi^2 X X^T ] + \mathbb{E} \left[ (\text{tr}(X^T A_0))^2 X X^T \right] \). Moreover, \( |\text{tr}(X^T A_0)| \leq \max_{i,j} |(A_0)_{i,j}| = \| A_0 \|_{\text{max}} \), and \( \| \mathbb{E} X X^T = \frac{1}{m} \|_F \), hence

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

Similarly,

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

Applying Theorem 3.1 (see remark 3) with

$$
\theta = \sqrt{\frac{2(t + \log(2(d_1 + d_2)))}{n}} \frac{1}{\| A_0 \|_{\text{max}} \vee \sqrt{\text{Var}(\xi)}} \left( \frac{(\text{Var}(\xi) \vee \| A_0 \|_{\text{max}}^2) \frac{2}{d_1 \wedge d_2}}{1/2} \right)^{1/2}
$$

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

we see that

$$
\| M \| \leq 2 \left( \| A_0 \|_{\text{max}} \vee \sqrt{\text{Var}(\xi)} \right) \sqrt{\frac{t + \log(2(d_1 + d_2))}{n(d_1 \wedge d_2)}}
$$

with probability \( \geq 1 - e^{-t} \). Final result now follows from the combination of this inequality with (4.4), (4.5) and lemma 4.2.
5. Optimal choice of $\theta$ and adaptation to the unknown second moment.

To make results of Theorem 3.1 and its versions useful, one has to choose the “optimal” value for the parameter $\theta$ which in turn depends on the (usually unknown) norm $\sigma_n^2 = \left\lVert \sum_{j=1}^{n} E Y_j^2 \right\rVert$. We will present two ways to address this problem. In section 5.1, we develop a simple adaptive solution based on Lepski’s method, and section 5.2 discusses an approach based on using the “plug-in estimator” of $\sigma_n^2$. The latter approach requires stronger assumptions but yields observable error bounds that can be used to construct conference balls for the unknown mean. As a side product, we obtain a version of PAC-Bayesian bound (see lemma 5.1) which extends result of proposition 2.1 in [Cat14] from the scalar case to matrices.

5.1. Adaptation via Lepski’s method.

Lepski’s method [Lep92] is a powerful general technique that allows to adapt to the unknown structure of the problem - for example, bandwidth selection in nonparametric estimation, or unknown second moment in our case. Let $Y_1, \ldots, Y_n \in \mathbb{C}^{d \times d}$ be independent self-adjoint random matrices with $\sigma_n^2 = \left\lVert \sum_{j=1}^{n} E Y_j^2 \right\rVert$, and assume that $\sigma_{\min}, \sigma_{\max}$ are such that

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

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

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

be a set of cardinality $|\mathcal{J}| \leq 1 + \log_2(\sigma_{\max}/\sigma_{\min})$, 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

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

(5.1)

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

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 $t > 0$:

$$\Pr \left( \|T_{n}^* - E Y\| \geq 6(\sigma_n/\sqrt{n}) \sqrt{\frac{2t}{n}} \right) \leq 2d \log_2 \left( \frac{2 \sigma_{\max}}{\sigma_{\min}} \right) e^{-t}.$$
Proof. Let \( \bar{j} = \min \{ j \in \mathcal{J} : \sigma_j \geq \frac{\sigma_n}{\sqrt{n}} \} \) (hence \( \sigma_{\bar{j}} \leq 2 \frac{\sigma_n}{\sqrt{n}} \)). First, we will show that \( j^* \leq \bar{j} \) with high probability. Indeed,

\[
\Pr (j^* > \bar{j}) \leq \Pr \left( \bigcup_{k \in \mathcal{J}, k > \bar{j}} \left\{ \| T_{n,k} - T_{n,\bar{j}} \| > 2 \sigma_k \sqrt{\frac{2t}{n}} \right\} \right)
\leq \Pr \left( \| T_{n,\bar{j}} - \mathbb{E} Y \| > \sigma_{\bar{j}} \sqrt{\frac{2t}{n}} \right) + \sum_{k \in \mathcal{J}, k > \bar{j}} \Pr \left( \| T_{n,k} - \mathbb{E} Y \| > \sigma_k \sqrt{\frac{2t}{n}} \right)
\leq 2de^{-t} + 2d \log_2 \left( \frac{\sigma_{\max}}{\sigma_{\min}} \right) e^{-t},
\]

where we used Theorem 3.1 to bound each of the probabilities in the sum. The display above implies that the event \( \mathcal{B} = \bigcap_{k \in \mathcal{J}, k \geq \bar{j}} \left\{ \| T_{n,k} - \mathbb{E} Y \| \leq \sigma_k \sqrt{\frac{2t}{n}} \right\} \) of probability \( \geq 1 - 2d \log_2 \left( \frac{2\sigma_{\max}}{\sigma_{\min}} \right) e^{-t} \) is contained in \( \mathcal{E} = \{ j^* \leq \bar{j} \} \). Hence, on \( \mathcal{B} \) we have that

\[
\| T_{n}^* - \mathbb{E} Y \| \leq \| T_{n}^* - T_{n,\bar{j}} \| + \| T_{n,\bar{j}} - \mathbb{E} Y \| \leq 2 \sigma_{\bar{j}} \sqrt{\frac{2t}{n}} + \sigma_{\bar{j}} \sqrt{\frac{2t}{n}}
\leq 4 \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}} + 2 \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}} = 6 \frac{\sigma_n}{n} \sqrt{\frac{2t}{n}},
\]

and result follows. \( \Box \)

Remark 9. 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 \( \{ j \in \mathbb{Z} : \sigma_{\min} \leq \kappa j \sigma_{\min} < \kappa \sigma_{\max} \} \) for some \( 1 < \kappa < 2 \), at the cost of replacing \( \log_2 \left( \frac{2\sigma_{\max}}{\sigma_{\min}} \right) \) by \( \log_2 \left( \frac{\kappa \sigma_{\max}}{\sigma_{\min}} \right) / \log_2 \kappa \).

5.2. Estimation of the second moment and PAC-Bayesian bounds.

One of the shortcomings of the adaptive estimator constructed in the previous section is that it does not allow to quantify the estimation error: we adapt to the unknown \( \sigma \) (hence, the desired error rate) but can’t say how large \( \sigma \) is.

In this section, we will construct an adaptive estimator that admits computable error bounds. As it often happens, stronger results are possible at the cost of slightly stronger assumptions – in particular, we will assume that the fourth moments of the entries of a random matrix are bounded.

For simplicity and clarity of presentation, we will assume that the observations \( Y_1, \ldots, Y_n \) are i.i.d. copies of a self-adjoint random matrix \( Y \in \mathbb{C}^{d \times d} \). We will first construct an estimator \( \hat{\sigma}^2 \) of \( \sigma^2 = \| \mathbb{E} Y^2 \| \) such that

\[
1 - \varepsilon \leq \frac{\hat{\sigma}^2}{\sigma^2} \leq 1 + \varepsilon
\quad (5.2)
\]
with high probability for a small fixed $\varepsilon > 0$. We will then extend PAC-Bayesian theorems and corresponding uniform deviation bounds to the noncommutative framework, which will allow us to use the estimator $\hat{\sigma}$ in place of unknown $\sigma$.

Define, for some function $\psi(\cdot)$ satisfying (3.1),

\[
\hat{\sigma}^2(\tau) := \left\| \frac{1}{n\tau} \sum_{j=1}^{n} \psi(\tau Y_j^2) \right\|
\] (5.3)

**Assumption 1.** Suppose that

(a) $\|EY^4\| \leq \sqrt{n}$,

(b) $\varepsilon \|EY^2\| \geq \sqrt{\frac{t}{n}}$,

where $\varepsilon > 0$ is a small fixed constant.

It follows from Theorem 3.1 that

\[
\Pr \left( \left\| \frac{1}{n\tau} \sum_{j=1}^{n} \psi(\tau Y_j^2) - EY^2 \right\| \geq R \sqrt{\frac{t}{n}} \right) \leq 2d \exp \left( -\tau R \sqrt{\frac{t}{n}} + \frac{n\tau^2 R^2}{2} \right)
\] (5.4)

for any $R^2 \geq \|EY^4\|$ and $\tau > 0$. Under assumption 1, choosing $\bar{\tau} = \sqrt{\frac{2}{n} \frac{1}{n^{1/4}}}$ and using the bound (5.4) with $R^2 = \sqrt{n}$ yields

\[
\Pr \left( \left\| \frac{1}{n\bar{\tau}} \sum_{j=1}^{n} \psi(\bar{\tau} Y_j^2) - EY^2 \right\| \geq \varepsilon \|EY^2\| \right) \leq 2de^{-t},
\] (5.5)

hence $\hat{\sigma}^2 = \hat{\sigma}^2(\bar{\tau})$ clearly satisfies (5.2).

### 5.2.1. PAC-Bayesian bounds.

Given $\theta > 0$, define $X_j(\theta) := \log \left( I + \theta Y_j + \frac{\theta^2}{2} Y_j^2 \right)$ for $j = 1, \ldots, n$. Recall that the Kullback-Leibler divergence between two probability distributions is defined via

\[
K(\rho\|\nu) = \begin{cases} \int \log \frac{d\rho}{d\nu} d\rho, & \text{if } \rho \text{ is absolutely continuous with respect to } \nu, \\ +\infty, & \text{otherwise}, \end{cases}
\]

where $\frac{d\rho}{d\nu}$ is the Radon-Nikodym derivative of $\rho$ with respect to $\nu$.

**Lemma 5.1.** Let $\nu$ be the “reference measure” - an arbitrary but fixed probability measure on $\mathbb{R}$. Then for all $t > 0$,

\[
\left\| \int_{\mathbb{R}} \left[ \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) \right] d\rho(\theta) \right\| \leq t \int_{\mathbb{R}} \theta d\rho(\theta) + K(\rho\|\nu)
\]

with probability $\geq 1 - 2d \int_{\mathbb{R}} \exp \left( -t \theta + \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} \theta EY_j^2 \right\| \right) d\nu(\theta)$ uniformly over the set of all distributions $\rho$ with $K(\rho\|\nu) < \infty$. 

Proof. See section A.3.

Remark 10. Conclusion of the lemma holds for non-identically distributed $Y_j$'s as well.

We are ready to state and prove the main result of this section. Let $Y_1, \ldots, Y_n$ be i.i.d. copies of $Y$, $\sigma^2 = \|EY^2\|$, and let assumption 1 be satisfied with some $0 < \varepsilon \leq \frac{1}{8}$. Let $\hat{\sigma} := \hat{\sigma}(\bar{\tau})$ be the estimator of $\sigma$ defined in (5.3) with $\bar{\tau} = \frac{\sqrt{2t}}{n^2\hat{\sigma}}$. Finally, set $\hat{\theta} = \sqrt{\frac{t}{n}}\frac{\sqrt{1 - \varepsilon}}{\sqrt{n\bar{\tau}}}$.

**Theorem 5.2.** With probability $\geq 1 - (2d + 1)e^{-t/2} - 2de^{-3t/8}$,

$$\left\| \frac{1}{n\theta} \sum_{j=1}^{n} X_j(\hat{\theta}) - EY \right\| \leq \frac{\hat{\sigma}}{\sqrt{1 - \varepsilon}} \left( \sqrt{\frac{1}{4n}} + 2\sqrt{\frac{t}{n}} + \frac{\log \left( 1 + 5\sqrt{nd}/8 \right)}{\sqrt{nt}} \right)$$

and

$$\left\| \frac{1}{n\theta} \sum_{j=1}^{n} X_j(\hat{\theta}) - EY \right\| \leq \sigma \left( \sqrt{\frac{1}{4n}} + (2 + 2\varepsilon) \sqrt{\frac{t}{n}} + (1 + 2\varepsilon) \frac{\log \left( 1 + 5\sqrt{nd}/8 \right)}{\sqrt{nt}} \right).$$

Proof. Set $\theta_* := \sqrt{\frac{t}{n}}\frac{1}{\sigma}$, and let $\nu(\cdot)$ be the uniform distribution on $[\theta_* - (2\varepsilon + \tau_n)\theta_*, \theta_* + \tau_n\theta_*]$, where $\tau_n > 0$ will be defined below. Furthermore, let $\rho(\cdot)$ be the uniform distribution on the (random) interval $[\hat{\theta} - \tau_n\theta_*, \hat{\theta} + \tau_n\theta_*]$. It is easy to check that on the event $\mathcal{E} = \left\{ 1 - \varepsilon \leq \frac{\hat{\sigma}^2}{\sigma^2} \leq 1 + \varepsilon \right\}$ of probability $\geq 1 - 2de^{-t/2}$ (this follows from (5.5)), the support of $\rho$ is contained in the support of $\nu$, hence on this event $K(\rho||\nu) = \log \left( 1 + \frac{\varepsilon}{\tau_n} \right)$.

We will now compare, for each $1 \leq j \leq n$, $\int_{\mathbb{R}} \log \left( I + \theta Y_j + \frac{\theta^2}{2} Y_j^2 \right) d\rho(\theta)$ with $\log \left( I + \hat{\theta} Y_j + \frac{\hat{\theta}^2}{2} Y_j^2 \right)$.

Using Lemma A.1, we deduce that

$$\left\| \int_{\mathbb{R}} \log \left( I + \theta Y_j + \frac{\theta^2}{2} Y_j^2 \right) d\rho(\theta) - \log \left( I + \hat{\theta} Y_j + \frac{\hat{\theta}^2}{2} Y_j^2 \right) \right\|$$

$$= \max_{i=1,\ldots,d} \left\| \int_{\mathbb{R}} \log \left( 1 + \theta \lambda_i(Y_j) + \frac{\theta^2}{2} \lambda_i^2(Y_j) \right) - \log \left( 1 + \hat{\theta} \lambda_i(Y_j) + \frac{\hat{\theta}^2}{2} \lambda_i^2(Y_j) \right) \right\|$$

$$\leq 2 \max_{i=1,\ldots,d} \int_{\mathbb{R}} \left\| \log \left( 1 + \lambda_i(Y_j)(\theta - \hat{\theta}) \right) \right\| d\rho(\theta) \leq 2 \log (1 + \tau_n\theta_*\|Y_j\|).$$
Similarly,
\[
\left\| \frac{1}{n} \sum_{j=1}^{n} \int_{\mathbb{R}} \log \left( I + \theta Y_j + \frac{\theta^2}{2} Y_j^2 \right) d\rho(\theta) - \frac{1}{n} \sum_{j=1}^{n} \log \left( I + \hat{\theta} Y_j + \frac{\hat{\theta}^2}{2} Y_j^2 \right) \right\| \leq \frac{2}{n} \sum_{j=1}^{n} \log (1 + \tau_n \theta_s ||Y_j||).
\]

The latter expression can be estimated using the following simple fact (its proof is included in section A.4):

**Lemma 5.2.** Let \( Z_1, \ldots, Z_n \) be a sequence of independent random variables such that \( Z_j > -1 \) almost surely and \( \mathbb{E}|Z_j| < \infty \) for \( j = 1, \ldots, n \). Then for any \( s > 0 \)

\[
\text{Pr} \left( \frac{1}{n} \sum_{j=1}^{n} \log (1 + Z_j) \geq \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}Z_j + \frac{s}{n} \right) \leq e^{-s}.
\]

Combined with an obvious bound \( \mathbb{E}\|Y\| \leq \sqrt{\mathbb{E}\|Y\|_F^2} = \sqrt{\text{tr } \mathbb{E}Y^2} \leq \sigma \sqrt{d} \) (where \( \| \cdot \|_F \) stands for the Frobenius norm), we see that for any \( s > 0 \),

\[
\left\| \frac{1}{n} \sum_{j=1}^{n} \int_{\mathbb{R}} X_j(\theta) d\rho(\theta) - \frac{1}{n} \sum_{j=1}^{n} X_j(\hat{\theta}) \right\| \leq 2\sigma \sqrt{d} \tau_n \theta_s + \frac{2s}{n}
\]

with probability \( \geq 1 - e^{-s} \).

Finally, we will combine the bounds above with the inequality of Lemma 5.1. Note that, replacing \( t \) by \( \sigma \sqrt{tn} \) in the statement of Lemma 5.1, we get that

\[
\left\| \int_{\mathbb{R}} \frac{1}{n} \left[ \sum_{j=1}^{n} (X_j(\theta) - \theta \mathbb{E}X_j) \right] d\rho(\theta) \right\| \geq \sigma \sqrt{\frac{t}{n} \int_{\mathbb{R}} \theta d\rho(\theta) + \frac{K(\rho||\nu)}{n}}
\]

with probability not exceeding

\[
2d \int_{\mathbb{R}} \exp \left( -\theta \|\mathbb{E}X\|_F^2 \right) \left( \sqrt{\text{tr } \mathbb{E}X^2} + \frac{n \theta^2}{2\sigma^2} \right) d\nu(\theta) \leq 2d \exp \left( \sup_{\theta \in \text{supp}(\nu)} \left\{ -\theta \sigma \sqrt{tn} + \frac{n \theta^2}{2\sigma^2} \right\} \right)
\]

\[
= 2d \exp \left( -\frac{t}{2} (1 - (\tau_n + 2\varepsilon)^2) \right)
\]

since the maximum is attained for \( \theta = \theta_s - (2\varepsilon + \tau_n) \theta_s \).

Noting that \( \int_{\mathbb{R}} \theta d\rho(\theta) = \hat{\theta} \) and combining (5.7) with (5.8), we conclude that with probability \( \geq 1 - 2d \exp \left( -\frac{t}{2} (1 - (\tau_n + 2\varepsilon)^2) \right) - 2de^{-t/2} - e^{-s} \)

\[
\left\| \frac{1}{n} \sum_{j=1}^{n} X_j(\hat{\theta}) - \hat{\theta} \mathbb{E}Y \right\| \leq 2\sigma \sqrt{d} \tau_n \theta_s + \frac{2s}{n} + \hat{\theta} \sigma \sqrt{\frac{t}{n} + \frac{\log (1 + \varepsilon \tau_n)}{n}}.
\]
Finally, observe that on the event \( \mathcal{E} = \{ 1 - \varepsilon \leq \frac{\sigma^2}{\hat{\sigma}^2} \leq 1 + \varepsilon \} \), \( \frac{\hat{\sigma}}{\hat{\theta}} \leq \sqrt{\frac{1 + \varepsilon}{1 - \varepsilon}} \leq 1 + 2\varepsilon \) for \( \varepsilon < 1/3 \), hence setting \( s = t/2 \) we obtain the inequality

\[
\left\| \frac{1}{n\hat{\theta}} \sum_{j=1}^{n} \log \left( I + \hat{\theta} Y_j + \frac{\hat{\theta}^2}{2} Y_j^2 \right) - EY \right\| \leq 2(1 + 2\varepsilon)\sigma\sqrt{d\tau_n} + \sigma\sqrt{\frac{t}{n}} + \frac{\dot{\sigma}}{\sqrt{1 - \varepsilon}} \left( \sqrt{\frac{t}{n}} + \frac{\log \left( 1 + \frac{\varepsilon}{\tau_n} \right)}{\sqrt{nt^2}} \right)
\]

which holds with probability \( \geq 1 - 2d \exp \left( -\frac{1}{2} \left( 1 - (\tau_n + 2\varepsilon)^2 \right) \right) - 2d e^{-t/2} - e^{-t/2} \). To get the final result, we set \( \tau_n = \frac{1}{4(1+2\varepsilon)} \sqrt{n} \) and use the bound \( \sigma \leq \frac{\dot{\sigma}}{\sqrt{1 - \varepsilon}} \) to get the first inequality, and \( \hat{\sigma} \leq \sigma \sqrt{1 + \varepsilon} \) to get the second. \( \square \)

6. From bounds depending on \( \|EY^2\| \) to bounds depending on \( \|E(Y - EY)^2\| \).

Assume that \( Y_1, \ldots, Y_n \) are i.i.d. copies of \( Y \in \mathbb{C}^{d \times d} \). Results of the previous sections apply to the whole family of estimators given by

\[
\left\{ T_n(S) = S + \sum_{j=1}^{n} \frac{1}{n\hat{\theta}_S} \psi(\theta_S(Y_j - S)), \ S \in S \subseteq \mathbb{C}^{d \times d}, \ S = S^* \right\}
\]

for a suitable choice of \( \theta_S := \theta(S) \) and a function \( \psi(\cdot) \) satisfying (3.1). In particular, bound (3.7) implies that for any self-adjoint \( S \) and \( \theta_S = \sqrt{\frac{2}{n} \|E(Y - S)^2\|^{1/2}} \),

\[
\Pr \left( \|T_n(S) - EY\| \geq \|E(Y - S)^2\|^{1/2} \sqrt{\frac{s}{n}} \right) \leq 2d \exp (-s/2).
\]

As suggested by the inequality above, to get an estimator with best deviation guarantees, one should pick \( S \) to minimize \( \mathbb{E}(Y - S)^2 \) yielding \( S_{\text{optimal}} = EY \) which is of course unknown. Instead, we will use a known but random \( S \) which is “not far” from \( EY \).

The idea is to define an iterative procedure starting with a “preliminary” estimator \( \hat{T}_0 \) of \( EY \) and then refine it by setting

\[
\hat{T}_1 := T_n(\hat{T}_0) = \hat{T}_0 + \frac{1}{n\theta_1} \sum_{j=1}^{n} \log \left( I + \theta_1(Y_j - \hat{T}_0) + \frac{\theta_1^2}{2} (Y_j - \hat{T}_0)^2 \right)
\]

for a suitable \( \theta_1 > 0 \). The process can be repeated if necessary.

 Everywhere in this section, we will assume that one has access to some known (possibly very crude) bounds for \( \sigma^2 = \|EY^2\| \) and \( \sigma_0^2 = \|E(Y - EY)^2\| \): \n
**Assumption 2.** Let \( \sigma_{\text{min}}, \sigma_{0,\text{min}} \) and \( \sigma_{\text{max}}, \sigma_{0,\text{max}} \) be known constants such that

\[
\sigma_{\text{min}} \leq \sigma \leq \sigma_{\text{max}} \text{ and } \sigma_{0,\text{min}} \leq \sigma_0 \leq \sigma_{0,\text{max}}.
\]
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 $Y_1, \ldots, Y_n$ into two disjoint subsets $G_1$ and $G_2$ of cardinality $\geq \lfloor n/2 \rfloor$ each. 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 5.1) 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 $G_2 - \hat{T}_0 := \{ Y_j - \hat{T}_0 : 1 \leq j \leq n, \ Y_j \in G_2 \}$.

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

$$J_1 = \{ j \in \mathbb{Z} : \sigma_{\min} \leq \sigma_j < 2 \sigma_{\max} \}$$

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

$$J_2 = \left\{ j \in \mathbb{Z} : \sigma_{0,\min} \leq \sigma_{0,j} < 2 \left( \sigma_{0,\max} + 12 \sigma_{\max} \sqrt{\frac{t}{n}} \right) \right\},$$

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

$$\hat{T}_0 = T_{|G_1|,j^*_1}(0),$$

where

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

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

$$j^*_1 := \min \left\{ j \in J_1 : \forall k \in J_1 \text{ s.t. } k > j, \ \| T_{|G_1|,k}(0; G_1) - T_{|G_1|,j}(0; G_1) \| \leq 2 \sigma_k \sqrt{\frac{2t}{|G_1|}} \right\}$$

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

$$\hat{T}_1 = \hat{T}_0 + T_{|G_2|,j^*_2}(\hat{T}_0; G_2),$$

where

$$T_{|G_2|,j}(S; G_2) = \frac{1}{|G_2|} \sum_{i=|G_1|+1}^n \psi(\theta_{0,j}(Y_i - S)),$$

$$\theta_{0,j} = \sqrt{\frac{2t}{n}} \frac{1}{\sigma_{0,j}}, \text{ and }$$

$$j^*_2 := \min \left\{ j \in J_2 : \forall k \in J_2 \text{ s.t. } k > j, \ \| T_{|G_2|,k}(\hat{T}_0; G_2) - T_{|G_2|,j}(\hat{T}_0; G_2) \| \leq 2 \sigma_{0,k} \sqrt{\frac{2t}{|G_2|}} \right\}.$$
Proof. Let $E_1$ be the event defined by

$$E_1 = \left\{ \| \hat{T}_0 - EY \| \leq 6\sigma \sqrt{\frac{2t}{n/2}} \right\}.$$ 

By Theorem 5.1,

$$\Pr(E_1) \geq 1 - 2d \log_2 \left( \frac{2\sigma_{\max}}{\sigma_{\min}} \right) e^{-t}. \quad (6.1)$$

Note that on this event,

$$\| E \left[ (Y - \hat{T}_0)^2 | \hat{T}_0 \right] \| = \| E(Y - EY)^2 + (EY - \hat{T}_0)^2 \| \leq \sigma_0^2 + 12\sigma \sqrt{\frac{t}{n}}.$$ 

(6.2)

In particular, on event $E_1$,

$$\| E \left[ (Y - \hat{T}_0)^2 | \hat{T}_0 \right] \|^{1/2} \leq \sigma_0 + 12\sigma \sqrt{\frac{t}{n}} \leq \sigma_{0,\max} + 12\sigma_{\max} \sqrt{\frac{t}{n}}. \quad (6.3)$$

Next,

$$\Pr \left( \| \hat{T}_1 - EY \| \geq 6 \left( \sigma_0 + 12\sigma \sqrt{\frac{t}{n}} \right) \sqrt{\frac{2t}{n/2}} \right) \quad (6.4)$$

$$= \Pr \left( \| T_{G_2,J_2}(\hat{T}_0; G_2) - (EY - \hat{T}_0) \| \geq 6 \left( \sigma_0 + 12\sigma \sqrt{\frac{t}{n}} \right) \sqrt{\frac{2t}{n/2}} \right)$$

$$\leq \Pr(E_1^c) + \Pr \left( \| T_{G_2,J_2}(\hat{T}_0; G_2) - (EY - \hat{T}_0) \| \geq 6 \left( \sigma_0 + 12\sigma \sqrt{\frac{t}{n}} \right) \sqrt{\frac{2t}{n/2}} | E_1 \right).$$

Define the new probability measure by $\tilde{\Pr}(A) = \Pr(A | E_1)$. Clearly, under this new measure, subsample $G_2$ is still independent of $G_1$ since $E_1 \in \sigma(G_1)$ - the sigma-algebra generated by $G_1$, and for any $B \in \sigma(G_2)$, $\tilde{\Pr}(B) = \Pr(B)$. Let $\hat{\Pr}$ be the expectation with respect to measure $\tilde{\Pr}(\cdot)$. We have

$$\Pr \left( \| T_{G_2,J_2}(\hat{T}_0; G_2) - (EY - \hat{T}_0) \| \geq 6 \left( \sigma_0 + 12\sigma \sqrt{\frac{t}{n}} \right) \sqrt{\frac{2t}{n/2}} | E_1 \right)$$

$$= \tilde{\Pr} \left( \| T_{G_2,J_2}(\hat{T}_0; G_2) - (EY - \hat{T}_0) \| \geq 6 \left( \sigma_0 + 12\sigma \sqrt{\frac{t}{n}} \right) \sqrt{\frac{2t}{n/2}} \right)$$

$$\leq \tilde{\Pr} \left( \| T_{G_2,J_2}(\hat{T}_0; G_2) - (EY - \hat{T}_0) \| \geq 6 \| E \left[ (Y - \hat{T}_0)^2 | \hat{T}_0 \right] \|^{1/2} \sqrt{\frac{2t}{n/2}} \right)$$

$$= \hat{\Pr} \left( \| T_{G_2,J_2}(\hat{T}_0; G_2) - (EY - \hat{T}_0) \| \geq 6 \| E \left[ (Y - \hat{T}_0)^2 | \hat{T}_0 \right] \|^{1/2} \sqrt{\frac{2t}{n/2}} \hat{T}_0 \right)$$

$$\leq 1 - 2d \log_2 \left( \frac{2(\sigma_{0,\max} + 12\sigma_{\max} \sqrt{t/n})}{\sigma_{0,\min}} \right) e^{-t}. \quad (6.5)$$
Here, we use the definition of $\tilde{\Pr}()$ on the first step and (6.3) on the second step. The last inequality follows from independence of $G_2$ from $\hat{T}_0$ (under $\tilde{\Pr}()$) and Theorem 5.1 applied conditionally on $\hat{T}_0$; indeed, this can be done since (6.3) holds on $E_1$. It remains to combine the last bound with (6.4) and (6.1).

6.2. Iterative estimation and equation $\sum_{j=1}^{n} \psi(\theta(Y_j - S)) = 0$.

We will next show how to design an estimator with deviations controlled by “correct” variance term without sample splitting (however, subject to the assumption that the sample size is sufficiently large). In what follows, let $\psi_0()$ be any function that satisfies (3.1) and is Lipschitz with Lipschitz constant bounded by 1. For example, we may take $\psi_0(x) = \log(1 + x + x^2/2)$. As before, let $t > 0$ be fixed, set $\sigma_0,j = 2^{j} \sigma_{0,\min}$, $\sigma_{0,\min} \leq \sigma_0,j < 2 \sigma_0,\max$, $\theta = \sqrt{\frac{2t}{n} \sigma_{\max}}$, and $\theta_j = \sqrt{\frac{2t}{n} \sigma_{0,j}}$ for $j \in J$.

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

$$
\delta_j^{(k)} = \frac{12}{5} \sigma_0,j \sqrt{\frac{2t}{n} + 6^{-k}} \left( \sigma_{\max} \sqrt{\frac{2t}{n}} - \frac{12}{5} \sigma_{0,j} \sqrt{\frac{2t}{n}} \right)
$$

for $k \geq 1$. Next, for each $j \in J$, we define

$$
T_{n,j}^{(0)} := T_n^{(0)} = \frac{1}{n \theta} \sum_{i=1}^{n} \psi_0(\theta Y_i),
$$

(independent of $j$)\(^4\), and

$$
T_{n,j}^{(k)} := T_n \left(T_{n,j}^{(k-1)}\right) = T_{n,j}^{(k-1)} + \frac{1}{n \theta_j} \sum_{i=1}^{n} \psi_0 \left(\theta_j \left(Y_i - T_{n,j}^{(k-1)}\right)\right)
$$

for $k \geq 1$. Finally, we apply Lepski’s method to the collection of estimators $\{T_{n,j}^{(k)} : j \in J\}$. To this end, define $\hat{T}_k := T_{n,j_k}^{(k)}$, where

$$
j_k = \min \left\{ j \in J : \forall l \in J \text{ s.t. } l > j, \|T_{n,l}^{(k)} - T_{n,j}^{(k)}\| \leq 2 \delta_l^{(k)} \right\}.
$$

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 “sufficiently small” with high probability.
Theorem 6.2. Let
\[ \tau = 1.1K \sqrt{\frac{d^3 + dt}{n}} + \sqrt{\frac{2t}{n} \frac{1}{2\sigma_0}}, \]
where \( K > 0 \) is an absolute constant, and assume that \( \tau \leq 1/6 \). Moreover, assume that
\[ \left( \frac{24}{5} \sigma_{0,\text{max}} \lor \sigma_{\text{max}} \right) \sqrt{\frac{2t}{n}} \leq 1. \tag{6.7} \]
Then for all \( k \geq 0 \) simultaneously,
\[ \| \hat{T}_k - \mathbb{E}Y \| \leq 3 \left[ (1 - 6^{-k}) \frac{24}{5} \sigma_0 \sqrt{\frac{2t}{n}} + 6^{-k} \sigma_{\text{max}} \sqrt{\frac{2t}{n}} \right] \]
with probability \( \geq 1 - 8d \left( 1 + 2 \log_2 \left( \frac{12\sigma_{\text{max}}}{\sigma_{0,\text{min}}} \right) \right) \log_2 \left( \frac{2\sigma_{0,\text{max}}}{\sigma_{0,\text{min}}} \right) e^{-t} \).

Before presenting the proof, let us discuss a corollary of the preceding result. Let \( A \) be the event of probability \( \Pr(A) \geq 1 - 8d \left( 1 + 2 \log_2 \left( \frac{12\sigma_{\text{max}}}{\sigma_{0,\text{min}}} \right) \right) \log_2 \left( \frac{2\sigma_{0,\text{max}}}{\sigma_{0,\text{min}}} \right) e^{-t} \) defined in Theorem 6.2.

On this event, for each \( j \in J \) s.t. \( \sigma_j \geq \sigma_0 \), the sequence \( \{ T_{n,j}^{(k)} \}_{k \geq 0} \) is uniformly bounded (it follows from the details of the proof presented below) - in particular, for each suitable \( j \in J \), we can select a converging subsequence (different for each \( \omega \in A \)) \( T_{n,j}^{(k_m)} \xrightarrow{m \to \infty} \hat{T}_{n,j} \). Then \( \hat{T}_{n,j} \) must satisfy the equation
\[ \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi_0 \left( \theta_j (Y_i - \hat{T}_{n,j}) \right) = 0. \]
In particular, the reasoning above implies that on event \( A \), equation
\[ \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi_0 \left( \theta_j (Y_i - S) \right) = 0 \tag{6.8} \]
has a solution in the set \( \mathcal{K} = \left\{ S : \| S - \mathbb{E}Y \| \leq (\sigma_{\text{max}} \lor \frac{24}{5} \sigma_{0,\text{max}}) \sqrt{\frac{2t}{n}} \right\} \). Next, if we take any solution \( \hat{S}_j \) of (6.8) that belongs to \( \mathcal{K} \) as a “preliminary" estimator \( T_{n,j}^{(0)} = \hat{S}_j \), then the sequence \( \{ T_{n,j}^{(k)} \}_{k \geq 0} \) is a constant sequence (meaning that \( T_{n,j}^{(k)} = \hat{S}_j \) for all \( k \)), and it follows from Theorem 6.2 that
\[ \| \hat{S}_j - \mathbb{E}Y \| \leq \inf_{k \geq 0} \delta_{j}^{(k)} \leq \frac{12}{5} \sigma_{0,j} \sqrt{\frac{2t}{n}} \]
with high probability. One can further apply Lepski’s method to the collection \( \{ \hat{S}_j \}_{j \in J} \) to obtain an estimator \( \tilde{S} \) that satisfies
\[ \| \tilde{S} - \mathbb{E}Y \| \leq \frac{72}{5} \sigma_0 \sqrt{\frac{2t}{n}} \]
with high probability. Details that can be easily recovered from the argument presented below.

\(^5\)We implicitly assume that \( \hat{S}_j \) is measurable, but detailed discussion of measurability issues is out of scope of this paper.
Proof of Theorem 6.2. We will first state several technical results that are required in the proof. Let \( j \in J \) be such that \( \sigma_{0,j} = 2^j \sigma_{0,\min} \geq \sigma_0 \), and define
\[
X_{j,i}(S) := \psi_0 \left( \theta_j (Y_i - S) \right), \quad i = 1, \ldots, n.
\]
Moreover, set
\[
L_n(\delta, j) := \sup_{S: \|S - \bar{E}Y\| \leq \delta} \left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(S) - \bar{E}X_{j,i}(S)) - \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(\bar{E}Y) - \bar{E}X_{j,i}(\bar{E}Y)) \right\|,
\]
and define the event
\[
\Omega(\delta, j) = \left\{ L_n(\delta, j) \leq K \delta \sqrt{\frac{d^3 + dt}{n}} \right\},
\]
where \( K > 0 \) is an absolute constant.

Lemma 6.1. For \( K \) large enough,
\[
\Pr(\Omega(\delta, j)) \geq 1 - 2de^{-t}.
\]
Proof. See section A.5 in the appendix. \( \square \)

Lemma 6.2. For any Hermitian \( S \),
\[
\left\| S - \bar{E}Y + \frac{1}{\theta_j} \bar{E}X_{j,1}(S) \right\| \leq \frac{\theta_j}{2} \left\| \bar{E}(Y - S)^2 \right\|.
\]
Proof. Note that
\[
X_{j,1}(S) = \psi_0 \left( \theta_j (Y_1 - S) \right) \leq \log \left( I + \theta_j (Y_1 - S) + \frac{\theta_j^2}{2} (Y_1 - S)^2 \right)
\]
\[
\leq \theta_j (Y_1 - S) + \frac{\theta_j^2}{2} (Y_1 - S)^2,
\]
which is a consequence of scalar inequality \( \log(1 + x) \leq x \), \( x > -1 \), and fact 2.1, hence we can deduce from fact 2.2 that
\[
\lambda_{\max} \left( S - \bar{E}Y + \frac{1}{\theta_j} \bar{E}X_{j,1}(S) \right) \leq \frac{\theta_j}{2} \left\| \bar{E}(Y - S)^2 \right\|.
\]
At the same time, \( -\lambda_{\min} \left( S - \bar{E}Y + \frac{1}{\theta_j} \bar{E}X_{j,1}(S) \right) = \lambda_{\max} \left( -\frac{1}{\theta_j} \bar{E}X_{j,1}(S) - (S - \bar{E}Y) \right) \). Since
\[
-X_{j,1}(S) = -\psi_0 \left( \theta_j (Y_1 - S) \right) \leq \log \left( I - \theta_j (Y_1 - S) + \frac{\theta_j^2}{2} (Y_1 - S)^2 \right)
\]
by (3.1), we conclude that
\[
-\lambda_{\min} \left( S - \bar{E}Y - \frac{1}{\theta_j} \bar{E}X_{j,1}(S) \right) \leq \frac{\theta_j}{2} \left\| \bar{E}(Y - S)^2 \right\|
\]
by (3.1), we conclude that
\[
\left\| S - \bar{E}Y + \frac{1}{\theta_j} \bar{E}X_{j,1}(S) \right\| \leq \frac{\theta_j}{2} \left\| \bar{E}(Y - S)^2 \right\|.
\]

Lemma 6.1. For \( K \) large enough,
\[
\Pr(\Omega(\delta, j)) \geq 1 - 2de^{-t}.
\]
Proof. See section A.5 in the appendix. \( \square \)
Lemma 6.3. With probability $\geq 1 - 2de^{-t}$,

$$\left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(EY) - \mathbb{E}[X_{j,i}(EY)]) \right\| \leq \sigma_{0,j} \sqrt{\frac{2t}{n}} + \frac{\theta_j}{2} \sigma_0^2.$$  

Proof. Result follows from Theorem 3.1 and the inequality

$$\left\| \frac{1}{\theta_j} \mathbb{E}X_{j,1}(EY) \right\| \leq \frac{\theta_j}{2} \sigma_0^2,$$

which is a consequence of lemma 6.2. Indeed,

$$\left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(EY) - \mathbb{E}[X_{j,i}(EY)]) \right\| \leq \left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} X_{j,i}(EY) \right\| + \frac{1}{\theta_j} \left\| \mathbb{E}X_{j,1}(EY) \right\|$$

$$\leq \sigma_{0,j} \sqrt{\frac{2t}{n}} + \frac{\theta_j}{2} \sigma_0^2$$

with probability $\geq 1 - 2de^{-t}$. We are ready to proceed with the proof of the theorem. Let $\mathcal{E}_0 = \left\{ \left\| T_{n}^{(0)} - EY \right\| \leq \sigma_{\text{max}} \sqrt{\frac{2t}{n}} \right\}$ (where $T_{n}^{(0)}$ was defined in (6.6)), and note that $\Pr(\mathcal{E}_0) \geq 1 - 2de^{-t}$ by Theorem 3.1. Let

$$k_{\text{max}} = 1 + \max \left\{ k \geq 0 : \sigma_{0,\text{max}} 1.1^k \leq \frac{12}{5} \sigma_{\text{max}} \right\}, \quad (6.9)$$

$$\gamma_l = 1.1^l \sigma_{0,\text{min}} \sqrt{\frac{2t}{n}}, \quad l \geq 0,$$

and note that $k_{\text{max}} \leq 1 + \left[ \frac{\log_2 \left( \frac{12\sigma_{\text{max}}}{5\sigma_{0,\text{min}}} \right) }{\log_2 1.1} \right] \leq 1 + 8 \log_2 \left( \frac{12\sigma_{\text{max}}}{5\sigma_{0,\text{min}}} \right).$ Define

$$\Omega_j := \left\{ \left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{i,EY} - \mathbb{E}X_{i,EY}) \right\| \leq \sigma_{0,j} \sqrt{\frac{2t}{n}} + \frac{\theta_j}{2} \sigma_0^2 \right\} \cap \bigcap_{l=0}^{k_{\text{max}}} \Omega (\gamma_l, j).$$

By lemma 6.1, lemma 6.3 and the union bound, $\Pr(\Omega_j) \geq 1 - 2d(2 + k_{\text{max}})e^{-t}.$ We will now show by induction that on the event $\mathcal{E}_0 \cap \Omega_j$, $\left\| T_{n,j}^{(k)} - EY \right\| \leq \delta_j^{(k)}$ for all $k \geq 0$. For $k = 0$, result follows from the definition of $\mathcal{E}_0$. In remains to complete the induction step $k - 1 \mapsto k$. Note that when $\left\{ \left\| T_{n,j}^{(k-1)} - EY \right\| \leq \delta_j^{(k-1)} \right\}$ occurs, we have

$$\left\| T_{n,j}^{(k)} - EY \right\| = \left\| T_{n,j}^{(k-1)} - EY + \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi_0 (\theta_j (Y_i - T_{n,j}^{(k-1)})) \right\|$$

$$\leq \sup_{S : \|S - EY\| \leq \delta_j^{(k-1)}} \left\| S - EY + \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi_0 (\theta_j (Y_i - S)) \right\|. \quad (6.10)$$
Expression under the supremum in (6.10) can be decomposed as follows:

\[
S - \mathbb{E}Y + \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi_0(\theta_j(Y_i - S)) = S - \mathbb{E}Y + \frac{1}{\theta_j} \mathbb{E}X_{j,1}(S) + \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(S) - \mathbb{E}X_{j,i}(S)) \\
- \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(\mathbb{E}Y) - \mathbb{E}X_{j,i}(\mathbb{E}Y)) + \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(\mathbb{E}Y) - \mathbb{E}X_{j,i}(\mathbb{E}Y)).
\]

We will treat 3 terms separately: first, it follows from lemma 6.2 that on \(\Omega_j\)

\[
\sup_{S: \|S - \mathbb{E}Y\| \leq \delta_j^{(k-1)}} \left\| S - \mathbb{E}Y + \frac{1}{\theta_j} \mathbb{E}X_{1,S} \right\| \leq \frac{\theta_j}{2} \left( \sigma_0^2 + (\delta_j^{(k-1)})^2 \right). 
\tag{6.11}
\]

Next,

\[
\left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{i,\mathbb{E}Y} - \mathbb{E}X_{i,\mathbb{E}Y}) \right\| \leq \sigma_{0,j} \sqrt{\frac{2t}{n}} + \frac{\theta_j}{2} \sigma_0^2, 
\tag{6.12}
\]

once again by the definition of \(\Omega_j\). Let \(\tilde{t} = \min \left\{ l \geq 0 : \gamma_{l} \geq \delta_j^{(k-1)} \right\} \) (where \(\gamma_{l}\) was defined in (6.9)), and note that \(\tilde{t} \leq k_{\max}\) and \(\gamma_{\tilde{t}} \leq 1.1 \delta_j^{(k-1)}\). We bound the third term as

\[
\sup_{S: \|S - \mathbb{E}Y\| \leq \delta_j^{(k-1)}} \left\| \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(S) - \mathbb{E}X_{j,i}(S)) - \frac{1}{n\theta_j} \sum_{i=1}^{n} (X_{j,i}(\mathbb{E}Y) - \mathbb{E}X_{j,i}(\mathbb{E}Y)) \right\| 
= L_n \left( \delta_j^{(k-1)}, j \right) \leq L_n \left( \gamma_{\tilde{t}}, j \right) \leq K \gamma_{\tilde{t}} \sqrt{\frac{d^3 + dt}{n}} \leq 1.1 K \delta_j^{(k-1)} \sqrt{\frac{d^3 + dt}{n}}.
\tag{6.13}
\]

Putting the bounds (6.11), (6.12), (6.13) together, we can estimate the supremum in (6.10) as

\[
\sup_{S: \|S - \mathbb{E}Y\| \leq \delta_j^{(k-1)}} \left\| S - \mathbb{E}Y + \frac{1}{n\theta_j} \sum_{i=1}^{n} \psi_0(\theta_j(Y_i - S)) \right\| 
\leq \frac{\theta_j}{2} \left( \sigma_0^2 + (\delta_j^{(k-1)})^2 \right) + \frac{\theta_j}{2} \sigma_0^2 + \sigma_{0,j} \sqrt{\frac{2t}{n}} + \delta_j^{(k-1)} \cdot 1.1 K \sqrt{\frac{d^3 + dt}{n}} 
\leq (\sigma_0 + \sigma_{0,j}) \sqrt{\frac{2t}{n}} + \delta_j^{(k-1)} \left( 1.1 K \sqrt{\frac{d^3 + dt}{n}} + \sqrt{\frac{2t}{n}} \frac{1}{2\sigma_0} \right). 
\tag{6.14}
\]

Note that we have used bounds \(\theta_j \sigma_0^2 \leq \sigma_0 \sqrt{\frac{2t}{n}}\) and \(\theta_j \left( \delta_j^{(k-1)} \right)^2 \leq \theta_j \delta_j^{(k-1)}\) (indeed, (6.7) implies that \(\delta_j^{(m)} \leq 1\) for all \(j\) and \(m\)) to get the second inequality above. Since \(j\) was chosen such that \(\sigma_{0,j} \geq \sigma_0\) and \(\tau = 1.1 K \sqrt{\frac{d^3 + dt}{n}} + \sqrt{\frac{2t}{n}} \frac{1}{2\sigma_0} \leq \frac{1}{6}\) by assumption, we have shown that

\[
\left\| T^{(k)}_{n,j} - \mathbb{E}Y \right\| \leq 2\sigma_{0,j} \sqrt{\frac{2t}{n}} + \frac{1}{6} \delta_j^{(k-1)} = \delta_j^{(k)},
\]
where the last equality follows from the fact that the sequence \( \delta_j^{(k)} \) defined in (6.5) satisfies the recursive relation
\[
\delta_j^{(0)} = \sigma_{\max} \sqrt{\frac{2t}{n}}, \quad \delta_j^{(k)} = 2\sigma_{0,j} \sqrt{\frac{2t}{n}} + \frac{1}{6} \delta_j^{(k-1)}.
\]

The union bound implies that To complete the proof, it is enough to follow the steps of the proof Theorem 5.1 applied to the collection of estimators \( \{ T_{n,j}^{(k)} : j \in J \} \); first, let \( \bar{j} = \min \{ j \in J : \sigma_{0,j} \geq \sigma_0 \} \), and note that the event
\[
\mathcal{E}_0 \cap \bigcap_{j \geq \bar{j}, \ j \in J} \Omega_j
\]
has probability \( \geq 1 - 8d \left( 1 + 2 \log_2 \left( \frac{12\sigma_{\max}}{5\sigma_{0,\min}} \right) \right) \log_2 \left( \frac{2\sigma_{0,\max}}{\sigma_{0,\min}} \right) e^{-t} \). Moreover, on this event \( j_k^* \leq \bar{j} \), hence
\[
\left\| \hat{T}_k - \mathbb{E}Y \right\| = \left\| T_{n,j_k^*}^{(k)} - \mathbb{E}Y \right\| \leq \left\| T_{n,j_k^*}^{(k)} - T_{n,j}^{(k)} \right\| + \left\| T_{n,j}^{(k)} - \mathbb{E}Y \right\|
\leq 3\delta_j^{(k)} \leq 3 \left[ (1 - 6^{-k}) \frac{24}{5} \sigma_0 \sqrt{\frac{2t}{n}} + 6^{-k} \sigma_{\max} \sqrt{\frac{2t}{n}} \right],
\]
where we used the fact that \( \sigma_{0,j} \leq 2\sigma_0 \) in the last inequality.

\( \square \)

**Acknowledgements.** I want to thank Larry Goldstein for carefully reading the preliminary version of the manuscript and making numerous helpful comments and suggestions.

**References**

[AMS96] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[BJD93] R. W. Butler, P. L. Davies, and M. Jhun. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, pages 1385–1400, 1993.

[Bha97] R. Bhatia. Matrix analysis. *Springer*, 1997.

[BJL15] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[Car10] E. Carlen. Trace inequalities and quantum entropy: an introductory course. Available at [http://www.mathphys.org/AZschool/material/AZ09-carlen.pdf](http://www.mathphys.org/AZschool/material/AZ09-carlen.pdf), 2010.

[Cat12] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185, 2012.

[Cat14] O. Catoni. PAC-Bayes learning bounds. *Lecture notes*, 2014. Available at [http://www.math.ens.fr/~catoni/homepage/catoni_ifcam.pdf](http://www.math.ens.fr/~catoni/homepage/catoni_ifcam.pdf).

[Cat16] O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.
[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Journal of the ACM (JACM), 58(3):11, 2011.

[CRZ16] T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. Electron. J. Statist., 10(1):1–59, 2016.

[CZZ10] T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics, 38(4):2118–2144, 2010.

[Dav92] L. Davies. The asymptotics of Rousseeuw’s minimum volume ellipsoid estimator. The Annals of Statistics, pages 1828–1843, 1992.

[Dir13] S. Dirksen. Tail bounds via generic chaining. arXiv preprint arXiv:1309.3522, 2013.

[DLLLO15] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. arXiv preprint arXiv:1509.05845, 2015.

[FWZ16] J. Fan, W. Wang, and Z. Zhu. Robust low-rank matrix recovery. arXiv preprint arXiv:1603.08315, 2016.

[GIU15] I. Giulini. PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. arXiv preprint arXiv:1511.06263, 2015.

[HR09] P. J. Huber and E. M. Ronchetti. Robust statistics. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009.

[HRVA08] M. Hubert, P. J. Rousseeuw, and S. Van Aelst. High-breakdown robust multivariate methods. Statistical Science, pages 92–119, 2008.

[HS16] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. Journal of Machine Learning Research, 17(18):1–40, 2016.

[Hub64] P. J. Huber. Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1):73–101, 1964.

[KLT11] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. arXiv preprint arXiv:1405.2468, 2014.

[KLT14] O. Klopp, K. Lounici, and A. B. Tsybakov. Robust matrix completion. arXiv preprint arXiv:1412.8132, 2014.

[Lep92] O. Lepski. Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. Theory of Probability & Its Applications, 36(4):682–697, 1992.

[Lie73] E. H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. Advances in Math., 11:267–288, 1973.

[LO11] Matthieu Lerasle and Roberto I Oliveira. Robust empirical mean estimators. arXiv preprint arXiv:1112.3914, 2011.

[Lou14] K. Lounici. High-dimensional covariance matrix estimation with missing observations. Bernoulli, 20(3):1029–1058, 2014.

[Mar76] R. A. Maronna. Robust M-estimators of multivariate location and scatter. Ann. Statist., 4(1):51–67, 01 1976.

[Min15] S. Minsker. Geometric median and robust estimation in Banach spaces. Bernoulli, 21(4):2308–2335, 2015.
Appendix A: Technical results.

Lemma A.1. For any $x, y \in \mathbb{R}$,

$$|\log(1 + x + x^2/2) - \log(1 + y + y^2/2)| \leq 2 \log(1 + |x - y|).$$

Proof. Without loss of generality, assume that $x > y$. Then

$$\log(1 + x + x^2/2) - \log(1 + y + y^2/2) = \log \left( \frac{1 + x + x^2/2}{1 + y + y^2/2} \right) = \log \left( 1 + \frac{(x - y)^2}{2(1 + y + y^2/2)} + (x - y) \frac{1 + y}{1 + y + y^2/2} \right).$$

It is easy to check that for any $y$, $2(1 + y + y^2/2) \geq 1$ and $\left| \frac{1 + y}{1 + y + y^2/2} \right| \leq 1$, so that

$$1 + \frac{(x - y)^2}{2(1 + y + y^2/2)} + (x - y) \frac{1 + y}{1 + y + y^2/2} \leq (1 + |x - y|)^2,$$

and the claim follows. \hfill \Box

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

$$-\log(1 + y + c_\alpha |y|^\alpha) \leq \log(1 - y + c_\alpha |y|^\alpha) \quad \text{for all } y \in \mathbb{R}.$$
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 $y_m = -\left(\frac{1}{ac_\alpha}\right)^{1/(a-1)}$. It is easy to check that $f(y_m) = 1 - y_m + \frac{y_m}{a}$, which implies that $f(y_m) > 0 \iff c_\alpha > \frac{a-1}{a^2}$ which always holds since $c_\alpha \geq \frac{a-1}{a}$ and $a > 1$.

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

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

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

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 $\left\| \begin{pmatrix} S & A \\ A^* & T \end{pmatrix}^2 \right\| \geq \|S^2 + AA^*\| \geq \|AA^*\|$ and $\left\| \begin{pmatrix} S & A \\ A^* & T \end{pmatrix}^2 \right\| \geq \|T^2 + A^*A\| \geq \|A^*A\|$. Since $\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}^2 = \begin{pmatrix} AA^* & 0 \\ 0 & A^*A \end{pmatrix}$, 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. \hfill \Box

A.2. Proof of Theorem 3.2.

Define $\phi(x) = e^x - 1$ and $X_j = \psi(\theta Y_j)$. Proceeding as in the proof of Theorem 3.1, we get

$$\Pr\left( \lambda_{\max}\left( \frac{1}{\theta} \sum_{j=1}^{n} (X_j - \theta EY_j) \right) \geq s \right) = \Pr\left( \phi\left( \lambda_{\max}\left( \sum_{j=1}^{n} (X_j - \theta EY_j) \right) \right) \geq \phi(s) \right) \leq \frac{1}{\phi(\theta s)} \E \phi\left( \sum_{j=1}^{n} (X_j - \theta EY_j) \right) = \frac{1}{\phi(\theta s)} \left( \E \exp\left( \sum_{j=1}^{n} (X_j - \theta EY_j) \right) - I \right).$$

It follows from Lemma 3.1 that

$$\E \exp\left( \sum_{j=1}^{n} (X_j - \theta EY_j) \right) \leq \exp\left( \frac{\theta^2}{2} \sum_{j=1}^{n} EY_j^2 \right).$$
Set $B_n^2 := \sum_{j=1}^n \mathbb{E} Y_j^2 \geq 0$, and note that

$$
\text{tr} \left[ \exp \left( \frac{\theta^2}{2} \sum_{j=1}^n \mathbb{E} Y_j^2 \right) - I \right] = \text{tr} \left[ \frac{\theta^2}{2} \sqrt{B_n^2} \left( I + \frac{\theta^2 B_n^2}{2!} + \cdots + \left( \frac{\theta^2 B_n^2}{k!} \right)^{k-1} \right) \right]
$$

$$
\leq \text{tr} \left[ \frac{\theta^2 B_n^2}{2} \left( 1 + \frac{\theta^2 \|B_n\|}{2} + \cdots + \left( \frac{\theta^2 \|B_n\|}{k!} \right)^{k-1} \right) \right] = \text{tr} B_n^2 \left( \exp \left( \frac{\theta^2 \|B_n\|}{2} \right) - 1 \right).
$$

Here we have used the fact that $A \preceq B$ implies $SAS^* \preceq SBS^*$ for $S = S^* := \sqrt{B_n^2}$, and the equality $\frac{e^{x-1}}{x} = \sum_{j=1}^\infty \frac{x^{j-1}}{j!}$. We have shown that

$$
\text{Pr} \left( \lambda_{\max} \left( \frac{1}{\theta} \sum_{j=1}^n (X_j - \theta \mathbb{E} Y_j) \right) \geq s \right) \leq \frac{\text{tr} B_n^2 \exp \left( \frac{\theta^2 \|B_n\|}{2} \right) - 1}{e^{\theta s} - 1}
$$

$$
\leq \frac{\text{tr} B_n^2}{\|B_n\|} \exp \left( \frac{\theta^2 \|B_n\|}{2} - \theta s \right) \frac{e^{\theta s}}{e^{\theta s} - 1} \leq \frac{\text{tr} B_n^2}{\|B_n\|} \exp \left( \frac{\theta^2 \|B_n\|}{2} - \theta s \right) \left( 1 + \frac{1}{\theta s} \right),
$$

where we used an elementary inequality $\frac{e^{\theta s}}{e^{\theta s} - 1} \leq 1 + \frac{1}{\theta s}$ on the last step.

Combining the same steps with fact 2.4 and the equality $-\lambda_{\min} (A) = \lambda_{\max} (-A)$, we get

$$
\text{Pr} \left( \lambda_{\min} \left( \frac{1}{\theta} \sum_{j=1}^n (X_j - \theta \mathbb{E} Y_j) \right) \leq -s \right) \leq \frac{\text{tr} B_n^2}{\|B_n\|} \exp \left( \frac{\theta^2 \|B_n\|}{2} - \theta s \right) \left( 1 + \frac{1}{\theta s} \right).
$$

Finally, replace $s$ by $t\sqrt{n}$ to get the bound in the required form.

### A.3. Proof of Lemma 5.1.

Note that

$$
\text{Pr} \left( \lambda_{\max} \left\{ \int_{\mathbb{R}} \left[ \sum_{j=1}^n (X_j(\theta) - \theta \mathbb{E} Y_j) \right] d\rho(\theta) \right\} > t \int_{\mathbb{R}} \theta d\rho(\theta) + K(\rho||\nu) \right)
$$

$$
= \text{Pr} \left( \lambda_{\max} \left\{ \int_{\mathbb{R}} \left[ \sum_{j=1}^n (X_j(\theta) - \theta \mathbb{E} Y_j) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta) I_d \right] d\rho(\theta) \right\} > 0 \right)
$$

$$
\leq \mathbb{E} \text{tr} \exp \left( \int_{\mathbb{R}} \left[ \sum_{j=1}^n (X_j(\theta) - \theta \mathbb{E} Y_j) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta) I_d \right] d\rho(\theta) \right).
$$
Note that \( Z \mapsto \text{tr} \exp(Z) \) (where \( Z \in \mathbb{R}^{d \times d} \) is Hermitian matrix) is a convex function by fact 2.6. Hence, Jensen’s inequality implies that for any \( \rho \) with \( K(\rho || \nu) < \infty \)

\[
\text{tr} \exp \left( \int_{\mathbb{R}} \left[ \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta)I_d \right] d\rho(\theta) \right) \\
\leq \int_{\mathbb{R}} \text{tr} \exp \left( \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta)I_d \right) d\rho(\theta) \\
= \int_{\mathbb{R}} e^{-\theta t} \text{tr} \exp \left( \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) \right) \exp \left( - \log \frac{d\rho}{d\nu}(\theta) \right) d\rho(\theta) \\
= \int_{\mathbb{R}} e^{-\theta t} \text{tr} \exp \left( \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) \right) \left( \frac{d\rho}{d\nu}(\theta) \right)^{-1} I \{ \frac{d\rho}{d\nu}(\theta) > 0 \} d\rho(\theta) \\
\leq \int_{\mathbb{R}} e^{-\theta t} \text{tr} \exp \left( \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) \right) d\nu(\theta).
\]

We can bound the expectation \( \mathbb{E} \text{tr} \exp \left( \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) \right) \) as in the proof of Lemma 3.1 to get

\[
\mathbb{E} \text{tr} \exp \left( \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) \right) \leq \text{tr} \exp \left( \frac{\theta^2}{2} \sum_{j=1}^{n} EY_j^2 \right) \\
\leq d \exp \left( \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} EY_j^2 \right\| \right),
\]

hence

\[
\mathbb{E} \text{tr} \exp \left( \int_{\mathbb{R}} \left[ \sum_{j=1}^{n} (X_j(\theta) - \theta EY_j) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta)I_d \right] d\rho(\theta) \right) \\
\leq d \int_{\mathbb{R}} \exp \left( -t\theta + \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} EY_j^2 \right\| \right) d\nu(\theta).
\]
Similarly,

\[
\Pr \left( \lambda_{\min} \left\{ \int_{\mathbb{R}} \sum_{j=1}^{n} (X_j(\theta) - \theta E Y_j) \, d\rho(\theta) \right\} < -t \int_{\mathbb{R}} \theta d\rho(\theta) - K(\rho \| \nu) \right) = \Pr \left( \lambda_{\max} \left\{ \int_{\mathbb{R}} \sum_{j=1}^{n} (\theta E Y_j - X_j(\theta)) \, d\rho(\theta) \right\} > t \int_{\mathbb{R}} \theta d\rho(\theta) + K(\rho \| \nu) \right)
\]

\[
\leq \mathbb{E} \text{tr} \exp \left( \int_{\mathbb{R}} \sum_{j=1}^{n} (\theta E Y_j - X_j(\theta)) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta) I_d \, d\rho(\theta) \right).
\]

Repeating the steps above, combined with the second bound of Lemma 3.1, we obtain

\[
\mathbb{E} \text{tr} \exp \left( \int_{\mathbb{R}} \sum_{j=1}^{n} (\theta E Y_j - X_j(\theta)) - t\theta I_d - \log \frac{d\rho}{d\nu}(\theta) I_d \, d\rho(\theta) \right) \leq d \int_{\mathbb{R}} \exp \left( -t\theta + \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} E Y_j^2 \right\| \right) d\nu(\theta).
\]

Combining inequalities (A.1) and (A.2), we deduce that for all probability measure \( \rho \) with \( K(\rho \| \nu) < \infty \) simultaneously,

\[
\left\| \int_{\mathbb{R}} \sum_{j=1}^{n} (X_j(\theta) - \theta E Y_j) \, d\rho(\theta) \right\| \leq t \int_{\mathbb{R}} \theta d\rho(\theta) + K(\rho \| \nu)
\]

with probability \( \geq 1 - 2d \int_{\mathbb{R}} \exp \left( -t\theta + \frac{\theta^2}{2} \left\| \sum_{j=1}^{n} E Y_j^2 \right\| \right) d\nu(\theta). \)

A.4. Proof of Lemma 5.2.

It is enough to note that

\[
\Pr \left( \frac{1}{n} \sum_{j=1}^{n} \log (1 + Z_j) \geq \frac{1}{n} \sum_{j=1}^{n} E Z_j + \frac{s}{n} \right) = \Pr \left( \exp \left( \sum_{j=1}^{n} \log (1 + Z_j) \right) \geq \exp \left( \sum_{j=1}^{n} E Z_j + s \right) \right)
\]

\[
\leq \mathbb{E} \left[ \prod_{j=1}^{n} (1 + Z_j) \right] e^{-\sum_{j=1}^{n} E Z_j - s} \leq e^{-s},
\]

where we used the inequality \( 1 + \mathbb{E} Z_j \leq e^{\mathbb{E} Z_j} \) for \( j = 1, \ldots, n \) on the last step.

A.5. Proof of Lemma 6.1.

To this end, we will use a chaining argument. Recall that the function \( \psi_0(\cdot) \) is Lipschitz with Lipschitz constant \( L = 1 \) by assumption. Recall that \( X_{j,i}(S) := \psi_0 (\theta_j (Y_i - S)) \), \( i = 1, \ldots, n \). It
follows from Lemma 2.2 that for any Hermitian $S_1, S_2$ and $1 \leq i \leq n$,

$$\|X_i, S_1 - X_i, S_2\| = \|\psi_0(\theta_j(Y_i - S_1)) - \psi_0(\theta_j(Y_i - S_2))\| \leq \|\psi_0(\theta_j(Y_i - S_1)) - \psi_0(\theta_j(Y_i - S_2))\|_F$$

$$\leq \theta_j \|S_1 - S_2\|_F \leq \sqrt{d} \theta_j \|S_1 - S_2\|.$$

Matrix Hoeffding’s inequality (lemma 2.3) applies with

$$Z_i = \frac{1}{n \theta_j} ((X_i, S_1 - \mathbb{E}X_i, S_1) - (X_i, S_2 - \mathbb{E}X_i, S_2)), \quad i = 1, \ldots, n,$$

and $M_i = \frac{2\sqrt{2}}{n} \|S_1 - S_2\|$, and yields that

$$\left\| \sum_{i=1}^{n} Z_i \right\| \leq \sqrt{32d} \|S_1 - S_2\| \sqrt{\frac{n}{d}}$$

with probability $\geq 1 - 2de^{-s}$.

**Lemma A.3** (Covering number in the operator norm). Let $B(r)$ be the ball of radius $r > 0$ in $\mathbb{R}^d$ with respect to the operator norm $\| \cdot \|$, centered at 0. Then the covering number $N(B(r), \varepsilon) := N(B(r), \| \cdot \|, \varepsilon)$ satisfies

$$N(B(r), \varepsilon) \leq \left( \frac{2r}{\varepsilon} + 1 \right)^d.$$

**Proof.** It is well known [Ver07] that

$$N(A, \varepsilon) \leq \frac{|A + B(\varepsilon/2)|}{|B(\varepsilon/2)|},$$

where $|C|$ denotes the Lebesgue measure of a set $C$, and $A + C$ stands for the Minkowski sum of the sets $A$ and $C$. For $A = B(r)$, we get $N(B(r), \varepsilon) \leq \frac{|B(r+\varepsilon/2)|}{|B(\varepsilon/2)|}$. The volume of the unit ball is given by

$$|B(r)| = c_d \int_{[-r,r]^d} \prod_{1 \leq i < j \leq d} \left| x_i^2 - x_j^2 \right| dx_1 \ldots dx_d,$$

where $c_d = d! 4^{-d} \left( \prod_{j=1}^{d} v_j^2 \right)^2$ and $v_j$ is the volume of the Euclidean unit ball in $\mathbb{R}^j$. From here, it is easy to see that

$$\frac{|B(r+\varepsilon/2)|}{|B(\varepsilon/2)|} = \left( \frac{2r}{\varepsilon} + 1 \right)^d.$$

Let $T(\delta_{k-1}) := \{ S \in \mathbb{C}^{d \times d} : \| S - \mathbb{E}Y \| \leq \delta_{k-1} \}$, and define the metric

$$\rho_d(S_1, S_2) := \sqrt{d} \| S_1 - S_2 \|, \quad S_1, S_2 \in \mathbb{C}^{d \times d}.$$

Viewing $S \mapsto \frac{1}{n \theta_j} \sum_{i=1}^{n} (X_i, S - \mathbb{E}X_i, S)$ as a $\mathbb{C}^{d \times d}$-valued stochastic process indexed by the elements of the metric space $(T(\delta_{k-1}), \rho_d)$, we can apply Lemma 2.4 which implies that there exists an absolute constant $C > 0$ such that for any $t \geq 1$,

$$L_n(\delta_{k-1}) \leq C \frac{1}{\sqrt{n}} \left( \gamma_2(T(\delta_{k-1}), \rho_d) + \sqrt{tD(T(\delta_{k-1}), \rho_d)} \right)$$

(A.3)
with probability $\geq 1 - 2de^{-t}$. Recall the Dudley’s entropy integral bound (2.2):

$$
\gamma_2(T(\delta_{k-1}), \rho_d) \leq \frac{1}{2\sqrt{2} - 1} \int_0^1 \sqrt{\log N(T(\delta_{k-1}), \rho_d, \varepsilon/4)} d\varepsilon.
$$

Noting that $D(T(\delta_{k-1}), \rho_d) = 2\delta_{k-1}\sqrt{d}$ and combining Dudley’s bound with the estimate of Lemma A.3, we get

$$
\gamma_2(T(\delta_{k-1}), \rho_d) \leq C_1\delta_{k-1}d^{3/2},
$$

where $C_1 = \frac{2}{2\sqrt{2} - 1} \int_0^1 \log^{1/2}(1 + 4/\varepsilon) d\varepsilon$. Bound (A.3) implies that with probability $\geq 1 - 2de^{-t}$,

$$
L_n(\delta_{k-1}) \leq \frac{C}{\sqrt{n}} \left( C_1\delta_{k-1}d^{3/2} + 2\delta_{k-1}\sqrt{d}^3 \right) \leq \delta_{k-1} \cdot K \sqrt{\frac{d^3 + dt}{n}}
$$

(A.4)

for some absolute constant $K > 0$.

**B. Sub-Gaussian bounds that hold simultaneously for multiple confidence levels.**

Tail guarantees of theorem 3.1 in the “sub-Gaussian” form do not hold for multiple confidence levels simultaneously. Here, we show how to improve on this aspect. Let $0 < t_{\text{min}} < t_{\text{max}}$ be fixed, and define $t_j = 2^jt_{\text{min}}$.

$$
\mathcal{J} = \{ j \in \mathbb{Z} : t_{\text{min}} < t_j < 2t_{\text{max}} \}.
$$

Assume that $Y_1, \ldots, Y_n \in \mathbb{C}^{d \times d}$ is a sequence of independent self-adjoint random matrices, and let $\sigma_n^2 \geq \left\| \sum_{j=1}^n E Y_j^2 \right\|$. For $j \in \mathcal{J}$, set $\theta_j = \sqrt{\frac{2t_j}{n} \frac{1}{\sigma_n/\sqrt{n}}}$, and

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

Define

$$
\textstyle j_* := \min \left\{ j \in \mathcal{J} : \forall k \in \mathcal{J} \text{ s.t. } k > j, \left\| T_{n,k} - T_{n,j} \right\| \leq 2 \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t_k}{n}} \right\},
$$

and $T^*_{n,j} := T_{n,j_*}$.

**Theorem B.1.** For any $0 < t_{\text{min}} \leq t \leq t_{\text{max}},$

$$
\Pr \left( \left\| T^*_{n} - EY \right\| \geq 6 \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{t}{n}} \right) \leq 2d e^{-t}.
$$

**Proof.** Let $t_{\text{min}} \leq t \leq t_{\text{max}}$, and define $j(t) = \min \{ j \in \mathcal{J} : t_j \geq t \}$; note that $t_j(t) \leq 2t$. Next, we have that

$$
\{ j_* > j(t) \} = \bigcup_{k \in \mathcal{J}} \left\{ \left\| T_{n,k} - T_{n,j(t)} \right\| > 2 \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t_k}{n}} \right\}
$$

$$
\subseteq \bigcup_{k \in \mathcal{J}} \left\{ \left\| T_{n,k} - EY \right\| > \sigma_n \sqrt{\frac{2t_k}{n}} \right\} \bigcup \left\{ \left\| T_{n,j(t)} - EY \right\| > \sigma_n \sqrt{\frac{2t_j(t)}{n}} \right\} := \mathcal{E}(t)
$$
By union bound, probability of the latter event $\mathcal{E}(t)$ is bounded by
\[
2de^{-t_j(t)} + \sum_{k \in J : k > j(t)} 2de^{-t_k} \leq 2d \sum_{t \geq 0} e^{-t_j(t)}t^2 \leq 2d \frac{e^{-t}}{1 - e^{-t}}.
\]

Next, on the event $\mathcal{E}(t) := (\mathcal{E}(t))^c \subseteq \{j_* \leq j(t)\}$, we have
\[
\|T_{n,j_*} - EY\| \leq \|T_{n,j_*} - T_{n,j(t)}\| + \|T_{n,j(t)} - EY\| \leq 2\frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t_j(t)}{n}} + \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t_j(t)}{n}}
\]
\[
\leq 2\sqrt{2} \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}} + \sqrt{2} \frac{\sigma_n}{\sqrt{n}} \sqrt{\frac{2t}{n}},
\]
and conclusion follows.

\[\square\]

C. Numerical simulation results.

Numerical simulation was performed for covariance estimation problem. Data was simulated as follows: let $U = (U^{(1)}, \ldots, U^{(100)})^T \in \mathbb{R}^{100}$ be a vector with i.i.d. coordinates such that $U^{(j)} \overset{d}{=} \frac{1}{\sqrt{2\xi_j} \text{d}(\xi_{j,1} - \xi_{j,2})}$, where $\xi_{j,1}$ and $\xi_{j,2}$, $j = 1, \ldots, 100$, are independent random variables with probability density function
\[
p_{\xi}(t; q) = \frac{q}{(1 + t)^{1+q}}I\{t \geq 0\}
\]
(which belongs to the Pareto family), $c(q) = \text{Var}(\xi) = \frac{q}{(q - 1)^2(q - 2)}$ and $q = 4.01$; in particular, $\text{Var}(U^{(j)}) = 1$. Finally, let $Z = \sqrt{\Sigma}U$, where $\Sigma$ is a diagonal matrix with $\Sigma_{11} = 10$, $\Sigma_{22} = 5$, $\Sigma_{33} = 1$, and $\Sigma_{jj} = \frac{1}{97}$, $j \geq 4$. In particular, $EZ = 0$ and $EZZ^T = \Sigma$.

The goal of numerical experiment was to evaluate the quality of estimation of the covariance matrix $\Sigma$ 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 $Z_1, \ldots, Z_n$, i.i.d. copies of $Z$, and centered the data via the spatial (or geometric) median defined as
\[
\widehat{M}_n = \arg\min_{y \in \mathbb{R}^{100}} \sum_{j=1}^{100} \|y - Z_j\|_2.
\]
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 5.1. We provide details for completeness: set $\sigma_{\text{max}} := 2\sqrt{\frac{1}{n} \sum_{j=1}^{n} \|Z_j^0\|_2^2 Z_j^0 Z_j^0 T}$, $\sigma_{\text{min}} = \frac{\sigma_{\text{max}}}{100}$, $J = \{ j \in \mathbb{Z} : \sigma_{\text{min}} < 1.3^j \leq \sigma_{\text{max}} \}$.
and let $\psi(\cdot)$ be the function defined in (3.3). Let $t = \log 10$, and for $j \in J$, set $\theta_j = \sqrt{\frac{2t}{n} \frac{1}{1.3^j}}$ and

$$
\hat{\Sigma}_{n,j} = \frac{1}{n \theta_j} \sum_{i=1}^{n} \psi\left(\theta_j z_i^0 z_i^0^T\right).
$$

Finally, define

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

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

Quality of covariance estimation was evaluated via comparing $\|\hat{S}_n - \Sigma\| / \|\Sigma\|$ with $\|\hat{\Sigma}_n - \Sigma\| / \|\Sigma\|$ over 500 runs of simulations. We also compared errors of estimation of projectors onto the first principal component, $\|u_1(\hat{S}_n)u_1(\hat{S}_n)^T - u_1(\Sigma)u_1(\Sigma)^T\|$ and $\|u_1(\hat{\Sigma}_n)u_1(\hat{\Sigma}_n)^T - u_1(\Sigma)u_1(\Sigma)^T\|$, where $u_1(\cdot)$ denotes the eigenvector corresponding to the largest eigenvalue of a matrix. Histograms illustrating performance of both estimators are presented in figures 1a and 1b (for the sample size $n = 100$), and in figures 2a and 2b (for the sample size $n = 1000$). It is clear from the graphs that in all scenarios, $\hat{\Sigma}_n$ performs significantly better than $\hat{S}_n$. 
Fig 1: Sample size $n = 100$, dimension $d = 100$. 

(a) Covariance matrix estimation error

(b) First principal component estimation error
Fig 2: Sample size $n = 1000$, dimension $d = 100$. 

(a) Covariance matrix estimation error

(b) First principal component estimation error