Bernoulli **23**(1), 2017, 110–133
DOI: 10.3150/15-BEJ730

Concentration inequalities and moment bounds for sample covariance operators

VLADIMIR KOLTCHINSKII* and KARIM LOUNICI**

School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA.
E-mail: *vlad@math.gatech.edu; **klounici@math.gatech.edu

Let $X, X_1, \dots, X_n, \dots$ be i.i.d. centered Gaussian random variables in a separable Banach space E with covariance operator Σ:

$$\Sigma : E^* \mapsto E, \qquad \Sigma u = \mathbb{E}\langle X, u\rangle X, \qquad u \in E^*.$$

The sample covariance operator $\hat{\Sigma} : E^* \mapsto E$ is defined as

$$\hat{\Sigma} u := n^{-1} \sum_{j=1}^{n} \langle X_j, u\rangle X_j, \qquad u \in E^*.$$

The goal of the paper is to obtain concentration inequalities and expectation bounds for the operator norm $\|\hat{\Sigma} - \Sigma\|$ of the deviation of the sample covariance operator from the true covariance operator. In particular, it is shown that

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n} \right),$$

where

$$\mathbf{r}(\Sigma) := \frac{(\mathbb{E}\|X\|)^2}{\|\Sigma\|}.$$

Moreover, it is proved that, under the assumption that $\mathbf{r}(\Sigma) \le n$, for all $t \ge 1$, with probability at least $1 - e^{-t}$

$$\big| \|\hat{\Sigma} - \Sigma\| - M \big| \lesssim \|\Sigma\| \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right),$$

where M is either the median, or the expectation of $\|\hat{\Sigma} - \Sigma\|$. On the other hand, under the assumption that $\mathbf{r}(\Sigma) \ge n$, for all $t \ge 1$, with probability at least $1 - e^{-t}$

$$\big| \|\hat{\Sigma} - \Sigma\| - M \big| \lesssim \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right).$$

Keywords: concentration inequalities; effective rank; moment bounds; sample covariance

1. Introduction

Let $(E, \|\cdot\|)$ be a separable Banach space with the dual space E^*. For $x \in E, u \in E^*$, $\langle x, u\rangle$ denotes the value of linear functional u at vector x. Let X be a centered random variable in E with $\mathbb{E}|\langle X, u\rangle|^2 < +\infty, u \in E^*$ (i.e., X is weakly square integrable). Let

$$\Sigma u := \mathbb{E}\langle X, u\rangle X, \qquad u \in E^*,$$

so that $\langle \Sigma u, v\rangle = \mathbb{E}\langle X, u\rangle\langle X, v\rangle$. It is well known that this defines a bounded symmetric nonnegatively definite operator $\Sigma : E^* \mapsto E$ that is called *the covariance operator* of random variable X. Moreover, if $\mathbb{E}\|X\|^2 < +\infty$ (so, X is strongly square integrable), then it is also well known that the covariance operator Σ is nuclear.

Let $X_1, \dots, X_n$ be i.i.d. copies of X. The sample (empirical) covariance operator based on the observations $(X_1, \dots, X_n)$ is defined as the operator $\hat{\Sigma} : E^* \mapsto E$ such that

$$\hat{\Sigma} u := n^{-1} \sum_{j=1}^{n} \langle X_j, u\rangle X_j, \qquad u \in E^*.$$

Clearly, this is an operator of rank at most n and it is an unbiased estimator of the covariance operator Σ.

In this paper, we are interested in the case when X is a centered Gaussian random vector in E with covariance Σ. This implies that $\mathbb{E}\|X\|^2 < +\infty$ (in fact, $\|X\|$ is even a random variable with a finite ψ_2-norm; see [8], Chapter 3) and, as a consequence, the covariance operator Σ is nuclear. For operators $A : E^* \mapsto E$, $\|A\|$ will denote the operator norm:

$$\|A\| := \sup_{u \in E^*, \|u\| \le 1} \|Au\| = \sup_{u, v \in E^*, \|u\| \le 1, \|v\| \le 1} \big|\langle Au, v\rangle\big|.$$

Several other definitions and notations will be used throughout the paper. In particular, the relationship $B_1 \lesssim B_2$ (for nonnegative B_1, B_2) means that there exists an absolute constant $c \in (0, \infty)$ such that $B_1 \le c B_2$. Similarly, $B_1 \gtrsim B_2$ means that $B_1 \ge c B_2$ for an absolute constant c. If both $B_1 \lesssim B_2$ and $B_1 \gtrsim B_2$, we write $B_1 \asymp B_2$. Sometimes, symbols $\lesssim, \gtrsim, \asymp$ are provided with subscripts indicating possible dependence of constant c on other constants (say, $B_1 \lesssim_a B_2$ would mean that $B_1 \le c B_2$ with c that might depend on a).

We will also use occasionally Orlicz norms (such as ψ_1- and ψ_2-norms) in the spaces of random variables. Given a convex nondecreasing function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $\psi(0) = 0$ and a random variable η on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, define its ψ-norm as

$$\|\eta\|_\psi := \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{C}\right) \le 1\right\}.$$

For $\psi(u) := u^p, u > 0,\ p \ge 1$, the ψ-norm coincides with the $L_p(\mathbb{P})$-norm. Consider also $\psi_2(u) := e^{u^2} - 1, u \ge 0$ and $\psi_1(u) = e^u - 1, u \ge 0$. Then $\|\eta\|_{\psi_2} < +\infty$ means that η has sub-Gaussian tails and $\|\eta\|_{\psi_1} < +\infty$ means that η has subexponential tails. Some well-known inequalities for ψ_1 random variables will be used in what follows. For instance, for arbitrary random variables $\xi_k, k = 1, \dots, N, N \geq 2$ with $\|\xi_k\|_{\psi_1} < +\infty$,

$$\mathbb{E} \max_{1\leq k\leq N} |\xi_k| \lesssim \max_{1\leq k\leq N} \|\xi_k\|_{\psi_1} \log N.$$

If $\xi, \xi_1, \dots, \xi_n$ are i.i.d. centered random variables with $\|\xi\|_{\psi_1} < +\infty$, then the sum $\xi_1 + \cdots + \xi_n$ satisfies the following version of Bernstein's inequality: for all $t \geq 0$ with probability at least $1 - e^{-t}$

$$\left| \frac{\xi_1 + \cdots + \xi_n}{n} \right| \lesssim \|\xi\|_{\psi_1} \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right).$$

Our goal is to obtain moment bounds and concentration inequalities for the operator norm $\|\hat{\Sigma} - \Sigma\|$. It turns out that both the size of the expectation of random variable $\|\hat{\Sigma} - \Sigma\|$ and its concentration around its mean can be characterized in terms of the operator norm $\|\Sigma\|$ and another parameter defined below.

Definition 1. *Assuming that X is a centered Gaussian random variable in E with covariance operator Σ, define*

$$\mathbf{r}(\Sigma) := \frac{(\mathbb{E}\|X\|)^2}{\|\Sigma\|}.$$

Note that, for a Gaussian vector X, $\mathbb{E}^{1/2}\|X\|^2 \asymp \mathbb{E}\|X\|$ implying that

$$\mathbf{r}(\Sigma) \leq \frac{\mathbb{E}\|X\|^2}{\|\Sigma\|} =: \tilde{\mathbf{r}}(\Sigma) \lesssim \mathbf{r}(\Sigma).$$

In the case when E is a Hilbert space, $\mathbb{E}\|X\|^2 = \mathrm{tr}(\Sigma)$ and

$$\tilde{\mathbf{r}}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}.$$

The last quantity has been already used in the literature under the name of "effective rank" (see [16]). Clearly, $\tilde{\mathbf{r}}(\Sigma) \leq \mathrm{rank}(\Sigma)$.

The main results of the paper include the following:

- under an assumption that $X, X_1, \dots, X_n$ are i.i.d. centered Gaussian random variables in E with covariance operator Σ, it will be shown that

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n} \right). \tag{1.1}$$

- Moreover, under an additional assumption that $\mathbf{r}(\Sigma) \lesssim n$, the following concentration inequality holds for some constant $C > 0$ and for all $t \geq 1$ with probability at least $1 - e^{-t}$:

$$\Big| \|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\| \Big| \leq C\|\Sigma\| \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right). \tag{1.2}$$

Under an assumption that $\mathbf{r}(\Sigma) \gtrsim n$, the concentration inequality becomes

$$\big| \|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\| \big| \leq C\|\Sigma\| \bigg(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \sqrt{\frac{t}{n}} \vee \frac{t}{n} \bigg) \tag{1.3}$$

and it holds with the same probability.

2. Main results

The problem of bounding the operator norm $\|\hat{\Sigma} - \Sigma\|$ has been intensively studied, especially, in the finite-dimensional case (see [16] and references therein). The focus has been on understanding of dependence of this norm on the dimension of the space and on the sample size n (that could be simultaneously large) as well as on the tails of linear forms $\langle X, u\rangle, u \in E$ and of the norm $\|X\|$ of random variable X. Many results that hold for Gaussian random variables are also true in a slightly more general sub-Gaussian case.

Definition 2. *A centered random variable X in E will be called sub-Gaussian iff, for all $u \in E^*$,*

$$\big\|\langle X, u\rangle\big\|_{\psi_2} \lesssim \big\|\langle X, u\rangle\big\|_{L_2(\mathbb{P})}.$$

We will also need the following definition (see [8], p. 260 for more details).

Definition 3. *A weakly square integrable centered random variable X in E with covariance operator Σ is called pre-Gaussian iff there exists a centered Gaussian random variable Y in E with the same covariance operator Σ.*

Suppose now that $E = \mathbb{R}^d$ for some $d \geq 1$. It will be viewed as a standard Euclidean space. Then the following result is well known (it is a slight modification of Theorem 5.39 in Vershynin [16] stated there for isotropic sub-Gaussian random variables, i.e., when Σ is the identity operator).

Theorem 1. *There exists an absolute constant $C > 0$ such that, for all $t \geq 1$, with probability at least $1 - e^{-t}$*

$$\|\hat{\Sigma} - \Sigma\| \leq C\|\Sigma\| \bigg(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n} \bigg).$$

The proof of this theorem is based on a simple ε-net argument that allows one to reduce bounding the operator norm $\|\hat{\Sigma} - \Sigma\|$ to bounding the finite maximum

$$\max_{u \in M} \big|\langle(\hat{\Sigma} - \Sigma)u, u\rangle\big| = \max_{u \in M} \bigg| n^{-1} \sum_{j=1}^{n} \langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2 \bigg|,$$

where $M \subset S^{d-1}$ is a 1/4-net of the unit sphere of cardinality $\mathrm{card}(M) \le 9^d$. The bounding of the finite maximum is based on a version of Bernstein inequality for the sum of independent ψ_1 random variables $\langle X_j, u\rangle^2$ combined with the union bound (see the proof of Theorem 5.39 in [16] and the comments after this theorem).

In the isotropic case (i.e., when $\Sigma = I_d$), the bound of Theorem 1 is sharp and it can be viewed as a nonasymptotic version of the well-known Bai–Yin theorem from the asymptotic theory of random matrices. In the cases when the distribution of X is far from being isotropic, this bound is no longer sharp and it clearly cannot be used in the infinite-dimensional case. If the covariance operator Σ is of a small rank, it is natural to expect that the rank of Σ rather than the dimension of the space E should be involved in the bound. It turns out that one can obtain bounds on the operator norm $\|\hat\Sigma - \Sigma\|$ in terms of the "effective rank" $\tilde{\mathbf{r}}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$ of the covariance operator Σ (i.e., always dominated by its actual rank). This could be done, for instance, using noncommutative Bernstein-type inequalities that go back to Ahlswede and Winter [2] (see also Tropp [15], Koltchinskii [6]). For instance, Lounici [9] showed that with some constant $C > 0$ and with probability at least $1 - e^{-t}$

$$\|\hat\Sigma - \Sigma\| \le C\|\Sigma\| \max\left\{\sqrt{\frac{\tilde{\mathbf{r}}(\Sigma)\log d + t}{n}}, \frac{(\tilde{\mathbf{r}}(\Sigma)\log d + t)\log n}{n}\right\}.$$

Another approach to bounding the operator norm $\|\hat\Sigma - \Sigma\|$ was developed by Rudelson [13] and it is based on a noncommutative Khintchine inequality due to Lust-Picard and Pisier [10]. This method can be used not only in sub-Gaussian, but also in "heavy tailed" cases and it leads, for instance, to the following expectation bound (see Vershynin [16], Theorem 5.48):

$$\mathbb{E}\|\hat\Sigma - \Sigma\| \lesssim \max\left\{\|\Sigma\|^{1/2}\mathbb{E}^{1/2}\max_{1\le j\le n}\|X_j\|^2\sqrt{\frac{\log d}{n}}, \mathbb{E}\max_{1\le j\le n}\|X_j\|^2\frac{\log d}{n}\right\}.$$

Note that, in the sub-Gaussian case,

$$\big\|\|X\|^2\big\|_{\psi_1} \lesssim \mathrm{tr}(\Sigma),$$

which implies that

$$\mathbb{E}\max_{1\le j\le n}\|X_j\|^2 \lesssim \mathrm{tr}(\Sigma)\log n = \|\Sigma\|\tilde{\mathbf{r}}(\Sigma)\log n.$$

Therefore, in this case we get

$$\mathbb{E}\|\hat\Sigma - \Sigma\| \lesssim \|\Sigma\| \max\left\{\sqrt{\frac{\tilde{\mathbf{r}}(\Sigma)\log d\log n}{n}}, \frac{\tilde{\mathbf{r}}(\Sigma)\log d\log n}{n}\right\}.$$

In each of the above approaches, the bounds are not dimension free (at least, with a straightforward application of noncommutative Bernstein or Khintchine inequalities) and they could not be directly used in the infinite-dimensional case. We will use below a different approach based on recent deep results on generic chaining bounds for empirical processes. The following facts about generic chaining complexities will be needed. Let $N_n := 2^{2^n}, n \ge 1$ and $N_0 := 1$. Given a metric space (T, d), an increasing sequence Δ_n of partitions of T is called admissible if $\mathrm{card}(\Delta_n) \le N_n$. For $t \in T$, $\Delta_n(t)$ denotes the unique set of the partition Δ_n that contains t. For $A \subset T$, $D(A)$ denotes the diameter of set A. Define

$$\gamma_2(T, d) = \inf \sup_{t \in T} \sum_{n=0}^{\infty} 2^{n/2} D\big(\Delta_n(t)\big),$$

where the infimum is taken over all admissible sequences.

The following fundamental result is due to Talagrand (see [14]; it was initially stated it terms of majorizing measures rather than generic complexities).

Theorem 2. *Let $X(t), t \in T$ be a centered Gaussian process and suppose that*

$$d(t, s) := \mathbb{E}^{1/2}\big(X(t) - X(s)\big)^2, \qquad t, s \in T.$$

Then there exists an absolute constant $K > 0$ such that

$$\mathbb{E} \sup_{t \in T} X(t) \ge K^{-1} \gamma_2(T; d).$$

In what follows, generic chaining complexities are used in the case when $T = \mathcal{F}$ is a function class on a probability space $(S, \mathcal{A}, P)$ and d is the metric generated by either $L_2(P)$-norm, or by the ψ_2-norm with respect to P. We will use the following result due to Mendelson [11] (although an earlier, simpler and weaker version, with $\sup_{f \in \mathcal{F}} \|f\|_{\psi_2}$ instead of $\sup_{f \in \mathcal{F}} \|f\|_{\psi_1}$ that goes back to Klartag and Mendelson [5] would suffice for our purposes).

Theorem 3. *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables in S with common distribution P and let $\mathcal{F}$ be a class of measurable functions on $(S, \mathcal{A})$ such that $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$ and $\mathbb{E} f(X) = 0$. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f^2(X_i) - \mathbb{E} f^2(X) \right| \lesssim \max\left\{ \sup_{f \in \mathcal{F}} \|f\|_{\psi_1} \frac{\gamma_2(\mathcal{F}; \psi_2)}{\sqrt{n}}, \frac{\gamma_2^2(\mathcal{F}; \psi_2)}{n} \right\}.$$

Assume again that E is an arbitrary separable Banach space. The next result provides a characterization of the size of $\mathbb{E}\|\hat{\Sigma} - \Sigma\|$ in terms of the parameters $\|\Sigma\|$ and $\mathbf{r}(\Sigma)$ for Gaussian random variable X (the upper bound also holds in the case when X is both sub-Gaussian and pre-Gaussian).

Theorem 4. *Let $X, X_1, \ldots, X_n$ be i.i.d. weakly square integrable centered random vectors in E with covariance operator Σ. If X is sub-Gaussian and pre-Gaussian, then*

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \lesssim \|\Sigma\| \max\left\{ \sqrt{\frac{\mathbf{r}(\Sigma)}{n}}, \frac{\mathbf{r}(\Sigma)}{n} \right\}.$$

Moreover, if X is Gaussian, then

$$\|\Sigma\| \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}, \frac{\mathbf{r}(\Sigma)}{n}\right\} \lesssim \mathbb{E}\|\hat{\Sigma}-\Sigma\| \lesssim \|\Sigma\| \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}, \frac{\mathbf{r}(\Sigma)}{n}\right\}.$$

Proof. The proof of the upper bound relies on the generic chaining bound of Theorem 3, while the proof of the lower bound is rather elementary.

Upper bound. We have

$$\begin{aligned}
\mathbb{E}\|\hat{\Sigma}-\Sigma\| &= \mathbb{E}\sup_{\|u\|\le 1, \|v\|\le 1} \big\langle(\hat{\Sigma}-\Sigma)u, v\big\rangle \\
&= \mathbb{E}\sup_{\|u\|\le 1, \|v\|\le 1}\left(\left\langle(\hat{\Sigma}-\Sigma)\frac{u+v}{2}, \frac{u+v}{2}\right\rangle - \left\langle(\hat{\Sigma}-\Sigma)\frac{u-v}{2}, \frac{u-v}{2}\right\rangle\right) \\
&\le 2\sup_{\|u\|\le 1}\big|\big\langle(\hat{\Sigma}-\Sigma)u, u\big\rangle\big| = 2\mathbb{E}\sup_{\|u\|\le 1}\left|\frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle^2 - \langle \Sigma u, u\rangle\right| \\
&= 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f^2(X_i) - \mathbb{E}f^2(X)\right|,
\end{aligned}$$

where $\mathcal{F} := \{\langle\cdot, u\rangle : u \in U_{E^*}\}$, $U_{E^*} := \{u \in E^* : \|u\| \le 1\}$ and P is the distribution of random variable X.

Since X is sub-Gaussian, the ψ_1- and ψ_2-norms of linear functionals $\langle X, u\rangle$ are both equivalent to the L_2-norm. This implies that

$$\sup_{f\in\mathcal{F}}\|f\|_{\psi_1} \lesssim \sup_{u\in U_{E^*}} \mathbb{E}^{1/2}\langle X, u\rangle^2 \le \|\Sigma\|^{1/2}.$$

Also, since X is pre-Gaussian, there exists a centered Gaussian random variable Y in E with the same covariance Σ. This means that

$$d_Y(u, v) = \mathbb{E}^{1/2}\big(\langle Y, u\rangle - \langle Y, v\rangle\big)^2 = \big\|\langle\cdot, u\rangle - \langle\cdot, v\rangle\big\|_{L_2(P)}, \qquad u, v \in U_{E^*}.$$

Using Talagrand's Theorem 2, we easily get that

$$\gamma_2(\mathcal{F}, \psi_2) \lesssim \gamma_2(\mathcal{F}, L_2) = \gamma_2(U_{E^*}; d_Y) \lesssim \mathbb{E}\sup_{u\in U_{E^*}}\langle Y, u\rangle \le \mathbb{E}\|Y\|.$$

Therefore, it follows that

$$\mathbb{E}\|\hat{\Sigma}-\Sigma\| \lesssim \max\left\{\|\Sigma\|^{1/2}\frac{\mathbb{E}\|Y\|}{\sqrt{n}}, \frac{(\mathbb{E}\|Y\|)^2}{n}\right\} \lesssim \|\Sigma\| \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}, \frac{\mathbf{r}(\Sigma)}{n}\right\},$$

which proves the upper bound.

Lower bound. To prove the lower bound, note that

$$\begin{aligned}\mathbb{E}\|\hat{\Sigma}-\Sigma\| &= \mathbb{E}\sup_{\|u\|\le 1}\Big\|n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle X_j-\mathbb{E}\langle X,u\rangle X\Big\| \\ &\ge \sup_{\|u\|\le 1}\mathbb{E}\Big\|n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle X_j-\mathbb{E}\langle X,u\rangle X\Big\|.\end{aligned} \tag{2.1}$$

For a fixed $u\in E^*$ with $\|u\|\le 1$ and $\langle\Sigma u,u\rangle>0$, denote

$$X' := X-\langle X,u\rangle\frac{\Sigma u}{\langle\Sigma u,u\rangle}.$$

By a straightforward computation, for all $v\in E^*$, the random variables $\langle X,u\rangle$ and $\langle X',v\rangle$ are uncorrelated. Since they are jointly Gaussian, it follows that $\langle X,u\rangle$ and X' are independent. Define

$$X_j' := X_j-\langle X_j,u\rangle\frac{\Sigma u}{\langle\Sigma u,u\rangle},\qquad j=1,\dots,n.$$

Then $\{X_j' : j=1,\dots,n\}$ and $\{\langle X_j,u\rangle : j=1,\dots,n\}$ are also independent. We easily get

$$\begin{aligned}&\mathbb{E}\Big\|n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle X_j-\mathbb{E}\langle X,u\rangle X\Big\| \\ &\quad=\mathbb{E}\Big\|n^{-1}\sum_{j=1}^{n}\big(\langle X_j,u\rangle^2-\mathbb{E}\langle X,u\rangle^2\big)\frac{\Sigma u}{\langle\Sigma u,u\rangle}+n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle X_j'\Big\|,\end{aligned} \tag{2.2}$$

where we used the fact that

$$\mathbb{E}\langle X,u\rangle X=\mathbb{E}\langle X,u\rangle^2\frac{\Sigma u}{\langle\Sigma u,u\rangle}+\mathbb{E}\langle X,u\rangle\mathbb{E}X'=\mathbb{E}\langle X,u\rangle^2\frac{\Sigma u}{\langle\Sigma u,u\rangle}.$$

Note that, conditionally on $\langle X_j,u\rangle$, $j=1,\dots,n$, the distribution of random variable

$$n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle X_j'$$

is Gaussian and it coincides with the distribution of the random variable

$$\Big(n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle^2\Big)^{1/2}\frac{X'}{\sqrt{n}}.$$

Denote now by $\mathbb{E}_u$ the conditional expectation given $\langle X_j, u\rangle$, $j = 1, \dots, n$ and by $\mathbb{E}'$ the conditional expectation given $X_1', \dots, X_n'$. Then we have

$$\begin{aligned}
&\mathbb{E}\left\| n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle X_j'\right\| \\
&= \mathbb{E}\mathbb{E}_u\left\| n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle X_j'\right\| \\
&= \mathbb{E}\mathbb{E}_u\left\| n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + \left(n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2\right)^{1/2}\frac{X'}{\sqrt{n}}\right\| \\
&= \mathbb{E}\left\| n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + \left(n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2\right)^{1/2}\frac{X'}{\sqrt{n}}\right\|.
\end{aligned}$$

Also

$$\begin{aligned}
&\mathbb{E}\left\| n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + \left(n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2\right)^{1/2}\frac{X'}{\sqrt{n}}\right\| \\
&= \mathbb{E}\mathbb{E}'\left\| n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + \left(n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2\right)^{1/2}\frac{X'}{\sqrt{n}}\right\| \\
&\geq \mathbb{E}\left\| \mathbb{E}' n^{-1}\sum_{j=1}^{n}\big(\langle X_j, u\rangle^2 - \mathbb{E}\langle X, u\rangle^2\big)\frac{\Sigma u}{\langle \Sigma u, u\rangle} + \mathbb{E}'\left(n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2\right)^{1/2}\frac{X'}{\sqrt{n}}\right\| \\
&= \mathbb{E}\left(n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2\right)^{1/2}\frac{\mathbb{E}\|X'\|}{\sqrt{n}}.
\end{aligned}$$

Note that

$$\mathbb{E}\big|\langle X, u\rangle\big| = \sqrt{\frac{2}{\pi}}\langle \Sigma u, u\rangle^{1/2}.$$

Therefore,

$$\mathbb{E}\big\|X'\big\| \geq \mathbb{E}\|X\| - \mathbb{E}\big|\langle X, u\rangle\big|\frac{\|\Sigma u\|}{\langle \Sigma u, u\rangle} = \mathbb{E}\|X\| - \sqrt{\frac{2}{\pi}}\frac{\|\Sigma u\|}{\langle \Sigma u, u\rangle^{1/2}}$$

and

$$\mathbb{E}\left\| n^{-1}\sum_{j=1}^{n}\bigl(\langle X_j,u\rangle^2-\mathbb{E}\langle X,u\rangle^2\bigr)\frac{\Sigma u}{\langle \Sigma u,u\rangle}+\left(n^{-1}\sum_{j=1}^{n}\langle X_j,u^2\rangle\right)^{1/2}\frac{X'}{\sqrt{n}}\right\|$$
$$\geq \langle \Sigma u,u\rangle^{1/2}\mathbb{E}\left(n^{-1}\sum_{j=1}^{n}Z_j^2\right)^{1/2}\frac{\mathbb{E}\|X\|-\sqrt{2/\pi}\,\|\Sigma u\|/\langle \Sigma u,u\rangle^{1/2}}{\sqrt{n}},$$

where

$$Z_j=\frac{\langle X_j,u\rangle}{\langle \Sigma u,u\rangle^{1/2}},\qquad j=1,\dots,n$$

are i.i.d. standard normal random variables. It is easy to check that

$$\mathbb{E}\left(n^{-1}\sum_{j=1}^{n}Z_j^2\right)^{1/2}\geq c_2$$

for a positive numerical constant c_2, implying that

$$\mathbb{E}\left\| n^{-1}\sum_{j=1}^{n}\bigl(\langle X_j,u\rangle^2-\mathbb{E}\langle X,u\rangle^2\bigr)\frac{\Sigma u}{\langle \Sigma u,u\rangle}+\left(n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle^2\right)^{1/2}\frac{X'}{\sqrt{n}}\right\|$$
$$\geq c_2\frac{\langle \Sigma u,u\rangle^{1/2}\mathbb{E}\|X\|-\sqrt{2/\pi}\,\|\Sigma u\|}{\sqrt{n}}.$$

We now combine this bound with (2.1) and (2.2) to get

$$\mathbb{E}\|\hat{\Sigma}-\Sigma\|\geq c_2\sup_{\|u\|\leq 1}\frac{\langle \Sigma u,u\rangle^{1/2}\mathbb{E}\|X\|-\sqrt{2/\pi}\,\|\Sigma u\|}{\sqrt{n}}$$
$$\geq c_2\frac{\|\Sigma\|^{1/2}\mathbb{E}\|X\|-\sqrt{2/\pi}\,\|\Sigma\|}{\sqrt{n}}\geq c_2\|\Sigma\|\left(\frac{\sqrt{\mathbf{r}(\Sigma)}-\sqrt{2/\pi}}{\sqrt{n}}\right).$$

We also have the following obvious bound:

$$\mathbb{E}\|\hat{\Sigma}-\Sigma\|\geq \sup_{\|u\|\leq 1}\mathbb{E}\left|n^{-1}\sum_{j=1}^{n}\langle X_j,u\rangle^2-\mathbb{E}\langle X,u\rangle^2\right|$$
$$=\sup_{\|u\|\leq 1}\langle \Sigma u,u\rangle\mathbb{E}\left|n^{-1}\sum_{j=1}^{n}Z_j^2-1\right|\geq c_3\frac{\|\Sigma\|}{\sqrt{n}}$$

for some numerical constant $c_3 > 0$. Thus, we get

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \geq c_2\|\Sigma\|\left(\frac{\sqrt{\mathbf{r}(\Sigma)} - \sqrt{2/\pi}}{\sqrt{n}}\right) \vee c_3\frac{\|\Sigma\|}{\sqrt{n}}$$
$$\geq \frac{1}{2}\left(c_2\|\Sigma\|\left(\frac{\sqrt{\mathbf{r}(\Sigma)} - \sqrt{2/\pi}}{\sqrt{n}}\right) + c_3\frac{\|\Sigma\|}{\sqrt{n}}\right) \geq \frac{c_2}{2}\|\Sigma\|\frac{\sqrt{\mathbf{r}(\Sigma)}}{\sqrt{n}},$$

provided c_2 is chosen to be small enough to satisfy $c_2\sqrt{\frac{2}{\pi}} \leq c_3$.

This completes the proof in the case when $\mathbf{r}(\Sigma) \leq 2n$ since in this case

$$\frac{\mathbf{r}(\Sigma)}{n} \lesssim \sqrt{\frac{\mathbf{r}(\Sigma)}{n}}.$$

On the other hand, under the assumption that $\mathbf{r}(\Sigma) \geq 2n$,

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \geq \mathbb{E}\|\hat{\Sigma}\| - \|\Sigma\| \geq \mathbb{E}\sup_{\|u\|\leq 1} n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle^2 - \|\Sigma\|$$
$$\geq \mathbb{E}\sup_{\|u\|\leq 1}\frac{\langle X_1, u\rangle^2}{n} - \|\Sigma\| \geq \frac{\mathbb{E}\|X\|^2}{n} - \|\Sigma\| \tag{2.3}$$
$$\geq \frac{(\mathbb{E}\|X\|)^2}{n} - \|\Sigma\| = \|\Sigma\|\left(\frac{\mathbf{r}(\Sigma)}{n} - 1\right) \geq \frac{1}{2}\|\Sigma\|\frac{\mathbf{r}(\Sigma)}{n},$$

which completes the proof in the case when $\mathbf{r}(\Sigma) \geq 2n$. □

Our next goal is to prove a concentration inequality for $\|\hat{\Sigma} - \Sigma\|$ around its median or around its expectation. In what follows, $\mathrm{Med}(\xi)$ denotes a median of a random variable ξ.

Theorem 5. *Let $X, X_1, \ldots, X_n$ be i.i.d. centered Gaussian random vectors in E with covariance Σ and let M be either the median, or the expectation of $\|\hat{\Sigma} - \Sigma\|$. Then there exists a constant $C > 0$ such that the following holds. If $\mathbf{r}(\Sigma) \leq n$, then for all $t \geq 1$, with probability at least $1 - e^{-t}$,*

$$\left|\|\hat{\Sigma} - \Sigma\| - M\right| \leq C\|\Sigma\|\left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right). \tag{2.4}$$

On the other hand, if $\mathbf{r}(\Sigma) \geq n$, then with the same probability

$$\left|\|\hat{\Sigma} - \Sigma\| - M\right| \leq C\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right). \tag{2.5}$$

In the case when M is the median, this result is an immediate consequence of Theorem 4 and Theorem 6 that is given below and that provides an equivalent concentration inequality written in a somewhat implicit form.

The bounds of Theorem 5 in the case when M is the median imply that

$$\big|\mathbb{E}\|\hat{\Sigma}-\Sigma\| - \mathrm{Med}\big(\|\hat{\Sigma}-\Sigma\|\big)\big| \lesssim \|\Sigma\|\frac{1}{\sqrt{n}}$$

when $\mathbf{r}(\Sigma)\le n$, and

$$\big|\mathbb{E}\|\hat{\Sigma}-\Sigma\| - \mathrm{Med}\big(\|\hat{\Sigma}-\Sigma\|\big)\big| \lesssim \|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\frac{1}{\sqrt{n}},$$

when $\mathbf{r}(\Sigma)\ge n$. This, in turn, implies the concentration bound in the case when M is the expectation.

Theorem 6. *Let $X, X_1, \dots, X_n$ be i.i.d. centered Gaussian random vectors in E with covariance Σ and let M be the median of $\|\hat{\Sigma}-\Sigma\|$. Then there exists a constant $C>0$ such that for all $t\ge 1$ with probability at least $1-e^{-t}$,*

$$\big|\|\hat{\Sigma}-\Sigma\| - M\big| \le C\bigg[\|\Sigma\|\bigg(\sqrt{\frac{t}{n}}\vee\frac{t}{n}\bigg)\vee\|\Sigma\|^{1/2}M^{1/2}\sqrt{\frac{t}{n}}\bigg]. \tag{2.6}$$

The proof of Theorem 6 is somewhat long and will be given in the next section. Here, we will state a couple corollaries of this theorem.

Corollary 1. *Under the assumptions and notation of Theorem 6, there exists a constant $C>0$ such that, for all $t\ge 1$, with probability at least $1-e^{-t}$,*

$$\|\hat{\Sigma}-\Sigma\| \le 2M + C\|\Sigma\|\bigg(\sqrt{\frac{t}{n}}\vee\frac{t}{n}\bigg). \tag{2.7}$$

Proof. The proof easily follows from the next simple bound: $2\|\Sigma\|^{1/2}M^{1/2}\sqrt{\frac{t}{n}} \le M + \|\Sigma\|\frac{t}{n}$. □

The following corollary can be viewed as an infinite-dimensional generalization of Theorem 1.

Corollary 2. *Under the assumptions and notation of Theorem 6, there exists a constant $C>0$ such that, for all $t\ge 1$, with probability at least $1-e^{-t}$,*

$$\|\hat{\Sigma}-\Sigma\| \le C\|\Sigma\|\bigg(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\vee\frac{\mathbf{r}(\Sigma)}{n}\vee\sqrt{\frac{t}{n}}\vee\frac{t}{n}\bigg). \tag{2.8}$$

This implies that for all $p\ge 1$

$$\mathbb{E}^{1/p}\|\hat{\Sigma}-\Sigma\|^p \lesssim_p \|\Sigma\|\max\bigg\{\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}, \frac{\mathbf{r}(\Sigma)}{n}\bigg\}. \tag{2.9}$$

Proof. Bound (2.8) follows immediately from Corollary 1 and Theorem 4. Bound (2.9) follows from (2.8) by integrating the tail probabilities. □

To derive Theorem 5 from Theorem 6 in the case when M is the median, it is enough to observe that $M \le 2\mathbb{E}\|\hat{\Sigma} - \Sigma\|$. Therefore, due to Theorem 4, $M^{1/2} \lesssim \|\Sigma\|^{1/2}$ when $\mathbf{r}(\Sigma) \le n$ and $M^{1/2} \lesssim \|\Sigma\|^{1/2}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}$ otherwise. Substituting these bounds on $M^{1/2}$ in the right-hand side of inequality of Theorem 6 immediately implies the bounds of Theorem 5 in the case when M is the median of $\|\hat{\Sigma} - \Sigma\|$. If now $M = \mathbb{E}\|\hat{\Sigma} - \Sigma\|$, use a well-known identity $\mathbb{E}|Z| = \int_0^\infty \mathbb{P}\{|Z| > t\}\,dt$ for $Z = |\|\hat{\Sigma} - \Sigma\| - \mathrm{Med}(\|\hat{\Sigma} - \Sigma\|)|$ and integrate out the tail probability bounds to get

$$\big|\mathbb{E}\|\hat{\Sigma} - \Sigma\| - \mathrm{Med}\big(\|\hat{\Sigma} - \Sigma\|\big)\big| \le \mathbb{E}Z \le C\|\Sigma\|\left[\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee 1\right]\sqrt{\frac{1}{n}}.$$

The result in the case of the mean follows.

3. Proof of the concentration inequality

In this section, we provide a proof of Theorem 6. We will use the following well-known fact (see, e.g., [7]).

Theorem 7. *Let X be a centered Gaussian random variable in a separable Banach space E. Then there exists a sequence $\{x_k : k \ge 1\}$ of vectors in E and a sequence $\{Z_k : k \ge 1\}$ of i.i.d. standard normal random variables such that*

$$X = \sum_{k=1}^{\infty} Z_k x_k,$$

where the series in the right-hand side converges in E a.s. and

$$\sum_{k=1}^{\infty} \|x_k\|^2 < +\infty.$$

Note that under the assumptions and notation of Theorem 7,

$$\Sigma u = \sum_{k=1}^{\infty} \langle x_k, u\rangle x_k, \qquad u \in E^*.$$

It easily follows from Theorem 7 that, for $X^{(m)} := \sum_{k=1}^{m} Z_k x_k$, we have

$$\mathbb{E}\big\|X^{(m)} - X\big\|^2 \to 0 \qquad \text{as } m \to \infty.$$

Let now $\Sigma^{(m)}$ be the covariance operator of $X^{(m)}$ and $\hat{\Sigma}^{(m)}$ be the sample covariance operator based on observations $(X_1^{(m)}, \dots, X_n^{(m)})$ (with the notation $X_j^{(m)}$ having an obvious meaning and the sample size n being fixed). Then

$$\big\| \Sigma^{(m)} - \Sigma \big\| \to 0 \quad \text{and} \quad \mathbb{E}\big\| \hat{\Sigma}^{(m)} - \hat{\Sigma} \big\| \to 0 \qquad \text{as } m \to \infty.$$

Thus, it is enough to prove the theorem only in the case when

$$X = X^{(m)} = \sum_{k=1}^{m} Z_k x_k.$$

The general case would then follow by a straightforward limiting argument.

The main ingredient of the proof is the classical Gaussian concentration inequality (see, e.g., Ledoux and Talagrand [8], p. 21).

Lemma 1. *Let $Z = (Z_1, \dots, Z_N)$ be a standard normal vector in $\mathbb{R}^N$ and let $f : \mathbb{R}^N \mapsto \mathbb{R}$ be a function satisfying the following Lipschitz condition with some $L > 0$:*

$$\big| f(z_1, \dots, z_N) - f\big(z_1', \dots, z_N'\big)\big| \le L \left(\sum_{j=1}^{N} \big| z_j - z_j' \big|^2 \right)^{1/2}, \qquad z_1, \dots, z_N, z_1', \dots, z_N' \in \mathbb{R}.$$

Then, for all $t > 0$,

$$\mathbb{P}\big\{ \big| f(Z) - \operatorname{Med}\big(f(Z)\big)\big| \ge t \big\} \le 2\left(1 - \Phi\left(\frac{t}{L}\right)\right),$$

where Φ is the distribution function of a standard normal random variable.

This result easily follows from the Gaussian isoperimetric inequality. We will also need another consequence of this inequality.

Lemma 2. *Under the assumptions of Lemma 1, suppose that for some M and for some $\alpha > 0$*

$$\mathbb{P}\big\{ f(Z) \ge M \big\} \ge \alpha.$$

Then there exists a constant $D > 0$ (possibly depending on α) such that, for all $t \ge 1$, with probability at least $1 - e^{-t}$,

$$f(Z) \ge M - DL\sqrt{t}.$$

Denote

$$g(X_1, \dots, X_n) := \|W\| \varphi\left(\frac{\|W\|}{\delta} \right),$$

where

$$W = \hat{\Sigma} - \Sigma, \qquad \hat{\Sigma}u = n^{-1}\sum_{j=1}^{n}\langle X_j, u\rangle X_j,$$

where φ is an arbitrary fixed Lipschitz function with constant 1 on $\mathbb{R}_+$, $0 \le \varphi(s) \le 1$, $\varphi(s) = 1, s \le 1$, $\varphi(s) = 0, s > 2$, and where $\delta > 0$ is a fixed number (to be chosen later). With a little abuse of notation, assume for now that $Z := (Z_{k,j}, k = 1, \dots, m, j = 1, \dots, n) \in \mathbb{R}^{mn}$, $Z' := (Z'_{k,j}, k = 1, \dots, m, j = 1, \dots, n) \in \mathbb{R}^{mn}$ are nonrandom vectors in $\mathbb{R}^{mn}$ and $X_1, \dots, X_n, X'_1, \dots, X'_n$ are nonrandom vectors in E defined as follows:

$$X_j = \sum_{k=1}^{m} Z_{k,j}x_k, \qquad X'_j = \sum_{k=1}^{m} Z'_{k,j}x_k.$$

Lemma 1 will be applied to the function $f(Z) = g(X_1, \dots, X_n)$. We have to check the Lipschitz condition for this function. To this end, we will prove the following lemma.

Lemma 3. *There exists a numerical constant $D > 0$ such that, for all $Z, Z' \in \mathbb{R}^{mn}$,*

$$\bigl|f(Z) - f\bigl(Z'\bigr)\bigr| \le D\frac{\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}}{\sqrt{n}}\Biggl(\sum_{j=1}^{n}\sum_{k=1}^{m}\bigl|Z_{k,j} - Z'_{k,j}\bigr|^2\Biggr)^{1/2}. \tag{3.1}$$

Proof. Obviously, $0 \le g(X_1, \dots, X_n) \le 2\delta$, $0 \le g(X'_1, \dots, X'_n) \le 2\delta$, implying that

$$\bigl|g(X_1, \dots, X_n) - g\bigl(X'_1, \dots, X'_n\bigr)\bigr| \le 2\delta. \tag{3.2}$$

Denote

$$W' = \hat{\Sigma}' - \Sigma, \qquad \hat{\Sigma}'u = n^{-1}\sum_{j=1}^{n}\bigl\langle X'_j, u\bigr\rangle X'_j, \qquad u \in E^*.$$

It is enough to consider the case when $\|W\| \le 2\delta$ or $\|W'\| \le 2\delta$ (otherwise, the claim of the lemma is obvious). To be specific, assume that $\|W\| \le 2\delta$. Then, using the assumption that φ is Lipschitz with constant 1, we get

$$\begin{aligned}
&\bigl|g(X_1, \dots, X_n) - g\bigl(X'_1, \dots, X'_n\bigr)\bigr| \\
&\quad = \biggl|\|W\|\varphi\biggl(\frac{\|W\|}{\delta}\biggr) - \bigl\|W'\bigr\|\varphi\biggl(\frac{\|W'\|}{\delta}\biggr)\biggr| \\
&\quad \le \bigl\|W - W'\bigr\| + \frac{\|W\|}{\delta}\bigl\|W - W'\bigr\| \\
&\quad \le 3\bigl\|W - W'\bigr\|.
\end{aligned} \tag{3.3}$$

We will now control $\|W - W'\|$. Note that

$$
\begin{aligned}
\big\|W - W'\big\| &= \sup_{\|u\|\le 1, \|v\|\le 1} \big|\big\langle (W - W')u, v\big\rangle\big| \\
&= \sup_{\|u\|\le 1, \|v\|\le 1} \left| n^{-1}\sum_{j=1}^{n} \langle X_j, u\rangle \langle X_j, v\rangle - \big\langle X'_j, u\big\rangle\big\langle X'_j, v\big\rangle \right| \\
&\le \sup_{\|u\|\le 1, \|v\|\le 1} \left| n^{-1}\sum_{j=1}^{n} \langle X_j, u\rangle \big\langle X_j - X'_j, v\big\rangle \right| \\
&\quad + \sup_{\|u\|\le 1, \|v\|\le 1} \left| n^{-1}\sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle \big\langle X'_j, v\big\rangle \right| \\
&\le \sup_{\|u\|\le 1} \left(n^{-1}\sum_{j=1}^{n} \langle X_j, u\rangle^2 \right)^{1/2} \sup_{\|v\|\le 1} \left(n^{-1}\sum_{j=1}^{n} \big\langle X_j - X'_j, v\big\rangle^2 \right)^{1/2} \\
&\quad + \sup_{\|u\|\le 1} \left(n^{-1}\sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle^2 \right)^{1/2} \sup_{\|v\|\le 1} \left(n^{-1}\sum_{j=1}^{n} \big\langle X'_j, v\big\rangle^2 \right)^{1/2} \\
&\le \frac{\|\hat\Sigma\|^{1/2} + \|\hat\Sigma'\|^{1/2}}{\sqrt{n}} \sup_{\|u\|\le 1} \left(\sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle^2 \right)^{1/2}.
\end{aligned}
$$

Since $\|W\| \le 2\delta$,

$$
\|\hat\Sigma\|^{1/2} + \big\|\hat\Sigma'\big\|^{1/2} \le 2\|\hat\Sigma\|^{1/2} + \big\|W - W'\big\|^{1/2} \le 2\|\Sigma\|^{1/2} + 2\sqrt{2\delta} + \big\|W - W'\big\|^{1/2}.
$$

Therefore,

$$
\begin{aligned}
\big\|W - W'\big\| \le{}& \frac{2\|\Sigma\|^{1/2} + 2\sqrt{2\delta}}{\sqrt{n}} \sup_{\|u\|\le 1} \left(\sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle^2 \right)^{1/2} \\
&+ \frac{\|W - W'\|^{1/2}}{\sqrt{n}} \sup_{\|u\|\le 1} \left(\sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle^2 \right)^{1/2},
\end{aligned}
$$

which easily implies

$$
\begin{aligned}
&\big\|W - W'\big\| \\
&\quad \le \frac{4\|\Sigma\|^{1/2} + 4\sqrt{2\delta}}{\sqrt{n}} \sup_{\|u\|\le 1} \left(\sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle^2 \right)^{1/2} \vee \frac{4}{n} \sup_{\|u\|\le 1} \sum_{j=1}^{n} \big\langle X_j - X'_j, u\big\rangle^2.
\end{aligned}
\tag{3.4}
$$

Substituting the last bound in (3.3), we get

$$\begin{aligned} &\left|g(X_1,\dots,X_n)-g\left(X_1',\dots,X_n'\right)\right| \\ &\le 12\frac{\|\Sigma\|^{1/2}+\sqrt{2\delta}}{\sqrt{n}}\sup_{\|u\|\le 1}\left(\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\right)^{1/2}\vee\frac{12}{n}\sup_{\|u\|\le 1}\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2. \end{aligned} \tag{3.5}$$

In view of (3.2), the left-hand side is also bounded from above by 2δ, which allows one to get from (3.5) that

$$\begin{aligned} &\left|g(X_1,\dots,X_n)-g\left(X_1',\dots,X_n'\right)\right| \\ &\le 12\frac{\|\Sigma\|^{1/2}+\sqrt{2\delta}}{\sqrt{n}}\sup_{\|u\|\le 1}\left(\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\right)^{1/2}\vee\left(\frac{12}{n}\sup_{\|u\|\le 1}\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\wedge 2\delta\right). \end{aligned} \tag{3.6}$$

In the case when

$$\sup_{\|u\|\le 1}\left(\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\right)^{1/2}\le\sqrt{\frac{\delta n}{6}},$$

we have

$$\frac{12}{n}\sup_{\|u\|\le 1}\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\wedge 2\delta\le\frac{12}{\sqrt{6}}\frac{\sqrt{\delta}}{\sqrt{n}}\sup_{\|u\|\le 1}\left(\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\right)^{1/2}.$$

It is easy to check that the same bound holds in the opposite case, also. As a consequence, (3.6) implies that with some numerical constant $D>0$,

$$\begin{aligned} &\left|g(X_1,\dots,X_n)-g\left(X_1',\dots,X_n'\right)\right| \\ &\le D\frac{\|\Sigma\|^{1/2}+\sqrt{\delta}}{\sqrt{n}}\sup_{\|u\|\le 1}\left(\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\right)^{1/2}. \end{aligned} \tag{3.7}$$

We will now upper bound

$$\sup_{\|u\|\le 1}\left(\sum_{j=1}^n\left\langle X_j-X_j',u\right\rangle^2\right)^{1/2}.$$

Note that

$$X_j-X_j'=\sum_{k=1}^m\left(Z_{k,j}-Z_{k,j}'\right)x_k,$$

implying that

$$\begin{aligned}\sup_{\|u\|\le 1}\left(\sum_{j=1}^{n}\langle X_j - X'_j, u\rangle^2\right)^{1/2} &\le \sup_{\|u\|\le 1}\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{m}(Z_{k,j}-Z'_{k,j})\langle x_k, u\rangle\right)^2\right)^{1/2}\\ &\le \sup_{\|u\|\le 1}\left(\sum_{j=1}^{n}\sum_{k=1}^{m}(Z_{k,j}-Z'_{k,j})^2\sum_{k=1}^{m}\langle x_k, u\rangle^2\right)^{1/2}\\ &\le \sup_{\|u\|\le 1}\left(\sum_{k=1}^{m}\langle x_k, u\rangle^2\right)^{1/2}\left(\sum_{j=1}^{n}\sum_{k=1}^{m}(Z_{k,j}-Z'_{k,j})^2\right)^{1/2}\\ &= \left(\sup_{\|u\|\le 1}\langle \Sigma u, u\rangle\right)^{1/2}\left(\sum_{j=1}^{n}\sum_{k=1}^{m}(Z_{k,j}-Z'_{k,j})^2\right)^{1/2}\\ &= \|\Sigma\|^{1/2}\left(\sum_{j=1}^{n}\sum_{k=1}^{m}(Z_{k,j}-Z'_{k,j})^2\right)^{1/2}.\end{aligned}$$

Combining this with bound (3.7) yields (3.1). □

In what follows, denote

$$M := \operatorname{Med}(\|\hat{\Sigma}-\Sigma\|) \quad \text{and} \quad M_g := \operatorname{Med}(g(X_1,\dots,X_n)).$$

It follows from Lemmas 1 and 3 that, for all $t \ge 1$ with probability at least $1-e^{-t}$,

$$\left|g(X_1,\dots,X_n)-M_g\right| \le D_1\left(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\right)\sqrt{\frac{t}{n}}, \tag{3.8}$$

where D_1 is a numerical constant. We will use this bound to get that, on the event where $\|W\| \le \delta$ and, at the same time, concentration bound (3.8) holds, we have

$$\begin{aligned}\|W\| = g(X_1,\dots,X_n) &\le M_g + D_1\left(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\right)\sqrt{\frac{t}{n}}\\ &\le M + D_1\left(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\right)\sqrt{\frac{t}{n}}.\end{aligned}$$

Denote

$$A := M + D_1\|\Sigma\|\sqrt{\frac{t}{n}}, \qquad B := D_1\|\Sigma\|^{1/2}\sqrt{\frac{t}{n}}.$$

Then we have

$$\mathbb{P}\left\{\delta \ge \|W\| \ge A + B\sqrt{\delta}\right\} \le e^{-t}. \tag{3.9}$$

We will need the following easy fact.

Lemma 4. *There exists a constant $D_2 > 0$ such that for all $t > 0$, with probability at least $1-e^{-t}$*

$$\|\hat{\Sigma} - \Sigma\| \leq D_2\|\Sigma\|\biggl[\mathbf{r}(\Sigma)\biggl(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\biggr) + \mathbf{r}(\Sigma) + 1\biggr]. \tag{3.10}$$

In particular, this implies that, for some constant $D_2 > 0$,

$$M \leq D_2\|\Sigma\|\bigl(\mathbf{r}(\Sigma) + 1\bigr). \tag{3.11}$$

Proof. To prove (3.10), note that

$$\|\hat{\Sigma} - \Sigma\| \leq \biggl|n^{-1}\sum_{j=1}^{n}\|X_j\|^2 - \mathbb{E}\|X\|^2\biggr| + \mathbb{E}\|X\|^2 + \|\Sigma\|.$$

It remains to observe that

$$\mathbb{E}\|X\|^2 \lesssim \bigl(\mathbb{E}\|X\|\bigr)^2 = \|\Sigma\|\mathbf{r}(\Sigma)$$

and, by Bernstein's inequality for ψ_1-random variables, with probability at least $1 - e^{-t}$

$$\biggl|n^{-1}\sum_{j=1}^{n}\|X_j\|^2 - \mathbb{E}\|X\|^2\biggr| \lesssim \bigl\|\|X\|^2\bigr\|_{\psi_1}\biggl(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\biggr)$$

$$\lesssim \bigl(\mathbb{E}\|X\|\bigr)^2\biggl(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\biggr) = \|\Sigma\|\mathbf{r}(\Sigma)\biggl(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\biggr).$$

The proof of (3.11) immediately follows by taking $t = \log 2$. □

Denote

$$\delta_0 := D_2\|\Sigma\|\biggl[\mathbf{r}(\Sigma)\biggl(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\biggr) + \mathbf{r}(\Sigma) + 1\biggr].$$

We will define δ_k for $k \geq 1$ as follows:

$$\delta_k = A + B\sqrt{\delta_{k-1}}.$$

It is easy to see that $\delta_1 \leq \delta_0$ (provided that constant D_2 is chosen to be sufficiently large). Note also that

$$\delta_k - \delta_{k+1} = B(\sqrt{\delta_{k-1}} - \sqrt{\delta_k}).$$

Thus, by induction, $\delta_k, k \geq 0$ is a nonincreasing sequence. In view of definition of δ_k, it follows from (3.9) that for all $k \geq 1$

$$\mathbb{P}\bigl\{\delta_{k-1} > \|W\| \geq \delta_k\bigr\} \leq e^{-t}. \tag{3.12}$$

Also, by Lemma 4,

$$\mathbb{P}\big\{\|W\| \geq \delta_0\big\} \leq e^{-t}. \tag{3.13}$$

Let

$$\bar{\delta} = \inf_{k\geq 1} \delta_k = \lim_{k\to\infty} \delta_k.$$

Then

$$\bar{\delta} = A + B\sqrt{\bar{\delta}}.$$

It is easy to check that

$$\bar{\delta} \lesssim \big(A \vee B^2\big). \tag{3.14}$$

In addition,

$$\delta_k - \bar{\delta} = B(\sqrt{\delta_{k-1}} - \sqrt{\bar{\delta}}) \leq B\sqrt{\delta_{k-1} - \bar{\delta}}.$$

Define $u_k, k \geq 0$ as follows: $u_0 = \delta_0$,

$$u_k = B\sqrt{u_{k-1}}.$$

Then

$$\delta_k - \bar{\delta} \leq u_k, \qquad k \geq 0.$$

It is also easy to check that

$$u_k = B^{1+2^{-1}+\cdots+2^{-k+1}} \delta_0^{2^{-k}} = B^2\left(\frac{\delta_0}{B^2}\right)^{2^{-k}}$$

implying

$$0 \leq \delta_k - \bar{\delta} \leq B^2\left(\frac{\delta_0}{B^2}\right)^{2^{-k}}.$$

Let

$$\bar{k} := \min\left\{k : \left(\frac{\delta_0}{B^2}\right)^{2^{-k}} \leq 2\right\}.$$

Clearly,

$$\delta_{\bar{k}} = \bar{\delta} + \delta_{\bar{k}} - \bar{\delta} \leq \bar{\delta} + 2B^2 \lesssim A \vee B^2,$$

where we also used (3.14). Taking into account (3.12) and (3.13), we get that for some constant $D_3 > 0$

$$\mathbb{P}\big\{\|W\| \geq D_3\big(A \vee B^2\big)\big\} \leq \mathbb{P}\big\{\|W\| \geq \delta_{\bar{k}}\big\} \leq (\bar{k}+1)e^{-t}. \tag{3.15}$$

Observe that

$$A \vee B^2 \lesssim M \vee \|\Sigma\|\left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)$$

and also that, for some constant $c_1 > 0$ and for $t \geq 1$

$$\bar{k} \lesssim \log\log \frac{\delta_0}{B^2} \lesssim \log\log\big(c_1 \mathbf{r}(\Sigma)\big) \vee \log\log(c_1 n).$$

Using now (3.15) with $t + \log(\bar{k}+1)$ instead of t, it is easy to get that with probability at least $1 - e^{-t}$

$$\|\hat{\Sigma} - \Sigma\| \lesssim M \vee \|\Sigma\|\Bigg[\sqrt{\frac{t}{n}} \vee \frac{t}{n} \vee \sqrt{\frac{\log^{[3]}(c_1\mathbf{r}(\Sigma))}{n}} \\ \vee \frac{\log^{[3]}(c_1\mathbf{r}(\Sigma))}{n} \vee \sqrt{\frac{\log^{[3]}(c_1 n)}{n}} \vee \frac{\log^{[3]}(c_1 n)}{n}\Bigg],$$

where we used the notation $\log^{[3]} x := \log\log\log x$. In the case when $\mathbf{r}(\Sigma) \lesssim n$, we have

$$\log^{[3]}\big(c_1\mathbf{r}(\Sigma)\big) \leq \log^{[3]}(2c_1 n).$$

Hence, doubling the value of the constant c_1 allows us to drop the two terms involving $\frac{\log^{[3]}(c_1\mathbf{r}(\Sigma))}{n}$. On the other hand, assume that $\mathbf{r}(\Sigma) \geq C'n$ with a sufficiently large constant C' (to be determined later). Observe that $\log^{[3]}(c_1\mathbf{r}(\Sigma)) \lesssim \mathbf{r}(\Sigma)$ and we can use a bound for the median M similar to (2.3):

$$\begin{aligned} M &\geq \mathrm{Med}\big(\|\hat{\Sigma}\|\big) - \|\Sigma\| \geq \mathrm{Med}\left(\sup_{\|u\|\leq 1} n^{-1}\sum_{j=1}^n \langle X_j, u\rangle^2\right) - \|\Sigma\| \\ &\geq \mathrm{Med}\left(\sup_{\|u\|\leq 1} \frac{\langle X_1, u\rangle^2}{n}\right) - \|\Sigma\| \geq \frac{\mathrm{Med}\,\|X\|^2}{n} - \|\Sigma\| \\ &= \frac{(\mathrm{Med}\,\|X\|)^2}{n} - \|\Sigma\| \geq \|\Sigma\|\left(\frac{c'\mathbf{r}(\Sigma)}{n} - 1\right) \geq \frac{c'}{2}\|\Sigma\|\frac{\mathbf{r}(\Sigma)}{n}, \end{aligned} \tag{3.16}$$

for some constants $c' > 0$ and for $C' \geq 2/c'$. We also used the fact that for Gaussian X

$$\mathrm{Med}\,\|X\| \asymp \mathbb{E}\|X\| = \big(\|\Sigma\|\mathbf{r}(\Sigma)\big)^{1/2}.$$

Thus, we get

$$\|\Sigma\|\left(\sqrt{\frac{\log^{[3]}(c_1\mathbf{r}(\Sigma))}{n}} \vee \frac{\log^{[3]}(c_1\mathbf{r}(\Sigma))}{n}\right) \lesssim \|\Sigma\|\frac{\mathbf{r}(\Sigma)}{n} \lesssim M.$$

Since also $\frac{\log^{[3]}(c_1 n)}{n} \lesssim 1$, this implies that with some constant C_1 and with the same probability

$$\begin{aligned} \|W\| &= \|\hat{\Sigma} - \Sigma\| \\ &\le C_1\Bigg[M \vee \|\Sigma\|\Bigg(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \vee \sqrt{\frac{\log^{[3]}(c_1 n)}{n}}\Bigg)\Bigg]. \end{aligned} \tag{3.17}$$

Take now δ to be equal to the expression in the right-hand side of bound (3.17) and use this value of δ to do another iteration of bound (3.9). This easily yields that with some constant $C > 0$ and with probability at least $1 - 2e^{-t}$

$$\|W\| = \|\hat{\Sigma} - \Sigma\| \le C\Bigg[M \vee \|\Sigma\|\Bigg(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\Bigg)\Bigg]. \tag{3.18}$$

To complete the proof of concentration inequality (2.6), note that, for an arbitrary $\delta > 0$, on the event where (3.8) holds and also $\|W\| \le \delta$,

$$\begin{aligned} \|\hat{\Sigma} - \Sigma\| &= g(X_1, \dots, X_n) \\ &\le M_g + D_1\big(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\big)\sqrt{\frac{t}{n}} \\ &\le M + D_1\big(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\big)\sqrt{\frac{t}{n}}. \end{aligned}$$

This bound will be used with

$$\delta := C\Bigg[M \vee \|\Sigma\|\Bigg(\sqrt{\frac{t+2n}{n}} \vee \frac{t+2n}{n}\Bigg)\Bigg]. \tag{3.19}$$

Then, in view of bound (3.18),

$$\mathbb{P}\big\{\|W\| \ge \delta\big\} \le 2e^{-t-2n} = 2e^{-2n-t} \le e^{-2n}$$

(provided that $t \ge 1$). Recall that $g(X_1, \dots, X_n) = \|W\|$ on the event where $\|W\| \le \delta$. Note also that

$$\begin{aligned} \mathbb{P}\big\{g(X_1, \dots, X_n) \ge M\big\} &\ge \mathbb{P}\big\{g(X_1, \dots, X_n) \ge M, \|W\| \le \delta\big\} \\ &\ge \mathbb{P}\big\{\|W\| \ge M, \|W\| \le \delta\big\} \ge 1/2 - \mathbb{P}\big\{\|W\| > \delta\big\} \\ &\ge 1/2 - e^{-2n} \ge 1/4. \end{aligned}$$

Then it follows from Lemma 2 that, for a sufficiently large constant D_1 and for all $t \ge 1$, with probability at least $1 - e^{-t}$, the following bound holds:

$$g(X_1, \dots, X_n) \ge M - D_1\big(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\big)\sqrt{\frac{t}{n}}.$$

Since $\mathbb{P}(\|W\| \le \delta) \ge 1 - 2e^{-t-2n} \ge 1 - e^{-t}$, we have that, with probability at least $1 - 3e^{-t}$,

$$\big|\|\hat{\Sigma} - \Sigma\| - M\big| \le D_1\big(\|\Sigma\| + \|\Sigma\|^{1/2}\sqrt{\delta}\big)\sqrt{\frac{t}{n}}. \tag{3.20}$$

The result now follows by substituting δ given by (3.19) into bound (3.20), doing simple algebra and adjusting the value of constant D_1 to get the probability bound $1 - e^{-t}$. □

Very recent exponential generic chaining bounds for empirical processes by Dirksen [4] (see Corollary 5.7) and by Bednorz [3] (see Theorem 1) imply the following (earlier, Mendelson [12], Theorem 3.1 obtained another version of exponential generic chaining bounds for the same class of processes).

Theorem 8. *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables in a measurable space $(S, \mathcal{A})$ with common distribution P and let $\mathcal{F}$ be a class of measurable functions on $(S, \mathcal{A})$. There exists a constant $C > 0$ such that for all $t \ge 1$ with probability at least $1 - e^{-t}$*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f^2(X_i) - \mathbb{E}f^2(X)\right|$$
$$\le C\max\left\{\sup_{f\in\mathcal{F}}\|f\|_{\psi_2}\frac{\gamma_2(\mathcal{F};\psi_2)}{\sqrt{n}}, \frac{\gamma_2^2(\mathcal{F};\psi_2)}{n}, \sup_{f\in\mathcal{F}}\|f\|_{\psi_2}^2\sqrt{\frac{t}{n}}, \sup_{f\in\mathcal{F}}\|f\|_{\psi_2}^2\frac{t}{n}\right\}.$$

This result together with the argument used in the proof of the upper bound of Theorem 4 easily implies the following generalization of Corollary 2.

Theorem 9. *Let $X, X_1, \ldots, X_n$ be i.i.d. weakly square integrable centered random vectors in E with covariance operator Σ. If X is sub-Gaussian and pre-Gaussian, then there exists a constant $C > 0$ such that, for all $t \ge 1$, with probability at least $1 - e^{-t}$,*

$$\|\hat{\Sigma} - \Sigma\| \le C\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right). \tag{3.21}$$

Note that the proof of concentration inequality of Theorem 6 does not rely on generic chaining bounds, it relies only on the Gaussian isoperimetric inequality. The bound of Theorem 9 (based on the generic chaining method) could be used to provide a shortcut in the proof of the concentration inequality. To this end, instead of using very rough initial bound δ_0 based on Lemma 4 one should use much more precise bound of Theorem 9. In this case, there is no need to implement an iterative argument improving the bound, the concentration inequality in its explicit form (Theorem 5) follows just by an application of the Gaussian isoperimetric inequality. Adamczak [1] suggested an alternative approach to the proof of Theorem 5. It is based on a uniform version of Hanson–Wright inequality and on some other tools (such as Gordon–Chevet inequality), but it does not rely on the generic chaining bounds.

Acknowledgments

The authors are very thankful to Sjoerd Dirksen for attracting their attention to paper [4]. Radek Adamczak pointed out that a similar result was proved in [3].

The authors are especially thankful to Radek Adamczak for providing an alternative proof of the concentration inequality and for very helpful discussions. The initial version of Theorem 5 was under an extra assumption that $\mathbf{r}(\Sigma) \lesssim e^{2n}$. We improved our argument after Adamczak had provided his alternative proof.

Vladimir Koltchinskii is supported in part by NSF Grants DMS-12-07808, CCF-0808863 and CCF-1415498. Karim Lounici is supported in part by Simons Grant 315477 and NSF Career Grant, DMS-14-54515.

References

[1] Adamczak, R. (2014). A note on the Hanson–Wright inequality for random vectors with dependencies. Preprint. Available at arXiv:1409.8457.

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

[3] Bednorz, W. (2014). Concentration via chaining method and its applications. Preprint. Available at arXiv:1405.0676v2.

[4] Dirksen, S. (2015). Tail bounds via generic chaining. *Electron. J. Probab.* **20** no. 53, 29. MR3354613

[5] Klartag, B. and Mendelson, S. (2005). Empirical processes and random projections. *J. Funct. Anal.* **225** 229–245. MR2149924

[6] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Heidelberg: Springer. MR2829871

[7] Kwapień, S. and Szymański, B. (1980). Some remarks on Gaussian measures in Banach spaces. *Probab. Math. Statist.* **1** 59–65. MR0591829

[8] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces. Isoperimetry and Processes*. Berlin: Springer. MR1102015

[9] Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437

[10] Lust-Piquard, F. and Pisier, G. (1991). Noncommutative Khintchine and Paley inequalities. *Ark. Mat.* **29** 241–260. MR1150376

[11] Mendelson, S. (2010). Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.* **20** 988–1027. MR2729283

[12] Mendelson, S. (2012). Oracle inequalities and the isomorphic method. Unpublished manuscript.

[13] Rudelson, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal.* **164** 60–72. MR1694526

[14] Talagrand, M. (2005). *The Generic Chaining*. Berlin: Springer. MR2133757

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

[16] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Y. Eldar and G. Kutyniok, eds.) 210–268. Cambridge: Cambridge Univ. Press. MR2963170

Received August 2014 and revised March 2015