The Annals of Statistics
2020, Vol. 48, No. 3, 1648–1664
https://doi.org/10.1214/19-AOS1862

ROBUST COVARIANCE ESTIMATION UNDER $L_4 - L_2$ NORM EQUIVALENCE

BY SHAHAR MENDELSON[1] AND NIKITA ZHIVOTOVSKIY[2]

[1]*Mathematical Sciences Institute, Australian National University, shahar.mendelson@upmc.fr*
[2]*Higher School of Economics, nikita.zhivotovskiy@phystech.edu*

Let X be a centered random vector taking values in $\mathbb{R}^d$ and let $\Sigma = \mathbb{E}(X \otimes X)$ be its covariance matrix. We show that if X satisfies an $L_4 - L_2$ norm equivalence (sometimes referred to as the bounded kurtosis assumption), there is a covariance estimator $\hat{\Sigma}$ that exhibits almost the same performance one would expect had X been a Gaussian vector. The procedure also improves the current state-of-the-art regarding high probability bounds in the sub-Gaussian case (sharp results were only known in expectation or with constant probability).

In both scenarios the new bounds do not depend explicitly on the dimension d, but rather on the effective rank of the covariance matrix Σ.

1. Introduction. The question of estimating the covariance of a random vector has been studied extensively in recent years (see, e.g., [2, 5, 11–13] and references therein). To formulate the problem, let X be a zero mean random vector taking its values in $\mathbb{R}^d$ and denote the covariance matrix by $\Sigma = \mathbb{E}(X \otimes X)$. Given a sample $X_1, \dots, X_N$ consisting of independent random vectors that are distributed according to X, the goal is to select a matrix $\hat{\Sigma}$ that approximates Σ. While there are various notions of approximation, the focus of this note is on approximation with respect to the $(\ell_2 \to \ell_2)$ operator norm, which from here on is denoted by $\| \ \|$.

One way of viewing the question of covariance estimation (with respect to any norm), is as a vector mean estimation problem. Indeed, if one sets $W = X \otimes X$, then $\mathbb{E}W = \Sigma$, and since one is given a sample $X_1, \dots, X_N$, the vectors $(X_i \otimes X_i)_{i=1}^N$ are N independent copies of W. Thus, a matrix $\hat{W}$ that is a good approximation of the mean $\mathbb{E}W$ with respect to the underlying norm is a solution to the problem of estimating the covariance of X with respect to that norm.

An immediate outcome of this simple observation is that the empirical mean

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N W_i = \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i,$$

which is the trivial choice for estimating the true mean, is a poor estimator unless the random vector W has a "nice" tail behaviour (see, e.g., the discussion in [6]). An example of a positive result of that flavour is Theorem 9 in [2], and to formulate it we need some definitions.

DEFINITION 1.1. The *effective rank* of a positive semidefinite square matrix $A \in \mathbb{R}^{d \times d}$ is given by

$$\mathbf{r}(A) = \frac{\mathrm{Tr}(A)}{\|A\|}. \tag{1}$$

Received October 2018; revised March 2019.
MSC2010 subject classifications. Primary 62G35; secondary 62G15.
Key words and phrases. Covariance estimation, robust estimation, median of means.

Clearly, $\mathbf{r}(A) \leq d$ but the gap between $\mathbf{r}(A)$ and d may be substantial. Recall that the ψ_2-norm of a centred real-valued random variable Y is defined by

$$\|Y\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}\exp(Y^2/c^2) \leq 2\},$$

and that there are absolute constants c and C such that

$$c\|Y\|_{\psi_2} \leq \sup_{p\geq 2} \frac{\|Y\|_{L_p}}{\sqrt{p}} \leq C\|Y\|_{\psi_2}.$$

DEFINITION 1.2. A random vector X with values in $\mathbb{R}^d$ and with the mean μ is L-sub-Gaussian if for every $t \in \mathbb{R}^d$ and every $p \geq 2$,

$$\big(\mathbb{E}|\langle X-\mu, t\rangle|^p\big)^{\frac{1}{p}} \leq L\sqrt{p}\big(\mathbb{E}\langle X-\mu, t\rangle^2\big)^{\frac{1}{2}}. \tag{2}$$

It is standard to verify that a centred random vector is L-sub-Gaussian if and only if its one-dimensional marginals $X_t = \langle X, t\rangle$ satisfy that $\|X_t\|_{\psi_2} \leq cL\|X_t\|_{L_2}$ where c is an absolute constant.

Among the class of L-sub-Gaussian random vectors are vectors whose distribution is multivariate normal (denoted by $\mathcal{N}(\mu, \Sigma)$) and in which case L is an absolute constant. Another simple example are vectors X whose components are independent copies of a zero mean random variable Y that satisfies $\|Y\|_{\psi_2} < \infty$. Indeed, it is standard to show that for such a random vector and any $t \in \mathbb{R}^d$

$$\sup_{p\geq 2} \frac{\big(\mathbb{E}|\langle X, t\rangle|^p\big)^{\frac{1}{p}}}{\sqrt{p}} \leq c\left(\sum_{i=1}^d t_i^2\|Y_i\|_{\psi_2}^2\right)^{\frac{1}{2}} = L\big(\mathbb{E}|\langle X, t\rangle|^2\big)^{1/2},$$

where c is an absolute constant and $L = c\|Y\|_{\psi_2}/\|Y\|_{L_2}$.

REMARK 1.3. Observe that a different notion of sub-Gaussian random vectors sometimes appears in literature: that a centred vector X is called sub-Gaussian if

$$\sup_{t\in S^{d-1}} \sup_{p\geq 2} \frac{\big(\mathbb{E}|\langle X, t\rangle|^p\big)^{\frac{1}{p}}}{\sqrt{p}} = C < \infty, \tag{3}$$

where S^{d-1} is the Euclidean unit sphere in $\mathbb{R}^d$. In other words, according to this notion, a centred random vector is sub-Gaussian if all its one-dimensional marginals have a finite ψ_2 norm, and those norms are all bounded by C. Unlike the notion in Definition 1.2, this does not imply a $\psi_2 - L_2$ norm equivalence of one-dimensional marginals of the random vector. As a result, the constant C in (3) may change dramatically under linear transformations of X, while the factor L in (2) does not.

Throughout this note the notion of a sub-Gaussian random vector that is used is the one from Definition 1.2.

With all the required definitions in place, one may formulate the covariance estimate from [2].

THEOREM 1.4. *For every $L \geq 1$ there exists a constant $c(L)$ for which the following holds. Let X be an L-sub-Gaussian random vector. Then with probability at least $1-\delta$*

$$\left\|\frac{1}{N}\sum_{i=1}^N X_i \otimes X_i - \Sigma\right\| \leq c(L)\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \frac{\mathbf{r}(\Sigma)}{N} + \sqrt{\frac{\log(2/\delta)}{N}} + \frac{\log(2/\delta)}{N}\right). \tag{4}$$

It was also shown in [2] that if G is a zero mean Gaussian vector (and in particular it satisfies the conditions of Theorem 1.4) with covariance Σ then

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} G_i \otimes G_i - \Sigma\right\| \gtrsim \|\Sigma\| \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)}{N}}, \frac{\mathbf{r}(\Sigma)}{N}\right\}.$$

Hence, there is no room for improvement in the deviation estimate of the empirical mean from the true one at the constant confidence level. Of course, that does not imply that the empirical mean is an optimal covariance estimator—even for a Gaussian vector, let alone for a general sub-Gaussian random vector. In fact, as we explain in what follows, there are far better covariance estimators than (4) when the confidence parameter δ is small.

Just as in the one-dimensional mean-estimation problem, once the problem is more "heavy-tailed" the performance of the empirical mean deteriorates quickly and a different procedure has to be used. And that is also the case for covariance estimation. The current state-of-the-art for covariance estimation in heavy-tailed situations is [13] (see Corollary 4.1 there and similar results in [11, 12]), in which X is assumed to satisfy an $L_4 - L_2$ norm equivalence.

DEFINITION 1.5. A random vector X with mean μ satisfies the $L_4 - L_2$ norm equivalence with a constant $L \geq 1$ if for every $t \in \mathbb{R}^d$,

$$\left(\mathbb{E}\langle X - \mu, t\rangle^4\right)^{\frac{1}{4}} \leq L\left(\mathbb{E}\langle X - \mu, t\rangle^2\right)^{\frac{1}{2}}.$$

Note that if X is L-sub-Gaussian then it satisfies an $L_4 - L_2$ norm equivalence with constant $2L$. At the same time, for an $L_4 - L_2$ equivalence the linear forms $\langle X, t\rangle$ need not have higher moments that the fourth one; in particular, X need not be L-sub-Gaussian. Another formulation of the same condition is that for any direction, the *kurtosis*[1] of the corresponding one-dimensional marginal is bounded by L.

REMARK 1.6. In the Appendix one can find two examples that demonstrate the difference between a random vector X being L-sub-Gaussian (which implies an $\psi_2 - L_2$ norm equivalence of the centred marginals of X) and X satisfying an $L_4 - L_2$ norm equivalence.

The current state of the art estimate for random vectors that satisfy Definition 1.5 is as follows:

THEOREM 1.7 ([13]). *For every $L \geq 1$ there are constants $c(L)$ and $c'(L)$ that depend only on L and for which the following holds. Let X satisfy an $L_4 - L_2$ norm equivalence with constant L. For $0 < \delta < 1$ there is an estimator $\tilde{\Sigma}_\delta$ that satisfies*

$$\|\tilde{\Sigma}_\delta - \Sigma\| \leq c(L)\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{N} \cdot \left(\log d + \log(1/\delta)\right)} \tag{5}$$

with probability at least $1 - \delta$, provided that $N \geq c'(L)\mathbf{r}(\Sigma)(\log d + \log(1/\delta))$.

REMARK 1.8. Let us mention that the procedure from [13] requires prior information on the values of $\|\Sigma\|$ and $\mathbf{r}(\Sigma)$ up to some absolute multiplicative constant—an assumption we shall return to in what follows. In fact, a significant part of our analysis is devoted to obtaining estimates on these parameters, and our approach is an alternative to Lepski's method used in [12, 13].

[1] The kurtosis of the random variable Y is equal to $\frac{\mathbb{E}(Y-\mathbb{E}Y)^4}{(\mathbb{E}(Y-\mathbb{E}Y)^2)^2}$.

Observe that if δ is smaller than $1/d$, the error guaranteed by Theorem 1.7 is of the order of

$$\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{N}}\sqrt{\log(1/\delta)}, \tag{6}$$

which turns out to be far from optimal as we now explain.

To put (6) in some perspective, let us examine possible benchmarks for general mean estimation problems and see how those compare with (4), (5) and (6) when applied to covariance estimation.

1.1. *Optimality in mean estimation.* Let W be a random vector with mean μ and set $|||\ |||$ to be an arbitrary norm. Let B° be the unit ball of the dual norm to $|||\ |||$, and denote by $\hat{\mu}$ a mean-estimator constructed using an independent sample $W_1, \dots, W_N$. As it happens, a lower bound on the performance of $\hat{\mu}$ is

$$\frac{R}{\sqrt{N}}\sqrt{\log(1/\delta)}, \tag{7}$$

where

$$R = \sup_{x^* \in B^\circ} \big(\mathbb{E}\big(x^*(W-\mu)\big)^2\big)^{\frac{1}{2}}. \tag{8}$$

Indeed, for every $x^* \in B^\circ$

$$|||\hat{\mu} - \mu||| \geq |x^*(\hat{\mu} - \mu)| = |x^*(\hat{\mu}) - x^*(\mu)|;$$

therefore, if there is a procedure for which $|||\hat{\mu} - \mu||| \leq \varepsilon$ with probability $1-\delta$, then on the same event the procedure automatically performs with accuracy ε and confidence $1-\delta$ for each one of the real-valued mean-estimation problems associated with the random variables $x^*(W)$, $x^* \in B^\circ$. By a lower bound (Proposition 6.1 from [1]) on real-valued mean estimation problems when W is a Gaussian vector, the best possible mean-estimation error for each $x^*(W)$ is

$$\sqrt{\frac{\operatorname{var}(x^*(W))}{N}}\sqrt{\log(1/\delta)},$$

and taking the "worst" $x^* \in B^\circ$ leads to (7).

Although (7) is part of the story, it is unlikely it is the whole story. Intuitively, (7) takes into account the effect of one-dimensional marginals of W rather than the entire geometry of the distribution. It stands to reason that an additional "global" parameter is called for—one that reflects the entire structure of W and the geometry of the norm. Moreover, that parameter should reflect the difficulty of the estimation problem at the constant confidence level.

To give an example of such a result, a (sharp) lower bound from [6] on the mean estimation problem when W is a Gaussian random vector is the following: if $|||\hat{\mu} - \mu||| \leq \varepsilon$ with probability at least $1-\delta$ then

$$\varepsilon \geq \frac{c}{\sqrt{N}}\big(\mathbb{E}|||W-\mu||| + R\sqrt{\log(1/\delta)}\big); \tag{9}$$

hence, the "global parameter" in the Gaussian case is just the mean $\mathbb{E}|||W-\mu|||$.

Let us examine (9) more carefully, in the hope that it would lead us towards the right answer for general random vectors. Note that by setting $\delta = \exp(-p)$, the Gaussian random variable W satisfies that

$$\sqrt{\log(1/\delta)}\big(\mathbb{E}\big(x^*(W-\mu)\big)^2\big)^{\frac{1}{2}} \sim \sqrt{p}\big(\mathbb{E}\big(x^*(W-\mu)\big)^2\big)^{\frac{1}{2}} \sim \big(\mathbb{E}\big|x^*(W-\mu)\big|^p\big)^{\frac{1}{p}}.$$

At the same time, the *strong-weak norm inequality*[2] for Gaussian vectors (see, e.g., [4]) implies that

$$\begin{aligned}
&\left(\mathbb{E}\left|\left|\left|\frac{1}{N}\sum_{i=1}^N W_i-\mu\right|\right|\right|^p\right)^{\frac{1}{p}}\\
&\quad\le \mathbb{E}\left|\left|\left|\frac{1}{N}\sum_{i=1}^N W_i-\mu\right|\right|\right| + c\sup_{x^*\in B^\circ}\left(\mathbb{E}\left|x^*\left(\frac{1}{N}\sum_{i=1}^N W_i-\mu\right)\right|^p\right)^{\frac{1}{p}}\\
&\quad= \frac{1}{\sqrt{N}}\Big(\mathbb{E}|||W-\mu||| + c\sup_{x^*\in B^\circ}\big(\mathbb{E}|x^*(W-\mu)|^p\big)^{\frac{1}{p}}\Big),\\
&\quad= \frac{1}{\sqrt{N}}\Big(\mathbb{E}|||W-\mu||| + c'\sqrt{p}\sup_{x^*\in B^\circ}\big(\mathbb{E}|x^*(W-\mu)|^2\big)^{\frac{1}{2}}\Big),
\end{aligned}$$

where c and c' are absolute constants. Thus, the lower bound of (9) implies that the best possible performance of a mean estimator of a Gaussian vector matches a strong-weak norm inequality. To see that these bounds are of the same order, one needs to use Markov's inequality and optimize with respect to p, where the right choice is indeed $p\sim\log(1/\delta)$.

This leads to a natural conjecture: that the best possible performance in a general mean estimation problem is given by a Gaussian-like strong-weak norm inequality, and that there is a procedure that performs with that accuracy/confidence tradeoff.

Recently, a general mean estimation procedure was introduced in [6] that exhibits this type of a "strong-weak" behaviour. To formulate the result, let W be an arbitrary random vector taking values in $\mathbb{R}^d$ and with mean μ, let G be the zero mean Gaussian random vector with the same covariance as W and set

$$Y_N=\frac{1}{N}\sum_{i=1}^N(W_i-\mu),$$

where $W_1,\dots,W_N$ are independent copies of W. Let $|||\ |||$ be a norm, set B° to be the unit ball of the dual norm, and put

$$R=\sup_{x^*\in B^\circ}\big(\mathbb{E}\big(x^*(W-\mu)\big)^2\big)^{\frac{1}{2}}.$$

THEOREM 1.9 ([6]). *For $0<\delta<1$ there is a procedure $\tilde{\mu}_\delta$ such that*

$$|||\tilde{\mu}_\delta-\mu|||\le c\max\left\{\mathbb{E}|||Y_N|||, \frac{\mathbb{E}|||G|||}{\sqrt{N}}+\frac{R}{\sqrt{N}}\sqrt{\log(1/\delta)}\right\}.$$

The mean estimation procedure is defined as follows: *let $T=\mathrm{ext}(B^\circ)$ to be the set of extreme points in B°.*

- *For the wanted confidence parameter $0<\delta<1$, let $n=\log(1/\delta)$ and set $m=N/n$.*
- *Let $(I_j)_{j=1}^n$ be the natural partition of $\{1,\dots,N\}$ to blocks of cardinality m and given a sample $W_1,\dots,W_N$ set $Z_j=\frac{1}{m}\sum_{i\in I_j}W_j$.*

[2]By "strong norm" we mean the L_1 norm of $|||W-\mu|||$, while the "weak norm" is just the largest L_p norm of a marginal $x^*(W-\mu)$ for $x^*\in B^\circ$.

- *For $x^* \in T$ and $\varepsilon > 0$, set*

$$S_{x^*}(\varepsilon) = \big\{y \in \mathbb{R}^d : \big|x^*(Y) - x^*(Z_j)\big| \le \varepsilon \text{ for more than } n/2 \text{ blocks}\big\},$$

and define

$$S(\varepsilon) = \bigcap_{x^* \in T} S_{x^*}(\varepsilon).$$

- *Set $\varepsilon_0 = \inf\{\varepsilon > 0 : S(\varepsilon) \neq \varnothing\}$, and let $\tilde{\mu}_\delta$ be any vector in $\bigcap_{\varepsilon > \varepsilon_0} S(\varepsilon)$.*

The main result of this note (which is formulated in the next section), is that the right application of Theorem 1.9 leads to an (almost) optimal covariance estimator: the procedure performs as if X were a Gaussian vector even if X only satisfies an $L_4 - L_2$ norm equivalence, and the accuracy/confidence tradeoff obeys the strong-weak inequality one would expect.

1.2. *From mean estimation to covariance estimation.* In what follows, we assume without loss of generality that X is symmetric and zero mean. We may do so because if X' is an independent copy of X then $Z = (X - X')/\sqrt{2}$ is symmetric and has the same covariance as X. It also satisfies an $L_4 - L_2$ norm equivalence if X does. Thus, given a random sample $X_1, \dots, X_N$ sampled independently according to X one may consider the sample

$$\frac{1}{\sqrt{2}}(X_1 - X_2), \dots, \frac{1}{\sqrt{2}}(X_{N-1} - X_N),$$

consisting of $N/2$ independent copies of Z, and perform the procedure with respect to that sample.

The natural choice of a random vector in Theorem 1.9 is $W = X \otimes X$, but as it happens, a better alternative is to use a truncated version of X instead of the original one:

DEFINITION 1.10. Let

$$\beta = \left(\frac{\operatorname{Tr}(\Sigma)\|\Sigma\| N}{\gamma}\right)^{\frac{1}{4}},$$

and let

$$\tilde{X} = X \mathbb{1}_{\{\|X\|_2 \le \beta\}}.$$

In the L-sub-Gaussian case set $\gamma = 1$ and when X only satisfies $L_4 - L_2$ norm equivalence, let $\gamma = \log \mathbf{r}(\Sigma)$. Also denote $\tilde{\Sigma} = \mathbb{E}(\tilde{X} \otimes \tilde{X})$.

DEFINITION 1.11. Given the random vector X taking its values in $\mathbb{R}^d$ define

$$R_X^2 = \sup_{u,v \in S^{d-1}} \mathbb{E}\big(v^T(X \otimes X - \mathbb{E}X \otimes X)u\big)^2. \tag{10}$$

The quantity R_X^2 is sometimes referred to as the *weak variance* of a random matrix.

As was mentioned previously, the main result of this note is the existence of an estimator whose performance improves both (4) and (5) and is an optimal (or very close to being optimal) covariance estimation procedure.

The estimator is constructed in three stages: the first stage leads to a data-dependent estimate on $\operatorname{Tr}(\Sigma)$; the second stage is based on the estimated value of $\operatorname{Tr}(\Sigma)$ established in the

first stage and its outcome is a data-dependent estimate on the value of $\|\Sigma\|$; the last stage receives as input the results of two first stages and the third part of the sample and returns the wanted estimator of Σ. A key point in the analysis of this procedure is that one only needs to estimate $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|$ up to absolute multiplicative constant factors and that simplifies the problem considerably.

The performance of the procedure is summarized in this, our main result.

THEOREM 1.12. *Let X be a zero mean random vector with (an unknown) covariance matrix Σ and let $\|\ \|$ be its operator norm. Using the notation of Definition 1.10 and Definition 1.11, for any $0 < \delta < 1$, there is a procedure that receives as data the sample $X_1, \dots, X_N$, returns a matrix $\hat{\Sigma}_\delta$ and satisfies*:

(1) *If X is L-sub-Gaussian and $N \geq c'(L)(\mathbf{r}(\Sigma) + \log(1/\delta))$, then with probability at least $1 - \delta$,*

$$\|\hat{\Sigma}_\delta - \Sigma\| \leq c(L)\left(\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \frac{R_{\tilde{X}}}{\sqrt{N}}\sqrt{\log(1/\delta)}\right);$$

(2) *If X satisfies an $L_4 - L_2$ norm equivalence and $N \geq c'(L)(\mathbf{r}(\Sigma)\log\mathbf{r}(\Sigma) + \log(1/\delta))$ then with probability at least $1 - \delta$,*

$$\|\hat{\Sigma}_\delta - \Sigma\| \leq c(L)\left(\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)\log(\mathbf{r}(\Sigma))}{N}} + \frac{R_{\tilde{X}}}{\sqrt{N}}\sqrt{\log(1/\delta)}\right). \tag{11}$$

In both cases $R_{\tilde{X}} \leq c(L)\|\Sigma\|$ and $c(L)$, $c'(L)$ are constants that depend only on L.

REMARK 1.13. Note that the estimates in Theorem 1.12 do not depend on the dimension d; instead, they depend only on $\mathbf{r}(\Sigma)$ which may be small even if d tends to infinity. This is important in view of the recent results on covariance estimation in Banach spaces [2].

The estimate in Theorem 1.12 is actually a strong-weak norm inequality—as if X were Gaussian (up to the logarithmic term in (11)). To see that, let G be the zero mean Gaussian random vector that has the same covariance as X and set $N \geq \mathbf{r}(\Sigma)$. As noted previously,

$$\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} \sim \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N G_i \otimes G_i - \Sigma\right\|,$$

with the left-hand side being the "strong term" from Theorem 1.12. Moreover, the term involving R_X is actually the natural weak term associated with the operator norm. Indeed, recall the well-known fact that the dual norm to the operator norm is the nuclear norm. And, since a linear functional z acts on the matrix x via trace duality—that is, $z(x) = [z, x] := \mathrm{Tr}(z^T x)$—it follows, for example, from [15] that the extreme points of the dual unit ball B° are

$$\{u \otimes v : u, v \in S^{d-1}\}.$$

Thus,

$$R_{\tilde{X}}^2 = \sup_{x^* \in B^\circ} \mathbb{E}\big(x^*(\tilde{X} \otimes \tilde{X} - \tilde{\Sigma})\big)^2 = \sup_{u,v \in S^{d-1}} \mathbb{E}\big(v^T(\tilde{X} \otimes \tilde{X} - \tilde{\Sigma})u\big)^2,$$

and in particular, by (7) the weak term $(R_{\tilde{X}}/\sqrt{N})\sqrt{\log(1/\delta)}$ appearing in Theorem 1.12 is sharp.

As a result, and up to the logarithmic factor in (2), Theorem 1.12 implies that the estimator $\hat{\Sigma}_\delta$ performs as if X were Gaussian, even though it can be very far from Gaussian.

Let us compare the outcome of Theorem 1.12 to the current state of the art that was mentioned previously. In the sub-Gaussian setup Theorem 1.12 improves Theorem 1.4 because there are situations in which $R_{\tilde{X}}$ is significantly smaller than $\|\Sigma\|$ (see such an example in what follows). And, under an $L_4 - L_2$ norm equivalence scenario the improvement is more dramatic: on top of an improvement in the logarithmic factor appearing in the "strong" term, the "weak" term, $(R_{\tilde{X}}/\sqrt{N})\sqrt{\log(1/\delta)}$ is significantly smaller than the corresponding estimate of $\|\Sigma\|\sqrt{\mathbf{r}(\Sigma)/N}\sqrt{\log(1/\delta)}$ from Theorem 1.7.

The proof of Theorem 1.12 is presented in the following section.

We end this introduction with some notation. Throughout, absolute constants are denoted by $c, c_1, \ldots, c', \ldots$ and their value may change from line to line. Constants that depend on a parameter L are denoted by $c(L)$, $a \lesssim b$ means that there is an absolute constant c such that $a \le cb$, and $a \sim b$ means that $cb \le a \le c_1 b$. When the constants depend on L we write $a \lesssim_L b$ and $a \sim_L b$ respectively.

2. Proof of Theorem 1.12. Consider the truncated vector $\tilde{X}$ introduced in Definition 1.10 but for now for an arbitrary level of truncation. Let $\alpha \ge 0$ and with a minor abuse of notation, redefine

$$\tilde{X} = X\mathbb{1}_{\{\|X\|_2 \le \alpha\}} \quad \text{and} \quad \tilde{\Sigma} = \mathbb{E}\tilde{X} \otimes \tilde{X}. \tag{12}$$

First, note that by the symmetry of X, $\tilde{X}$ is symmetric as well. Second, for every $p \ge 2$ and any $u \in \mathbb{R}^d$,

$$\|\langle \tilde{X}, u\rangle\|_{L_p} = \big(\mathbb{E}|\langle \tilde{X}, u\rangle|^p\big)^{\frac{1}{p}} \le \big(\mathbb{E}|\langle X, u\rangle|^p\big)^{\frac{1}{p}}.$$

Hence, if X is L-sub-Gaussian then $\|\langle \tilde{X}, u\rangle\|_{L_p} \le L\sqrt{p}\|\langle X, u\rangle\|_{L_2}$, and if X satisfies $L_4 - L_2$ norm equivalence with constant L then $\|\langle \tilde{X}, u\rangle\|_{L_4} \le L\|\langle X, u\rangle\|_{L_2}$.

More important features of $\tilde{X}$ have to do with its covariance matrix $\tilde{\Sigma}$ and trace $\operatorname{Tr}(\Sigma)$:

LEMMA 2.1. *If X is zero mean and satisfies an $L_4 - L_2$ norm equivalence with constant L, then*

$$\|\tilde{\Sigma} - \Sigma\| \le c(L)\frac{\|\Sigma\|\operatorname{Tr}(\Sigma)}{\alpha^2}, \tag{13}$$

and

$$\big|\operatorname{Tr}(\tilde{\Sigma}) - \operatorname{Tr}(\Sigma)\big| \le c(L)\frac{\operatorname{Tr}^2(\Sigma)}{\alpha^2}, \tag{14}$$

where $c(L)$ is a constant that depends only on L.

PROOF. Observe that

$$\begin{aligned}
\|\tilde{\Sigma} - \Sigma\| &= \sup_{u,v \in S^{d-1}} \big|u^T\big(\mathbb{E}(X \otimes X) - \mathbb{E}(\tilde{X} \otimes \tilde{X})\big)v\big| \\
&= \sup_{u,v \in S^{d-1}} \big|\mathbb{E}\langle X, u\rangle\langle X, v\rangle \mathbb{1}_{\{\|X\|_2 > \alpha\}}\big| \\
&\le \sup_{u,v \in S^{d-1}} \big(\mathbb{E}\langle X, u\rangle^4\big)^{\frac{1}{4}} \cdot \big(\mathbb{E}\langle X, v\rangle^4\big)^{\frac{1}{4}} \cdot \Pr^{\frac{1}{2}}\big(\|X\|_2 \ge \alpha\big).
\end{aligned}$$

By the $L_4 - L_2$ norm equivalence,

$$\sup_{u \in S^{d-1}} \big(\mathbb{E}\langle X, u\rangle^4\big)^{\frac{1}{4}} \le L \sup_{u \in S^{d-1}} \big(\mathbb{E}\langle X, u\rangle^2\big)^{\frac{1}{2}} = L\|\Sigma\|^{\frac{1}{2}}$$

and

$$\begin{aligned}\mathbb{E}\|X\|_2^4 &= \mathbb{E}\left(\sum_{i=1}^d \langle X, e_i\rangle^2\right)^2 \le \mathbb{E}\sum_{i,j}\langle X, e_i\rangle^2\langle X, e_j\rangle^2 \\ &\le \sum_{i,j}\big(\mathbb{E}\langle X, e_i\rangle^4\big)^{\frac{1}{2}}\big(\mathbb{E}\langle X, e_j\rangle^4\big)^{\frac{1}{2}} \\ &\le L^2\sum_{i,j}\mathbb{E}\langle X, e_i\rangle^2\cdot\mathbb{E}\langle X, e_j\rangle^2 \\ &= L^2\sum_{i,j}\Sigma_{ii}\Sigma_{jj} = L^2\big(\mathrm{Tr}(\Sigma)\big)^2.\end{aligned} \tag{15}$$

Clearly,

$$\Pr^{\frac{1}{2}}\big(\|X\|_2 \ge \alpha\big) \le \left(\frac{\mathbb{E}\|X\|_2^4}{\alpha^4}\right)^{\frac{1}{2}} \le L\frac{(\mathrm{Tr}(\Sigma))}{\alpha^2} \tag{16}$$

and combining the two observations,

$$\|\tilde{\Sigma} - \Sigma\| \le c(L)\frac{\|\Sigma\|\,\mathrm{Tr}(\Sigma)}{\alpha^2}, \tag{17}$$

as claimed. Turning to the second part of the lemma, note that

$$\mathrm{Tr}(\Sigma) = \sum_{i=1}^d \mathbb{E}\langle X, e_i\rangle^2 \quad \text{and} \quad \mathrm{Tr}(\tilde{\Sigma}) = \sum_{i=1}^d \mathbb{E}\langle X, e_i\rangle^2 \mathbb{1}_{\{\|X\|_2\le\alpha\}}.$$

Therefore, by the $L_4 - L_2$ norm equivalence and (16),

$$\begin{aligned}\big|\mathrm{Tr}(\tilde{\Sigma}) - \mathrm{Tr}(\Sigma)\big| &= \sum_{i=1}^d \mathbb{E}\langle X, e_i\rangle^2 \mathbb{1}_{\{\|X\|_2>\alpha\}} \le \sum_{i=1}^d \mathbb{E}\big(\langle X, e_i\rangle^4\big)^{\frac{1}{2}}\Pr^{\frac{1}{2}}\big(\|X\|_2 > \alpha\big) \\ &\le L^2\left(\sum_{i=1}^d \mathbb{E}\langle X, e_i\rangle^2\right)\Pr^{\frac{1}{2}}\big(\|X\|_2 > \alpha\big) \le c(L)\frac{\mathrm{Tr}^2(\Sigma)}{\alpha^2}.\end{aligned}$$

□

The core component in the estimation procedure is denoted by $\hat{\Sigma}_{\delta,\alpha}$, and its definition for a truncation parameter $\alpha > 0$ is as follows:

The estimator $\hat{\Sigma}_{\delta,\alpha}$

Let $\alpha > 0, 0 < \delta < 1$ and consider the given sample $X_1, \dots, X_N$. Set $\tilde{X}_i = X_i \mathbb{1}_{\{\|X_i\|_2\le\alpha\}}$.

- Let $n = \log(1/\delta)$ and split the sample to n blocks I_j, each one of cardinality $m = N/n$; set $M_j = \frac{1}{m}\sum_{i\in I_j}\tilde{X}_i \otimes \tilde{X}_i$.
- Let $T = \{(u, v) : u, v \in S^{d-1}\}$ and for $\varepsilon > 0$ and a pair (u, v) let

$$S_{u,v}(\varepsilon) = \big\{Y \in \mathbb{R}^{d\times d} : \big|v^T(M_j - Y)u\big| \le \varepsilon \text{ for more than } n/2 \text{ blocks}\big\}.$$

- Set

$$S(\varepsilon) = \bigcap_{(u,v)\in T} S_{u,v}(\varepsilon).$$

- Let $\varepsilon_0 = \inf\{\varepsilon > 0 : S(\varepsilon) \neq \varnothing\}$ and choose $\hat{\Sigma}_{\delta,\alpha}$ to be any matrix that satisfies

$$\hat{\Sigma}_{\delta,\alpha} \in \bigcap_{\varepsilon > \varepsilon_0} S(\varepsilon). \tag{18}$$

While the right truncation level is given in Definition 1.10, namely

$$\beta = \left(\frac{\operatorname{Tr}(\Sigma)\|\Sigma\|N}{\gamma}\right)^{\frac{1}{4}},$$

its definition depends on the identities of $\operatorname{Tr}(\Sigma)$ and $\|\Sigma\|$, which are unknown. To address this issue one first invokes a median-of-means estimator, denoted by $\hat{\varphi}_1$, and show that with high probability,

$$\frac{1}{2}\operatorname{Tr}(\Sigma) \leq \hat{\varphi}_1 \leq 2\operatorname{Tr}(\Sigma).$$

Then $\hat{\Sigma}_{\delta,\alpha}$ is performed on an independent part of the sample and at a truncation level of $\alpha \sim \hat{\varphi}_1$, that is, of the order of $\operatorname{Tr}(\Sigma)$. The outcome in an estimator $\hat{\varphi}_2$ that satisfies

$$\frac{\|\Sigma\|}{2} \leq \hat{\varphi}_2 \leq 2\|\Sigma\|$$

with high probability.

The combination of $\hat{\varphi}_1$ and $\hat{\varphi}_2$ allows one to identify β up to an absolute constant. With that information, $\hat{\Sigma}_{\delta,\alpha}$ is preformed again, this time at the "correct level", resulting in a matrix that is a fine approximation of Σ.

With that in mind, the core of the proof of Theorem 1.12 is the next lemma.

LEMMA 2.2. *Using the notation introduced previously, the following holds for* $\hat{\Sigma}_{\delta,\alpha}$:

(1) *If X is L-sub-Gaussian, then with probability at least* $1-\delta$,

$$\|\hat{\Sigma}_{\delta,\alpha} - \tilde{\Sigma}\| \leq c(L)\left(\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \frac{\mathbf{r}(\Sigma)}{N}\right) + \frac{R_{\tilde{X}}}{\sqrt{N}}\sqrt{\log(1/\delta)}\right).$$

(2) *If X satisfies an $L_4 - L_2$ norm equivalence, $N \geq c'(L)\mathbf{r}(\Sigma)\log\mathbf{r}(\Sigma)$ and*

$$c_1(L)\sqrt{\operatorname{Tr}(\Sigma)} \leq \alpha \leq c_2(L)\left(\frac{\operatorname{Tr}(\Sigma)\|\Sigma\|N}{\log\mathbf{r}(\Sigma)}\right)^{\frac{1}{4}}$$

then with probability at least $1-\delta$,

$$\|\hat{\Sigma}_{\delta,\alpha} - \tilde{\Sigma}\| \leq c(L)\left(\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)\log\mathbf{r}(\Sigma)}{N}} + \frac{R_{\tilde{X}}}{\sqrt{N}}\sqrt{\log(1/\delta)}\right),$$

where $R_{\tilde{X}}$ is as in (10).

In both cases $R_{\tilde{X}} \leq c(L)\|\Sigma\|$ and $c(L), c'(L), c_1(L), c_2(L)$ are constants that depend only on L.

The proof of the lemma is presented in Section 3. Assuming its validity let us complete the proof of Theorem 1.12. From this point on and without the loss of generality, assume that the given sample is of cardinality $3N$, as that only affects the constant factors appearing in the bounds.

Stage 1. *Estimation of* $\mathrm{Tr}(\Sigma)$. The goal is to use the first N observations $X_1, \dots, X_N$ to construct the estimator $\hat{\varphi}_1$, for which, with high probability $\hat{\varphi}_1 \sim \mathrm{Tr}(\Sigma)$. Since

$$\mathrm{Tr}(\Sigma) = \mathbb{E} \sum_{i=1}^{d} \langle X, e_i \rangle^2,$$

a standard median-of-means estimator $\hat{\varphi}_1$ of $\mathbb{E} \sum_{i=1}^{d} \langle X, e_i \rangle^2$ (see [14] for what is by now a standard argument) satisfies that with probability at least $1 - \delta$,

$$\left| \hat{\varphi}_1 - \mathrm{Tr}(\Sigma) \right| \leq c \sqrt{\mathrm{Var}\left(\sum_{i=1}^{d} \langle X, e_i \rangle^2 \right) \frac{\log(1/\delta)}{N}}.$$

Using (15),

$$\mathrm{Var}\left(\sum_{i=1}^{d} \langle X, e_i \rangle^2 \right) \leq (L^2 - 1) \, \mathrm{Tr}(\Sigma)^2,$$

and therefore,

$$\left| \hat{\varphi}_1 - \mathrm{Tr}(\Sigma) \right| \leq c(L) \, \mathrm{Tr}(\Sigma) \sqrt{\frac{\log(1/\delta)}{N}}.$$

Hence, if $N \geq c'(L) \log(1/\delta)$, then with probability at least $1 - \delta$ one has

$$\frac{1}{2} \mathrm{Tr}(\Sigma) \leq \hat{\varphi}_1 \leq 2 \, \mathrm{Tr}(\Sigma). \tag{19}$$

Stage 2. *Estimation of* $\|\Sigma\|$. In this stage, the second part of the sample $X_{N+1}, \dots, X_{2N}$ is utilized, and the procedure receives as an additional input $\hat{\varphi}_1$ that satisfies (19). To ease notation, one may assume that $\mathrm{Tr}(\Sigma)$ is known and set $\alpha = \kappa(L) \sqrt{\mathrm{Tr}(\Sigma)}$, where $\kappa(L)$ is a constant that depends only on L.

Using the notation from (12) and by Lemma 2.1 it follows that

$$\|\tilde{\Sigma} - \Sigma\| \leq c(L) \frac{\|\Sigma\|}{\kappa^2(L)},$$

and

$$\left| \mathrm{Tr}(\tilde{\Sigma}) - \mathrm{Tr}(\Sigma) \right| \leq c(L) \frac{\mathrm{Tr}(\Sigma)}{\kappa^2(L)}.$$

In the L-sub-Gaussian case, invoking Lemma 2.2 and the triangle inequality,

$$\begin{aligned} \|\hat{\Sigma}_{\delta,\alpha} - \Sigma\| &\leq \|\tilde{\Sigma} - \Sigma\| + \|\hat{\Sigma}_{\delta,\alpha} - \tilde{\Sigma}\| \\ &\leq \frac{\|\Sigma\|}{10} + c(L) \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \frac{\mathbf{r}(\Sigma)}{N} + \sqrt{\frac{\log(1/\delta)}{N}} \right) \\ &\leq \frac{\|\Sigma\|}{2}, \end{aligned}$$

provided that $N \geq c'(L)(\mathbf{r}(\Sigma) + \log(1/\delta))$ for a large enough constant $c'(L)$ and $c(L)/\kappa^2(L) \leq \frac{1}{10}$. In that case, setting $\hat{\varphi}_2 = \|\hat{\Sigma}_{\delta,\alpha}\|$, it follows that

$$\frac{\|\Sigma\|}{2} \leq \hat{\varphi}_2 \leq 2 \|\Sigma\|. \tag{20}$$

Finally, in the case of $L_4 - L_2$ norm equivalence, and again by Lemma 2.2, one has that (20) holds as long as $N \geq c'(L)(\mathbf{r}(\Sigma)\log\mathbf{r}(\Sigma) + \log(1/\delta))$. Indeed, since $R_{\tilde{X}} \leq c(L)\|\Sigma\|$, one has

$$\|\hat{\Sigma}_{\delta,\alpha} - \Sigma\| \leq \frac{1}{10}\|\Sigma\| + c(L)\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)\log(\mathbf{r}(\Sigma))}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right) \leq \frac{\|\Sigma\|}{2},$$

as required.

Stage 3. *Estimation of* Σ. The final step uses the sample $X_{2N+1}, \ldots, X_{3N}$. Consider a truncation level β as in Definition 1.10, and which, thanks to the first two stages, can be estimated by $\hat{\beta}$ up to an absolute multiplicative factor. Therefore, to ease notation again, simplicity, assume that β itself is known.

For that choice of truncation parameter consider $\tilde{X}$ and $\tilde{\Sigma}$ as in (12) and let $\hat{\Sigma}_\delta = \hat{\Sigma}_{\delta,\beta}$.

By the triangle inequality,

$$\|\hat{\Sigma}_\delta - \Sigma\| \leq \|\hat{\Sigma}_\delta - \tilde{\Sigma}\| + \|\tilde{\Sigma} - \Sigma\|,$$

and by Lemma 2.1 the quantity $\|\tilde{\Sigma} - \Sigma\|$ is smaller than the wanted accuracy for the chosen level β. The required bound on $\|\hat{\Sigma}_\delta - \Sigma\|$ follows immediately from Lemma 2.2, and Theorem 1.12 follows by taking the union bound over the events analyzed in three stages and combining the conditions on N.

3. Proof of Lemma 2.2. Thanks to Theorem 1.9, the proof of Lemma 2.2 follows once one establishes sufficient control on $\mathbb{E}\|Y_N\|$, $\mathbb{E}\|G\|$ and $R_{\tilde{X}}$.

Controlling $R_{\tilde{X}}$. The required estimate on $R_{\tilde{X}}$ for an arbitrary truncation level α is presented in the next Lemma.

LEMMA 3.1. *Assume that* X *is zero mean and satisfies an* $L_4 - L_2$ *norm equivalence with constant* L. *Setting* $\mathbf{v}^2(X) = \sup_{v\in S^{d-1}} \mathbb{E}\langle X, v\rangle^4$ *one has that*

$$R_{\tilde{X}} \leq \mathbf{v}(X) \lesssim_L \|\Sigma\|.$$

PROOF. For every $u, v \in S^{d-1}$, $\mathbb{E}\langle\tilde{X}, v\rangle\langle\tilde{X}, u\rangle = v^T\tilde{\Sigma}u$; therefore,

$$\mathbb{E}\big(v^T(\tilde{X}\otimes\tilde{X} - \tilde{\Sigma})u\big)^2 = \mathbb{E}\langle\tilde{X}, v\rangle^2\langle\tilde{X}, u\rangle^2 - \big(v^T\tilde{\Sigma}u\big)^2 \leq \mathbb{E}\langle\tilde{X}, v\rangle^2\langle\tilde{X}, u\rangle^2$$

$$\leq \big(\mathbb{E}\langle\tilde{X}, v\rangle^4\big)^{\frac{1}{2}} \cdot \big(\mathbb{E}\langle\tilde{X}, u\rangle^4\big)^{\frac{1}{2}},$$

implying that $R_{\tilde{X}} \leq \mathbf{v}(X)$.

Also, recalling that X satisfies and $L_4 - L_2$ norm equivalence,

$$\mathbb{E}\langle X, v\rangle^4 \leq L^4\big(\mathbb{E}\langle X, v\rangle^2\big)^2 \leq L^4\|\Sigma\|^2$$

implying that $\mathbf{v}(X) \leq L^2\|\Sigma\|$, as claimed. □

Controlling $\mathbb{E}\|G\|$ *and* $\mathbb{E}\|Y_N\|$. In the context of Theorem 1.9, G is the zero mean Gaussian vector on $\mathbb{R}^{d\times d}$ whose covariance coincides with that of $W = \tilde{X}\otimes\tilde{X}$. Instead of dealing with that vector directly, note that

$$\mathbb{E}\|G\| \leq \liminf_{N\to\infty} \sqrt{N}\mathbb{E}\|Y_N\|. \tag{21}$$

Indeed, for every finite set T',

$$\mathbb{E}\|G\| = \sup_{T'\subset B^\circ, T' \text{ is finite}} \mathbb{E}\max_{x^*\in T'} x^*(G),$$

and by the multivariate CLT, for every finite set T', $\{N^{-1/2}\sum_{i=1}^N x^*(W_i - \mathbb{E}W) : x^* \in T'\}$ converges weakly to $\{x^*(G) : x^* \in T'\}$. Hence, (21) follows from tail integration.

Thanks to (21), all that remains is to bound $\mathbb{E}\|Y_N\|$.

The sub-Gaussian case. Fix an integer N and note that

$$\left\| \frac{1}{N}\sum_{i=1}^N \tilde{X}_i \otimes \tilde{X}_i - \tilde{\Sigma} \right\| = \sup_{u\in S^{d-1}} \left| \frac{1}{N}\sum_{i=1}^N \langle \tilde{X}_i, u\rangle^2 - \mathbb{E}\langle \tilde{X}_i, u\rangle^2 \right|, \tag{22}$$

which is the supremum of a quadratic empirical process indexed by S^{d-1}. Such empirical processes have been studied extensively (see, e.g., [7–9]), mainly using chaining methods. As it happens, quadratic *sub-Gaussian* processes may be controlled in terms of a natural metric invariant of the indexing class—the so-called γ_2 *functional*.[3] In the case of (22), the indexing class is S^{d-1} whose elements are viewed as linear functionals on $\mathbb{R}^d$, and the underlying metric is the ψ_2 norm endowed by the random vector $\tilde{X}$. By Corollary 1.9 from [9] it follows that

$$\begin{aligned} &\mathbb{E}\sup_{u\in S^{d-1}} \left| \frac{1}{N}\sum_{i=1}^N \langle \tilde{X}_i, u\rangle^2 - \mathbb{E}\langle \tilde{X}_i, u\rangle^2 \right| \\ &\qquad \le c\left(\mathcal{D}\frac{\gamma_2(S^{d-1}, \psi_2(\tilde{X}))}{\sqrt{N}} + \frac{\gamma_2^2(S^{d-1}, \psi_2(\tilde{X}))}{N} \right), \end{aligned} \tag{23}$$

where c is an absolute constant and

$$\mathcal{D} = \mathcal{D}(S^{d-1}, \psi_2) = \sup_{u\in S^{d-1}} \|\langle \tilde{X}, u\rangle\|_{\psi_2} \sim \sup_{u\in S^{d-1}} \sup_{p\ge 2} \frac{(\mathbb{E}|\langle \tilde{X}, u\rangle|^p)^{\frac{1}{p}}}{\sqrt{p}}.$$

To estimate (23) one requires two facts (see, e.g., [16] for more details). Firstly, a general property of the γ_2 functional is monotonicity in d: if (T, d) is a metric space and d' is another metric on T which satisfies that for every $t_1, t_2 \in T$, $d(t_1, t_2) \le \kappa d'(t_1, t_2)$, then

$$\gamma_2(T, d) \le \kappa \gamma_2(T, d').$$

Here, for every $p \ge 2$ and $u \in \mathbb{R}^d$,

$$\left(\mathbb{E}|\langle \tilde{X}, u\rangle|^p\right)^{\frac{1}{p}} \le \left(\mathbb{E}|\langle X, u\rangle|^p\right)^{\frac{1}{p}} \le L\sqrt{p}\left(\mathbb{E}|\langle X, u\rangle|^2\right)^{\frac{1}{2}},$$

implying that

$$\|\langle \tilde{X}, u\rangle\|_{\psi_2} \le L\|\langle X, u\rangle\|_{L_2};$$

hence, $\gamma_2(S^{d-1}, \psi_2(\tilde{X})) \le L\gamma_2(S^{d-1}, L_2(X))$.

Secondly, by Talagrand's majorizing measures theorem, if G is a zero mean Gaussian random vector with the same covariance as X then

$$\gamma_2\left(S^{d-1}, L_2(X)\right) \le c\mathbb{E}\sup_{u\in S^{d-1}} \langle G, u\rangle \le c\left(\mathbb{E}\|G\|_2^2\right)^{\frac{1}{2}} = c\sqrt{\mathrm{Tr}(\Sigma)},$$

for a some absolute constant c.

Finally, again thanks to the fact that X is L-sub-Gaussian,

$$\mathcal{D} \le L\sup_{u\in S^{d-1}} \|\langle X, u\rangle\|_{L_2} = L\|\Sigma\|^{\frac{1}{2}}.$$

Therefore, by (23), for every N,

$$\mathbb{E}\|Y_N\| \le c(L)\left(\|\Sigma\|^{1/2}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \frac{\mathrm{Tr}(\Sigma)}{N} \right),$$

and in particular, $\liminf_{N\to\infty} \sqrt{N}\mathbb{E}\|Y_N\| \le c(L)\|\Sigma\|^{1/2}\sqrt{\mathrm{Tr}(\Sigma)}$.

This completes the proof of the first part of Lemma 2.2.

[3] Rather than defining the γ_2 functional, we refer the reader to [16] for a detailed exposition on the topic, and to [7–9] for the study of the quadratic empirical process in this and more general situations.

$L_4 - L_2$ norm equivalence. Just as in the sub-Gaussian case, the key issue is finding a suitable estimate on $\mathbb{E}\|Y_N\|$. Thanks to the fact that $\tilde{X}$ is a truncated random vector, one may apply a version of the matrix Bernstein inequality.

We invoke Corollary 7.3.2 from the survey [17] (which is a slightly modified version of the original result from [10]): if Z is a random vector which satisfies that $\|Z \otimes Z\| \le \beta$ almost surely, and $B = \mathbb{E}(Z \otimes Z)^2$, then

$$\mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^{N} Z_i \otimes Z_i - \mathbb{E}(Z \otimes Z) \right\| \le c\left(\sqrt{\frac{\|B\| \log(\mathbf{r}(B))}{N}} + \frac{\beta \log(\mathbf{r}(B))}{N} \right). \tag{24}$$

Here, $Z = X\mathbb{1}_{\{\|X\| \le \alpha\}}$ for α as in Definition 1.10, and all that remains is to estimate $\|B\|$ and $\mathbf{r}(B)$.

It is straightforward to verify that

$$c\|\tilde{\Sigma}\| \operatorname{Tr}(\tilde{\Sigma}) \le \|B\| \le c_1(L)\|\Sigma\| \operatorname{Tr}(\Sigma) \quad \text{and} \quad \operatorname{Tr}(B) \le c_1(L)(\operatorname{Tr}(\Sigma))^2 :$$

the upper estimates on $\|B\|$ and $\operatorname{Tr}(B)$ follow from a direct computation and the fact that X satisfies an $L_4 - L_2$ norm equivalence (see, e.g., Lemma 4.1 in [13]); the lower estimate is an outcome of the FKG inequality (see Corollary 5.1 in the supplementary material to [12]).

Turning to the upper bound on $\mathbf{r}(B)$, by Lemma 2.1 and using its notation, both $\|\tilde{\Sigma}\|$ and $\operatorname{Tr}(\tilde{\Sigma})$ are equivalent up to multiplicative constant factors to $\|\Sigma\|$ and $\operatorname{Tr}(\Sigma)$ respectively, as long as $\alpha \ge c_2(L)\sqrt{\operatorname{Tr}(\Sigma)}$; hence, $\mathbf{r}(B) \lesssim_L \mathbf{r}(\Sigma)$.

Finally, observe that $\|Z \otimes Z\| = \|Z\|_2^2 \le \alpha^2$. By (24) and the fact that $N \gtrsim_L \mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)$,

$$\begin{aligned} \mathbb{E}\|Y_N\| &\le c(L)\left(\|\Sigma\|^{1/2} \sqrt{\frac{\operatorname{Tr}(\Sigma)\log \mathbf{r}(\Sigma)}{N}} + \alpha^2 \frac{\log \mathbf{r}(\Sigma)}{N} \right) \\ &= c(L)\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)}{N}} + \frac{\alpha^2 \mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)}{\operatorname{Tr}(\Sigma)N} \right). \end{aligned} \tag{25}$$

In particular,

$$\liminf_{N\to\infty} \sqrt{N}\mathbb{E}\|Y_N\| \le c'(L)\|\Sigma\| \sqrt{\mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)},$$

provided that $\alpha = \alpha(N)$ satisfies

$$\liminf_{N\to\infty} \frac{\alpha^2 \mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)}{\operatorname{Tr}(\Sigma)\sqrt{N}} \le c''(L)\sqrt{\mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)}.$$

That is the case if

$$\alpha \le c_2(L)\left(\frac{\operatorname{Tr}(\Sigma)\|\Sigma\|N}{\log \mathbf{r}(\Sigma)} \right)^{\frac{1}{4}}. \tag{26}$$

Finally, observe that when (26) holds,

$$\frac{\alpha^2 \mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)}{\operatorname{Tr}(\Sigma)N} \le c_2^2(L)\sqrt{\frac{\mathbf{r}(\Sigma)\log \mathbf{r}(\Sigma)}{N}},$$

and combined with (25) this completes the proof of second part of the lemma. □

Concluding remarks. We start this section with an alternative way of estimating $\|\Sigma\|$ that does not require the knowledge of either $\operatorname{Tr}(\Sigma)$ or L, and does not have the extra factor $\log \mathbf{r}(\Sigma)$ appearing in the condition on N when X satisfies an $L_4 - L_2$ norm equivalence. The drawback of this approach is that the bound depends on the dimension d, rather than on $\mathbf{r}(\Sigma)$.

Sketch of the argument. Let $\mathcal{N}$ be a minimal $1/4$ cover of S^{d-1} with respect to the Euclidean norm. Thus, $\|\Sigma\| \sim \sup_{u\in\mathcal{N}} u^T\Sigma u$. For any fixed u, the median of means estimator $\hat{\varphi}_{2,u}$ of $\mathbb{E}u^T X \otimes Xu$ satisfies that with probability at least $1-\delta$,

$$\left|\hat{\varphi}_{2,u} - u^T\Sigma u\right| \le c(L)\|\Sigma\|\sqrt{\frac{\log(1/\delta)}{N}},$$

because $\mathrm{Var}(u^T X \otimes Xu) \le L^4\|\Sigma\|^2$. Finally, recalling that $|\mathcal{N}| \le 9^d$, the union bound shows that with probability at least $1-\delta$

$$\sup_{u\in\mathcal{N}}\left|\hat{\varphi}_{2,u} - u^T\Sigma u\right| \le c_1(L)\|\Sigma\|\sqrt{\frac{d+\log(1/\delta)}{N}}.$$

Therefore, when $N \ge c_1'(L)(d+\log(1/\delta))$, one has that $\sup_{u\in\mathcal{N}}\hat{\varphi}_{2,u} \sim \|\Sigma\|$ with probability at least $1-\delta$.

We end this note with an example showing that there could be a substantial gap between R_X and $\|\Sigma\|$ (and in a similar way, between R_X and $\mathbf{v}(X)$), which is a reason for the suboptimality of Theorem 1.4 (Theorem 9 in [2]).

EXAMPLE 3.2. Let $(\varepsilon_i)_{i=1}^d$ be independent, symmetric, $\{-1,1\}$-valued random variables, and set $\alpha_1 > \cdots > \alpha_d \ge 0$. Let $X^{(i)} = \alpha_i\varepsilon_i$ and consider $X = (X^{(1)},\dots,X^{(d)})$. Since the $X^{(i)}$'s are centered, independent and sub-Gaussian with a constant sub-Gaussian parameter, then X is a centered, L-sub-Gaussian random vector for some absolute constant L.

Let $\Sigma = \mathbb{E}(X\otimes X)$ and note that $\|\Sigma\| = \alpha_1^2$, $\mathbf{r}(\Sigma) = \sum_{i=1}^d \alpha_i^2/\alpha_1^2$ and

$$\begin{aligned}
\mathbb{E}\big(v^T(X\otimes X-\Sigma)u\big)^2 &= \mathbb{E}\Big(\sum_{i\neq j} v_iu_jX^{(i)}X^{(j)}\Big)^2\\
&= \sum_{i\neq j}\alpha_i^2\alpha_j^2\big(v_i^2u_j^2 + v_iv_ju_iu_j\big)\\
&\le (\alpha_1\alpha_2)^2\Big(\sum_{i,j}(v_iu_j)^2 + |v_iv_ju_iu_j|\Big)\\
&\le (\alpha_1\alpha_2)^2\big(\|v\|^2\|u\|^2 + \langle|v|,|u|\rangle^2\big) \le 2(\alpha_1\alpha_2)^2.
\end{aligned}$$

Hence,

$$R_X \le \sqrt{2}\alpha_1\alpha_2 \le \alpha_1^2 = \|\Sigma\|, \tag{27}$$

and the gap between R_X and $\|\Sigma\|$ may be arbitrary large.

Inequality (27) is the best one can hope for in general. Indeed, let Y be a centered random vector taking its values in $\mathbb{R}^d$, set $\Sigma = \mathbb{E}(Y\otimes Y)$ and consider R_Y. It follows that

$$\begin{aligned}
\big\|\mathbb{E}(Y\otimes Y-\Sigma)^2\big\| &= \Big\|\mathbb{E}(Y\otimes Y-\Sigma)\sum_{i=1}^d e_ie_i^T(Y\otimes Y-\Sigma)\Big\|\\
&\le \sum_{i=1}^d\sup_{v\in S^{d-1}}\mathbb{E}\big(e_i^T(Y\otimes Y-\Sigma)v\big)^2 \le dR_Y^2.
\end{aligned}$$

As before, Corollary 5.1 in [12] implies that $\|\mathbb{E}(Y\otimes Y)^2\| \ge \mathrm{Tr}(\Sigma)\|\Sigma\|$. Therefore,

$$dR_Y^2 \ge \big\|\mathbb{E}(Y\otimes Y-\Sigma)^2\big\| \ge \big\|\mathbb{E}(Y\otimes Y)^2\big\| - \big\|\Sigma^2\big\| \ge \big(\mathrm{Tr}(\Sigma)\big)\|\Sigma\| - \|\Sigma\|^2$$

and

$$R_Y \geq \sqrt{\frac{\mathbf{r}(\Sigma)-1}{d}}\|\Sigma\|, \tag{28}$$

which is optimal when $\mathbf{r}(\Sigma) \sim d$.

APPENDIX: SUB-GAUSSIAN VS. NORM EQUIVALENCE

The first example we present is the class of *L-subexponential* random vectors. These vectors satisfy

$$\big(\mathbb{E}|\langle X-\mu, t\rangle|^p\big)^{\frac{1}{p}} \leq Lp\big(\mathbb{E}\langle X-\mu, t\rangle^2\big)^{\frac{1}{2}}$$

for every $p \geq 2$; in particular, X satisfies an $L_4 - L_2$ norm equivalence with constant $4L$. On the other hand, there are obvious examples in which some marginals of X need not be sub-Gaussian. For example, if X has independent components that are distributed according to an exponential random variable y, then for every $1 \leq i \leq d$, $\|\langle X, e_i\rangle\|_{\psi_2} = \|y\|_{\psi_2} = \infty$.

Another simple example are of random vectors with a multivariate t-distribution,[4] which, in some cases, satisfy an $L_4 - L_2$ norm equivalence but are not L-sub-Gaussian for any L. The bad sub-Gaussian behaviour is an immediate consequence of the observation that when $d = 1$ and the random variable has ν degrees of freedom, its νth moment does not exist.

EXAMPLE A.1. Assume that Z has a multivariate normal distribution $\mathcal{N}(0, \Sigma')$ and V is a random variable independent of Z that has a χ^2_ν distribution for some $\nu \geq 1$. Consider the random vector $X = \frac{Z}{\sqrt{V/\nu}}$, which is centred and has a multivariate t-distribution with parameters (ν, Σ'). Fix $t \in \mathbb{R}^d \setminus \{0\}$ and consider the random variable $\langle X, t\rangle = \frac{\langle Z,t\rangle}{\sqrt{V/\nu}}$. Observe that $\langle Z, t\rangle$ is normal with mean zero and variance $t^T\Sigma' t$ and is independent of V, and therefore has a t distribution with ν degrees of freedom. A straightforward calculation shows that its kurtosis is $\frac{3\nu-6}{\nu-4}$ for $\nu > 4$ [3]. Hence, X satisfies an $L_4 - L_2$ norm equivalence with $L = (\frac{3\nu-6}{\nu-4})^{\frac{1}{4}}$ provided that $\nu > 4$, but clearly X is not sub-Gaussian.

Acknowledgment. The second author was supported by RSF grant No. 18-11-00132.

REFERENCES

[1] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

[2] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

[3] KOTZ, S. and NADARAJAH, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge Univ. Press, Cambridge. MR2038227 https://doi.org/10.1017/CBO9780511550683

[4] LATAŁA, R. and WOJTASZCZYK, J. O. (2008). On the infimum convolution inequality. *Studia Math.* **189** 147–187. MR2449135 https://doi.org/10.4064/sm189-2-5

[5] LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437 https://doi.org/10.3150/12-BEJ487

[6] LUGOSI, G. and MENDELSON, S. (2019). Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields*. To appear.

[7] MENDELSON, S. (2010). Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.* **20** 988–1027. MR2729283 https://doi.org/10.1007/s00039-010-0084-5

[8] MENDELSON, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.* **126** 3652–3680. MR3565471 https://doi.org/10.1016/j.spa.2016.04.019

[4] See, for example, [3] for an extensive survey on multivariate t-distributions and their properties.

[9] MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.* **17** 1248–1282. MR2373017 https://doi.org/10.1007/s00039-007-0618-7

[10] MINSKER, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statist. Probab. Lett.* **127** 111–119. MR3648301 https://doi.org/10.1016/j.spl.2017.03.020

[11] MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

[12] MINSKER, S. and WEI, X. (2017). Estimation of the covariance structure of heavy-tailed distributions. In *NIPS*.

[13] MINSKER, S. and WEI, X. (2018). Robust modifications of U-statistics and applications to covariance estimation problems. Preprint. Available at arXiv:1801.05565.

[14] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication*. Wiley, New York. MR0702836

[15] SO, W. (1990). Facial structures of Schatten p-norms. *Linear Multilinear Algebra* **27** 207–212. MR1064897 https://doi.org/10.1080/03081089008818012

[16] TALAGRAND, M. (2014). *Upper and Lower Bounds for Stochastic Processes*: *Modern Methods and Classical Problems*. **60**. Springer, Heidelberg. MR3184689 https://doi.org/10.1007/978-3-642-54075-2

[17] TROPP, J. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* **8** 1–230.