# Covariance estimation under missing observations and L4 − L2 moment equivalence

Electronic Journal of Statistics
Vol. 18 (2024) 2665–2686
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2264

# Covariance estimation under missing observations and $L_4 - L_2$ moment equivalence

**Pedro Abdalla**

*Department of Mathematics, ETH Zürich,*
*e-mail:* pedro.abdallateixeira@ifor.math.ethz.ch

**Abstract:** We consider the problem of estimating the covariance matrix of a random vector by observing i.i.d samples, where entry of the sampled vector is missing with probability $p$. Under the standard $L_4 - L_2$ moment equivalence assumption, we construct the first estimator that simultaneously achieves optimality with respect to the parameter $p$ and recovers the optimal convergence rate for the classical covariance estimation problem when $p = 1$.

**MSC2020 subject classifications:** Primary 62A99.
**Keywords and phrases:** Covariance estimation, missing observations, heavy tails.

Received June 2023.

## Contents



## 1. Introduction

High-dimensional covariance estimation is one of the most fundamental problems in the intersection of probability and statistics. On the applied side, it is a fundamental task for PCA or linear regression [31]. On the theoretical side, the non-asymptotic properties of isotropic sample covariance matrices have been extensively studied [2, 20, 30, 29, 28, 10] due to a famous question by Kannan, Lovász and Simonovits [11] and further generalized to the anisotropic case [16, 14, 1]. Although the sample covariance matrix seems to be the most natural choice of estimator, its performance is suboptimal when the input data lacks a

strong decay in the tail. Specifically, the convergence rate with respect to the confidence level $\delta$ is quite slow.

Motivated by this fact, a line of work in robust statistics, pioneered by Catoni [5], studied the so-called sub-Gaussian estimators. These estimators are defined to be estimators that perform as good as the empirical mean under the Gaussian distribution. Many estimators have been proposed for the covariance estimation problem (see [12] for a survey), in particular, there are now sub-Gaussian estimators under minimal assumptions on the data distribution [1, 23].

On the other hand, data may be corrupted by noise. In [18], Lounici addressed the so-called covariance estimation problem with *missing observations*, motivated by applications in climate change, gene expression and cosmology. His work considers i.i.d observations, where each entry is "missing" with probability $p$. We highlight that the missing observations model is a standard notion in the literature, extending beyond the covariance estimation setting, see [9, 17] and the references therein.

The goal of this work is to design an estimator that simultaneously achieves the following properties:

- **Missing Observations:** We allow the data to have missing observations and heavy tails. We construct an estimator with minimax optimal convergence rate without assuming any knowledge of $p$. Remarkably, we show that dependence on $p$ is universal, meaning that it does not depend on the distribution of the data.
- **Dimension-Free:** The convergence rate scales with the effective rank $\mathbf{r}(\Sigma)$ rather than the dimension $d$,

$$\mathbf{r}(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|}.$$

  This is an important aspect in high dimensional settings when the dimension $d$ is at least the sample size $N$.
- **Heavy-Tails:** We allow the distribution to have heavy tails, requiring only the existence of four moments satisfying minimal assumptions. Moreover, the result is as sharp as if the data were Gaussian (up to an absolute constant).

We begin with the rigorous definition of the model. We say that a centred random vector $X$ satisfies the $L_4 - L_2$ *moment equivalence* (*hypercontractivity*) with constant $\kappa \geq 1$, if for all $v \in S^{d-1}$,

$$(\mathbb{E}\langle X, v\rangle^4)^{1/4} \leq \kappa(\mathbb{E}\langle X, v\rangle^2)^{1/2}.$$

Here we always assume that the data satisfies the $L_4 - L_2$ moment equivalence with an absolute constant $\kappa > 0$, i.e, the constant $\kappa$ is a fixed real number that does not depend on any other parameter. A vast class of distributions satisfies the moment equivalence assumption mentioned above, with $\kappa$ being a small absolute constant. Examples include sub-gaussian random vectors, sub-exponential

random vectors with bounded $\psi_\alpha$ norm, as well $t$-student distributions with a sufficiently large degree of freedom [21].

We say that the sample $Y_1, \dots, Y_N$ is *$p$-sparsified* if it is obtained from the sample $X_1, \dots X_N$ of independent copies of $X$ by multiplying each entry of the $X_i$'s by an independent $0/1$ Bernoulli random variable with mean $p$. In concise terminology, we say that the data is sampled from $X \odot \mathbf{p}$, where $\mathbf{p} \in \{0,1\}^d$ is a random vector with i.i.d entries Bernoulli $p$, and the notation $\odot$ simply denotes the standard entrywise product. The choice of zero to represent missing information is merely for convenience and could be replaced by any other value. Now, we present the main result of this manuscript.

**Theorem 1** (Main result)**.** *Assume that $X$ is a zero mean random vector in $\mathbb{R}^d$ with covariance matrix $\Sigma$ satisfying the $L_4 - L_2$ moment equivalence assumption with an absolute constant $\kappa$. Fix the confidence level $\delta \in (0,1)$. Suppose that $Y_1, \dots, Y_N$ are i.i.d samples distributed as $X \odot p$, where $\mathbf{p} = (p_1, \dots, p_d) \in \mathbb{R}^d$ is a random vector with i.i.d Bernoulli entries with parameter $p$. Then there exists an estimator $\widehat{\Sigma}(N, \delta)$ depending only on the sample $Y_1, \dots, Y_N$ and $\delta$ satisfying that, with probability at least $1 - \delta$,*

$$\|\widehat{\Sigma} - \Sigma\| \leq \frac{C(\kappa)}{p}\|\Sigma\| \left( \sqrt{\frac{r(\Sigma) + \log(1/\delta)}{N}} \right).$$

*Here $C(\kappa) > 0$ is an absolute constant depending only on $\kappa$.*

**Literature review:** We remark that several results for covariance estimation under missing observations were obtained in the literature, for example [18, 27, 26, 25, 13, 3]. However, none of the previous results has been able to simultaneously scale correctly with the factor of $p$ and recover a sub-Gaussian estimator when $p = 1$, as established in [1, 23], even when the data is Gaussian. Moreover, the convergence rate is optimal up to an absolute constant: When $p = 1$, a classical result by Lounici and Koltchinskii [14, Theorem 4] states that if $G_1, \dots, G_N$ are i.i.d mean zero Gaussian vectors with covariance matrix $\Sigma$ and $N \geq r(\Sigma)$, then

$$c\|\Sigma\|\sqrt{\frac{r(\Sigma)}{N}} \leq \mathbb{E}\left\| \frac{1}{N}\sum_{i=1}^N G_i \otimes G_i - \Sigma \right\| \leq C\|\Sigma\|\sqrt{\frac{r(\Sigma)}{N}}.$$

This essentially shows optimality with respect to the effective rank, as we expect that the empirical covariance for Gaussian distributions to be sharp in expectation. They also showed that that the expectation is tightly concentrated around the mean, and our quantitative convergence rate with respect to $\delta$ matches their result up to an absolute constant. Both are indeed optimal among all (measurable) estimators for the covariance; see [21, 1] for a more technical discussion. Intuitively, it should be not surprising that we cannot beat the Gaussian decay.

In addition to this, the dependence with respect to $p$ is also optimal thanks to a minimax lower bound from Lounici [18, Theorem 2]. In a nutshell, his result

shows that there exist absolute constants $c_1, c_2 > 0$ for which

$$\inf_{\widehat{\Sigma}} \sup_{\mathbb{P}} \mathbb{P}\left( \|\widehat{\Sigma} - \Sigma\| \geq \frac{c_1}{p} \|\Sigma\| \sqrt{\frac{r(\Sigma)}{N}} \right) \geq c_2.$$

Here, the infimum is taken with respect to all estimators that depend only on the data, and the supremum is taken over all possible distributions with covariance matrix $\Sigma$. This implies that our main result captures the optimal dependence with respect to $p$ as well.

It is important to note that the main drawback of our result is that our estimator is not computationally tractable. The primary focus of this work is on the information-theoretic limits of covariance estimators, specifically our main contribution is demonstrating the possibility of constructing an optimal data-driven estimator for covariance under minimal assumptions on the data (albeit it is computationally infeasible).

To the best of our knowledge, there are no computable estimators for the covariance matrix under heavy tails, even in the case without missing observations. We leave it as an important open problem.

**Proposed estimator:** The startpoint to construct our estimator is the following observation: The expectation of the covariance matrix of $Y$ scales differently for the diagonal part and the off-diagonal part of its covariance matrix. More accurately,

$$\mathbb{E}Y \otimes Y = p \operatorname{Diag}(\Sigma) + p^2 \operatorname{Off}(\Sigma).$$

We can "invert" the equality above to get the dependence between the true covariance and the data, namely

$$\Sigma = p^{-1} \operatorname{Diag}(\mathbb{E}Y \otimes Y) + p^{-2} \operatorname{Off}(\mathbb{E}Y \otimes Y).$$

A natural approach would be to replace the unknown term $\mathbb{E}Y \otimes Y$ by its sample covariance, but this is not enough when we consider heavy tailed data $X_1, \ldots, X_N$, as discussed above. In fact, we define the truncation function

$$\psi(x) = \begin{cases} x, & \text{for } x \in [-1, 1], \\ \operatorname{sign}(x), & \text{for } |x| > 1, \end{cases} \tag{1.1}$$

to robustify our estimator in each direction of the sphere. The idea here is to estimate the matrix through its quadratic form. Next, we describe the estimator's final form. We estimate the diagonal and off-diagonal part separately,

$$\widehat{\Sigma_1}(\lambda_1) := \operatorname*{argmin}_{\Sigma_1 \in \mathbb{S}_+^d \,|\, \operatorname{Off}(\Sigma_1) = 0} \sup_{v \in S^{d-1}} |v^T \Sigma_1 v - \frac{1}{n\lambda_1} \sum_{i=1}^n \psi(\lambda_1 v^T \operatorname{Diag}(Y_i \otimes Y_i) v)|,$$

$$\widehat{\Sigma_2}(\lambda_2) := \operatorname*{argmin}_{\Sigma_2 \in \mathbb{S}_+^d \,|\, \operatorname{Diag}(\Sigma_2) = 0} \sup_{v \in S^{d-1}} |v^T \Sigma_2 v - \frac{1}{n\lambda_2} \sum_{i=1}^n \psi(\lambda_2 v^T \operatorname{Off}(Y_i \otimes Y_i) v)|,$$

where $\mathbb{S}_+^d$ is the set of $d$ by $d$ positive semi-definite matrices and $S^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$, and the final estimator becomes

$$\widehat{\Sigma} = \frac{1}{\widehat{p}} \operatorname{Diag}(\widehat{\Sigma_1}) + \frac{1}{\widehat{p}^2} \operatorname{Off}(\widehat{\Sigma_2}).$$

Here, $\widehat{p}$ is an estimator for the parameter $p$, and the choice of the truncation levels $\lambda_1, \lambda_2$ will be clarified in what follows.

As mentioned before, the main drawback of our estimation is that it is not computational tractable. Indeed, $\widehat{\Sigma}_1, \widehat{\Sigma}_2$ does not seem to be computable in polynomial time as a (sub)-gradient descent/ascent type method might get stuck in local optima, and analyzing it is out of the scope of this text. We remark that a similar optimization problem appears in [15], with the quadratic forms replaced by linear forms. Unfortunately, even in that case it is an open problem to come up with an (analyzable) algorithm.

From a more practical perspective, it might be possible that, under some stronger concentration assumption on the data, we can avoid evaluating the truncation function $\psi$ in each direction $v$ on the Euclidean sphere, and replace the supremum over the Euclidean sphere in the definition of $\widehat{\Sigma}_1, \widehat{\Sigma}_2$ by other (more tractable) quantity. This would make the optimization problem much easier to be solved in polynomial time. We leave this as an interesting question to pursue in a future work.

The construction of the estimator and its analysis share similarities with the "trimmead covariance" estimator proposed by Zhivotovskiy and the author [1]. However, we need to break into diagonal and off-diagonal parts to take in account the different scales with $p$. Indeed, the main technical difficulty arises in controlling the random quadratic form to get the optimal dependence with respect to $p$, mainly in the off-diagonal case. A direct approach faces the difficulty that we no longer have a positive semidefinite matrix, making it challenging to capture cancellations. Conversely, an indirect approach, expressing the off-diagonal part as the total part minus the diagonal part, leads to sub-optimality with respect to $p$. Thus, we need to carefully balance these two approaches.

**Organization** The rest of the paper is organized as follows: In Section 2, we assume the knowledge of certain parameters to simplify the analysis of the estimator and derive sharp convergence rates. We then systematically relax these assumptions in Section 3 by estimating each parameter separately in individual subsections. The last subsection of Section 3 is devoted to the formal construction of the estimator and the proof of the main result.

**Notation** Throughout this text $C, c > 0$ denote an absolute constant that may change from line to line. For an integer $N$, we set $[N] = \{1, \dots, N\}$. For any two functions (or random variables) $f, g$ defined in some common domain, the notation $f \lesssim g$ means that there is an absolute constant $c$ such that $f \leq cg$ and $f \sim g$ means that $f \lesssim g$ and $g \lesssim f$. Let $\mathbb{S}_+^d$ denote the set of $d$ by $d$ positive-definite matrices. The symbols $\|\cdot\|, \|\cdot\|_F$ denote the operator norm

and the Frobenius norm of a matrix, respectively. Let $\mathcal{KL}(\rho,\mu)=\int \log\left(\frac{d\rho}{d\mu}\right)d\rho$ denote the Kullback-Leibler divergence between a pair of measures $\rho$ and $\mu$. We write $\rho \ll \mu$ to indicate that the measure $\rho$ is absolutely continuous with respect to the measure $\mu$. For a vector $X\in\mathbb{R}^d$, the tensor product $\otimes$ is defined as $X\otimes X := XX^T \in \mathbb{R}^{d\times d}$.

## 2. Oracle estimator

In this section, we prove our main result under the assumption that we know the effective rank of the covariance matrix $r(\Sigma)$, the trace of the covariance matrix $\mathrm{Tr}(\Sigma)$, and the sparsifying factor $p$. These assumptions will be further relaxed in the next section. Our main goal is to prove the following result.

**Proposition 1.** *Assume that $X$ is a mean zero random vector in $\mathbb{R}^d$ with covariance matrix $\Sigma$ satisfying the $L_4-L_2$ moment equivalence assumption. Fix the confidence level $\delta\in(0,1)$. Suppose that $Y_1,\dots,Y_N$ are i.i.d samples from $X\odot\mathbf{p}$. Then there exists $\lambda_1,\lambda_2>0$ depending only on $\mathrm{Tr}(\Sigma),\|\Sigma\|$ and $p$ for which, with probability at least $1-\delta$,*

$$\begin{aligned}&\max\{\|p^{-1}\widehat{\Sigma_1}(\lambda_1)-\mathrm{Diag}(\Sigma)\|,\|p^{-2}\widehat{\Sigma_2}(\lambda_2)-\mathrm{Off}(\Sigma)\|\}\\&\le\frac{C(\kappa)}{p}\|\Sigma\|\left(\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}\right).\end{aligned}$$

*Here $C(\kappa)>0$ is an absolute constant depending only on $\kappa$.*

Our analysis is based on the variational principle pioneered by O. Catoni [5, 4, 8] and further developed in many applications related to high dimensional probability and statistics [4, 6, 7, 8, 32, 22]. In most of the applications of the variational principle, the following lemma serves as a key stepping stone.

**Lemma 1.** *Assume that $X_i$ are i.i.d. random variables defined on some measurable space. Let $\Theta$ be a subset of $\mathbb{R}^p$ for some $p\ge1$, $\mu$ be a a fixed distribution on $\Theta$, and $\rho$ be any distribution on $\Theta$ satisfying that $\rho\ll\mu$. Then, simultaneously for any such $\rho$, with probability at least $1-\delta$,*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\rho f(X_i,\theta)\le\mathbb{E}_\rho\log(\mathbb{E}_X e^{f(X,\theta)})+\frac{\mathcal{KL}(\rho,\mu)+\log(1/\delta)}{N}.$$

*Here $\theta$ is distributed according to $\rho$.*

The proof can be found in [4, 32] and will be omitted. The next lemma is a technical fact that allow to "convexify" the truncation function $\psi$. Indeed, it is easy to see that the function $e^{\psi}(x)$ is bounded by $(1+x+x^2)$ that still not convex, but if we add a suitable quadratic term, then it becomes convex.

**Lemma 2.** *Let $\psi$ be the truncation function from (1.1), and let $Z$ be a random variable with finite second moment. Then the following holds*

$$\psi(\mathbb{E}Z) \leq \mathbb{E}\log(1+Z+Z^2) + \min\left\{1, \frac{\mathbb{E}Z^2}{6}\right\}.$$

*Moreover, for any $a > 0$,*

$$\mathbb{E}\log(1+Z+Z^2) + a\min\left\{1, \frac{\mathbb{E}Z^2}{6}\right\} \leq \mathbb{E}\log\left(1+Z+\left(1+\frac{(7+\sqrt{6})(\exp(a)-1)}{6}\right)Z^2\right).$$

This result has been previously used in [4, 32, 1]. For the sake of completeness, we include a proof at the end of this section. Now we start with the facts specifically derived for the missing observation case. The next result is crucial to establish the right dependence on $p$. The proof is deferred to the end of this section.

**Lemma 3.** *Let $Y$ as above. For every $v \in S^{d-1}$, we have*

$$\mathbb{E}(v^T \operatorname{Diag}(Y \otimes Y)v)^2 \leq 2p\kappa^4 \|\operatorname{Diag}(\Sigma)\|^2$$

*and*

$$\mathbb{E}(v^T \operatorname{Off}(Y \otimes Y)v)^2 \leq 4p^2\kappa^4\|\Sigma\|^2.$$

The main idea behind the proof of Proposition 1 consists in using the variational principle twice, one for the diagonal part and the other for the more delicate off-diagonal part.

*Proof.* **Diagonal Part**: We start by defining the parameter space of interest, namely

$$\Theta = \mathbb{R}^d \times \mathbb{R}^d.$$

Choose $\mu$ to be a product of two zero mean multivariate Gaussians with mean zero and covariance $\beta^{-1}I_d$, where $\beta > 0$ will be chosen later. For each $v \in S^{d-1}$, let $\rho_v$ be the product of two multivariate Gaussian distribution with mean $v$ and covariance $\beta^{-1}I_d$. By construction, $(\theta, \nu)$ is distributed according to $\rho_v$, therefore it satisfies that $\mathbb{E}_{\rho_v}(\theta, \nu) = (v, v)$. The standard formula for the $\mathcal{KL}$-divergence between two Gaussian measures [24] implies that

$$\mathcal{KL}(\rho_v, \mu) = \beta.$$

Let $\lambda_1 > 0$ be a free parameter to be optimized later. By the first part of Lemma 2, we have

$$\psi\left(\lambda_1 v^T \operatorname{Diag}(Y \otimes Y)v\right) = \psi\left(\lambda_1 \mathbb{E}_{\rho_v}\theta^T \operatorname{Diag}(Y \otimes Y)\nu\right)$$
$$\leq \mathbb{E}_{\rho_v}\log(1 + \lambda_1\theta^T \operatorname{Diag}(Y \otimes Y)\nu + \lambda_1^2(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2) + R.$$

where $R := \min\{1, \lambda_1^2 \mathbb{E}_{\rho_v}(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2/6\}$. Notice that $\mathbb{E}\theta_i^2 = \beta^{-1} + v_i^2$ and $\mathbb{E}\theta_i\theta_j = v_i v_j$ for all $i \in [d]$, therefore

$$\begin{aligned}
&\mathbb{E}_{\rho_v}(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2 = \mathbb{E}_{\rho_v}\left(\sum_{i=1}^d \langle Y, e_i\rangle^2 \theta_i \nu_i\right)^2 \\
&= \beta^{-2}\sum_{i=1}^d \langle Y, e_i\rangle^4 + \sum_{i,j=1}^d \langle Y, e_i\rangle^2 \langle Y, e_j\rangle^2 v_i^2 v_j^2 + 2\beta^{-1}\sum_{i=1}^d \langle Y, e_i\rangle^4 v_i^2 \\
&= \beta^{-2}\|\operatorname{Diag}(Y \otimes Y)\|_F^2 + (v^T \operatorname{Diag}(Y \otimes Y)v)^2 + 2\beta^{-1}\|\operatorname{Diag}(Y \otimes Y)v\|_2^2,
\end{aligned}$$

By symmetry, $\mathbb{P}\left(\theta^T \operatorname{Diag}(Y \otimes Y)\nu \geq v^T \operatorname{Diag}(Y \otimes Y)v\right) \geq \frac{1}{4}$. To see this, observe that it is equal to

$$\mathbb{P}(\langle \operatorname{Diag}(Y \otimes Y)\theta, (\nu - v)\rangle + \langle \operatorname{Diag}(Y \otimes Y)v, (\theta - v)\rangle \geq 0).$$

The second term is positive with probability one half. Conditioned on this event, the first term is positive with probability one half, and it is independence from the first term. We obtain that the probability of both are positive is at least one quarter. Therefore,

$$\min\left\{1, \frac{\lambda_1^2}{6}(v^T \operatorname{Diag}(Y \otimes Y)v)^2\right\} \leq 4\mathbb{E}_{\rho_v} \min\left\{1, \frac{\lambda^2}{6}(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2\right\}.$$

By the second part of Lemma 2, we have

$$\begin{aligned}
&\psi(\lambda_1 v^T \operatorname{Diag}(Y \otimes Y)v) \\
&\leq \mathbb{E}_{\rho_v} \log(1 + \lambda\theta^T \operatorname{Diag}(Y \otimes Y)\nu + C_1\lambda^2(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2) + R(Y, \beta),
\end{aligned}$$

where $R(Y,\beta) := \min\{1, 2\lambda_1^2\beta^{-1}\|\operatorname{Diag}(Y \otimes Y)v\|_2^2/6\} + \min\{1, \lambda_1^2\beta^{-2}\|Y \otimes Y\|_F^2/6\}$. For instance, let us focus on the first term. The goal is to apply Lemma 1 to the function $f$ defined below

$$f(Y, \theta, \nu) := \log(1 + \lambda_1\theta^T \operatorname{Diag}(Y \otimes Y)\nu + C_1\lambda_1^2(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2).$$

Using the numeric inequality $\log(1 + y) \leq y$, valid for all $y \geq -1$, followed by Fubini's theorem and Lemma 3, we have

$$\begin{aligned}
&\mathbb{E}_{\rho_v} \log \mathbb{E}\left(1 + \lambda_1\theta^T \operatorname{Diag}(Y \otimes Y)\nu + C_1\lambda_1^2(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2\right) \\
&\leq \mathbb{E}_{\rho_v}\mathbb{E}\left(\lambda_1\theta^T \operatorname{Diag}(Y \otimes Y)\nu + C_1\lambda_1^2(\theta^T \operatorname{Diag}(Y \otimes Y)\nu)^2\right) \\
&\leq p\lambda_1 v^T \operatorname{Diag}(\Sigma)v + C_1\lambda_1^2(p\beta^{-2}\kappa^4 \operatorname{Tr}^2(\Sigma) + 2\beta^{-1}p\kappa^4\|\operatorname{Diag}\Sigma\|^2 + p\kappa^4\|\Sigma\|^2) \\
&\leq p\lambda_1 v^T \operatorname{Diag}(\Sigma)v + C_1\lambda_1^2(p\beta^{-2}\kappa^4 \operatorname{Tr}^2(\Sigma) + 2\beta^{-1}p\kappa^4\|\Sigma\|^2 + p\kappa^4\|\Sigma\|^2).
\end{aligned}$$

Next, setting $\beta := r(\Sigma)$ (which is at least one) and applying Lemma 1, it follows that with probability at least $1 - \delta$, for all $v \in S^{d-1}$,

$$\begin{aligned}
&\frac{1}{N\lambda_1}\sum_{i=1}^N \psi(\lambda_1 v^T \operatorname{Diag}(Y \otimes Y)v) \\
&\leq pv^T \operatorname{Diag}(\Sigma)v + C\lambda_1 p\|\Sigma\|^2\kappa^4 + \sum_{i=1}^n \frac{R_i}{N\lambda_1} + \frac{r(\Sigma) + \log(1/\delta)}{\lambda_1 N}.
\end{aligned}$$

Here $R_i$ is an independent copy of $R$. We proceed to estimate the third term in the right-hand side. Clearly, since $\min\{1, 2\lambda_1^2\beta^{-1}\|\operatorname{Diag}(Y \otimes Y)v\|_2^2/6\}$ is bounded by one, its variance is bounded by its expectation. Therefore, by Bernstein's inequality it follows that with probability $1-\delta$,

$$\begin{aligned}\frac{1}{\lambda_1 N}\sum_{i=1}^N \min\{1, \beta^{-2}\lambda_1^2\|Y\otimes Y\|_F^2/6\} &\lesssim \mathbb{E}\beta^{-2}\|Y\otimes Y\|_F^2 + \frac{\log(1/\delta)}{\lambda_1 N}\\ &\lesssim \lambda_1 p\kappa^4\|\Sigma\|^2 + \frac{\log(1/\delta)}{\lambda_1 N}.\end{aligned}$$

An analogous computation shows the same estimate holds (up to an absolute constant) for the term $\min\{1, 2\beta^{-1}\|\operatorname{Diag}(Y\otimes Y)v\|_2^2/6\}$. Finally we conclude that, there exists an absolute constant $C > 0$ such that, with probability at least $1-\delta$,

$$\begin{aligned}&\frac{1}{N\lambda_1}\sum_{i=1}^N \psi(\lambda_1 v^T \operatorname{Diag}(Y_i\otimes Y_i)v)\\ &\le pv^T\operatorname{Diag}(\Sigma)v + C\left(\lambda_1 p\|\Sigma\|^2\kappa^4 + \frac{r(\Sigma)+\log(1/\delta)}{\lambda_1 N}\right).\end{aligned}$$

We optimize the right-hand side over $\lambda_1 > 0$. More accurately, setting

$$\lambda_1 = \frac{1}{\|\Sigma\|\kappa^2 p}\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}$$

we obtain that, with probability at least $1-\delta$,

$$\frac{1}{N\lambda_1}\sum_{i=1}^N \psi(\lambda_1 v^T\operatorname{Diag}(Y_i\otimes Y_i)v) \le pv^T\operatorname{Diag}(\Sigma)v + C\kappa^2\sqrt{p}\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

We repeat the same arguments above for $\rho_{2,v}$ being a product measure between two Gaussians $\theta \sim N(v, \beta^{-1}I_d)$ and $\nu \sim N(-v, \beta^{-1}I_d)$. The argument follows exactly the same steps because $\psi$ is symmetric. Therefore, it also holds that with probability $1-\delta$,

$$-\frac{1}{N\lambda_1}\sum_{i=1}^N \psi(\lambda_1 v^T\operatorname{Diag}(Y_i\otimes Y_i)v) \le -pv^T\operatorname{Diag}(\Sigma)v + C\kappa^2\sqrt{p}\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

By union bound, we obtain a two-sided bound: With probability at least $1-\delta$,

$$\|p^{-1}\widehat{\Sigma}_1 - \operatorname{Diag}(\Sigma)\| \lesssim \frac{1}{\sqrt{p}}\kappa^2\|\Sigma\|\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

**Off-diagonal part**: We now proceed to the second part of the proof to deal with the off-diagonal part. We choose $\mu$ and $\rho(v)$ as before, and write

$$\begin{aligned}&\psi\left(\lambda_2 v^T\operatorname{Off}(Y\otimes Y)v\right) = \psi\left(\lambda_2\mathbb{E}_{\rho_v}\theta^T\operatorname{Off}(Y\otimes Y)\nu\right)\\ &\le \mathbb{E}_{\rho_v}\log(1+\lambda_2\theta^T\operatorname{Off}(Y\otimes Y)\nu + \lambda_2^2(\theta^T\operatorname{Off}(Y\otimes Y)\nu)^2) + R.\end{aligned}$$

where $R_2 := \min\{1, \lambda_2^2 \mathbb{E}_{\rho_v}(\theta^T \operatorname{Off}(Y \otimes Y)\nu)^2/6\}$. We have to deal with the quadratic form of the off-diagonal that requires a more delicate analysis. In fact,

$$\mathbb{E}_{\rho_v}(\theta^T \operatorname{Off}(Y \otimes Y)\nu)^2 = \mathbb{E}_{\rho_v} \sum_{i \neq j; k \neq l} \langle Y, e_i\rangle\langle Y, e_j\rangle\langle Y, e_k\rangle\langle Y, e_l\rangle \theta_i \nu_j \theta_k \nu_l$$

By independence between $\theta$ and $\nu$, it remains to analyze the term $\mathbb{E}\theta_i\theta_k\mathbb{E}\nu_j\nu_l$. We split the analysis in three cases: The first one when $k = i$ and $j = l$, the second when either $k = i$ and $j \neq l$ or $k \neq i$ and $j = l$, and finally the third one when $k \neq i$ and $j \neq l$. In the first case, the summation becomes

$$\sum_{i \neq j} \langle Y, e_i\rangle^2 \langle Y, e_j\rangle^2 (\beta^{-1} + v_i^2)(\beta^{-1} + v_j^2).$$

In the second case, the summation becomes

$$\sum_{i \neq j \neq l} \langle Y, e_i\rangle^2 \langle Y, e_j\rangle\langle Y, e_l\rangle (\beta^{-1} + v_i^2) v_j v_l + \sum_{i \neq j \neq k} \langle Y, e_i\rangle\langle Y, e_j\rangle^2 \langle Y, e_k\rangle (\beta^{-1} + v_j^2) v_i v_k.$$

The third case is simpler,

$$\sum_{i \neq j \neq k \neq l} \langle Y, e_i\rangle\langle Y, e_j\rangle\langle Y, e_k\rangle\langle Y, e_l\rangle v_i v_j v_k v_l.$$

Observe that summing all terms that do not contain any $\beta$ factor, we obtain $(v^T \operatorname{Off}(Y \otimes Y)v)^2$. As before, the goal is to apply Lemma 1 to the function $f$,

$$f(Y, \theta, \nu) := \log(1 + \lambda_2 \theta^T \operatorname{Off}(Y \otimes Y)\nu + C_2 \lambda_2^2 (\theta^T \operatorname{Off}(Y \otimes Y)\nu)^2),$$

where $C_2 > 0$ is a sufficiently large absolute constant. Using again the numeric inequality $\log(1 + y) \leq y$, Fubini's theorem, and Lemma 3, we have

$$\begin{aligned} &\mathbb{E}_{\rho_v} \log \mathbb{E}\left(1 + \lambda_2 \theta^T \operatorname{Off}(Y \otimes Y)\nu + C_2 \lambda_2^2 (\theta^T \operatorname{Off}(Y \otimes Y)\nu)^2\right) \\ &\leq \mathbb{E}_{\rho_v} \mathbb{E}\left(\lambda_2 \theta^T \operatorname{Off}(Y \otimes Y)\nu + C_2 \lambda_2^2 (\theta^T \operatorname{Off}(Y \otimes Y)\nu)^2\right). \end{aligned}$$

The first term is equal to $p^2 \lambda_2 v^T \operatorname{Off}(\Sigma) v$. We know that all terms in the expansion of $\theta^T \operatorname{Off}(Y \otimes Y)\nu$ without a $\beta$ factor add up $v^T \operatorname{Off}(Y \otimes Y)v$ and its expectation is at most $4p^2\kappa^4\|\Sigma\|^2$ by Lemma 3. Next, we estimate the terms containing $\beta$ systematically. Using Cauchy-Schwarz inequality together with the moment equivalence for $X$, we obtain

$$\begin{aligned} &\beta^{-2} \sum_{i \neq j} \mathbb{E}\langle Y, e_i\rangle^2 \langle Y, e_j\rangle^2 = \beta^{-2} p^2 \sum_{i \neq j} \mathbb{E}\langle X, e_i\rangle^2 \langle X, e_j\rangle^2 \leq p^2 \beta^{-2} \kappa^4 \sum_{i \neq j} \Sigma_{ii}\Sigma_{jj} \\ &\leq p^2 \beta^{-2} \kappa^4 \operatorname{Tr}^2(\Sigma). \end{aligned}$$

Similarly, we obtain

$$\mathbb{E} \sum_{i \neq j} \langle Y, e_i\rangle^2 \langle Y, e_j\rangle^2 (\beta^{-1} + v_i^2)(\beta^{-1} + v_j^2) \lesssim p^2 \kappa^4 \beta^{-2} \operatorname{Tr}^2(\Sigma) + \beta^{-1} p^2 \kappa^4 \|\Sigma\| \operatorname{Tr}(\Sigma).$$

It remains to analyze

$$\mathbb{E}\sum_{i\neq j\neq l}\langle Y,e_i\rangle^2\langle Y,e_j\rangle\langle Y,e_l\rangle\beta^{-1}v_jv_l+\sum_{i\neq j\neq k}\langle Y,e_i\rangle\langle Y,e_j\rangle^2\langle Y,e_k\rangle\beta^{-1}v_iv_k.$$

We estimate the first term on the right-hand side as the second term is identically distributed. To this end, we apply Hölder's inequality with conjugate exponents $4/3$ and $4$, and the moment equivalence to obtain that

$$\begin{aligned}
&\mathbb{E}\sum_{i\neq j\neq l}\langle Y,e_i\rangle^2\langle Y,e_j\rangle\langle Y,e_l\rangle\beta^{-1}v_jv_l \le p^3\sum_{i\neq j\neq l}\mathbb{E}\langle X,e_i\rangle^2\langle X,e_j\rangle\langle X,e_l\rangle\beta^{-1}v_jv_l\\
&\le p^3\sum_{i;j\neq l}\mathbb{E}\langle X,e_i\rangle^2\langle X,e_j\rangle\langle X,e_l\rangle\beta^{-1}v_jv_l+\left|2p^3\sum_{j\neq l}\mathbb{E}\langle X,e_j\rangle^3\langle X,e_l\rangle\beta^{-1}v_jv_l\right|\\
&\lesssim\frac{p^3}{\beta}\left(\mathbb{E}\left[(v^T\operatorname{Off}(X\otimes X)v)\left(\sum_{i=1}^d\langle X,e_i\rangle^2\right)\right]+\kappa^4\sum_{j\neq l}(\Sigma_{ll})^{1/2}(\Sigma_{jj})^{3/2}v_lv_j\right)\\
&\lesssim p^3\left[(v^T\operatorname{Off}(X\otimes X)v)\left(\sum_{i=1}^d\beta^{-1}\langle X,e_i\rangle^2\right)\right]+p^3\beta^{-1}\kappa^4\|\Sigma\|\operatorname{Tr}(\Sigma)\\
&\le p^3\mathbb{E}(v^TX\otimes Xv-v^T\operatorname{Diag}(X\otimes X)v)(\beta^{-1}\operatorname{Tr}(X\otimes X))+2p^3\frac{\kappa^4}{\beta}\|\Sigma\|\operatorname{Tr}(\Sigma)\\
&\le p^3\mathbb{E}\langle X,v\rangle^2\beta^{-1}\operatorname{Tr}(X\otimes X)+2p^3\beta^{-1}\kappa^4\|\Sigma\|\operatorname{Tr}(\Sigma)\\
&\lesssim p^3\kappa^4(\|\Sigma\|^2+\beta^{-2}\operatorname{Tr}^2(\Sigma)+\beta^{-1}\|\Sigma\|\operatorname{Tr}(\Sigma)),
\end{aligned}$$

where the last inequality follows from the arithmetic-geometric inequality. Putting all together, we conclude that

$$\mathbb{E}\mathbb{E}_{\rho_v}(\theta^T\operatorname{Off}(Y\otimes Y)\nu)^2\lesssim p^2\beta^{-2}\kappa^4\operatorname{Tr}^2(\Sigma)+p^2\kappa^4\|\Sigma\|^2+p^3\beta^{-1}\kappa^4\operatorname{Tr}(\Sigma)\|\Sigma\|.$$

Therefore, setting $\beta=r(\Sigma)$, it follows that

$$\begin{aligned}
&\mathbb{E}_{\rho_v}\log\mathbb{E}\left(1+\lambda_2\theta^T\operatorname{Diag}(Y\otimes Y)\nu+C_2\lambda_2^2(\theta^T\operatorname{Diag}(Y\otimes Y)\nu)^2\right)\\
&\le\lambda_2v^T\operatorname{Off}(Y\otimes Y)v+C_2\lambda_2^2\delta^2\kappa^4\|\Sigma\|^2.
\end{aligned}$$

Finally we conclude that there exists an absolute constant $C_2'>0$ for which, with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{N\lambda_2}\sum_{i=1}^N\psi(\lambda_2v^T\operatorname{Off}(Y_i\otimes Y_i)v)\\
&\le p^2v^T\operatorname{Off}(\Sigma)v+C_2'\left(\lambda_2p^2\|\Sigma\|^2\kappa^4+\sum_{i=1}^N\frac{R_2(Y_i)}{\lambda_2N}+\frac{r(\Sigma)+\log(1/\delta)}{\lambda_2N}\right).
\end{aligned}$$

By Bernstein inequality the remainder terms $R_2(Y_i)$ are absorbed by the last term in the sum exactly in the same way as in the diagonal case. We optimize

over $\lambda_2 > 0$ by setting

$$\lambda_2 := \frac{1}{p\|\Sigma\|\kappa^2}\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

Therefore, with probability $1-\delta$,

$$\frac{1}{N\lambda_2}\sum_{i=1}^{N}\psi(\lambda_2 v^T \operatorname{Off}(Y_i\otimes Y_i)v) \le p^2 v^T \operatorname{Off}(\Sigma)v + C\kappa^2 p\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

We repeat the arguments by changing the mean of $\nu$ to $-v$. This gives the other side of the inequality in the same way it was done for the diagonal part. We conclude that, with probability $1-\delta$,

$$\|p^{-2}\widehat{\Sigma}(\lambda_2) - \operatorname{Off}(\Sigma)\| \lesssim \frac{1}{p}\kappa^2\|\Sigma\|\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

By triangular inequality, union bound and re-scaling the multiplicative constant in $\delta$, the following holds. The estimator $\widehat{\Sigma}$ satisfies, with probability $1-\delta$,

$$\|\widehat{\Sigma}-\Sigma\| \lesssim \frac{\kappa^2}{p}\|\Sigma\|\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}}.$$

□

To end this section, we prove some technical facts, Lemma 2 and 3. We start with the proof of Lemma 3.

*Proof.* We start with the diagonal case. Observe that

$$\mathbb{E}(v^T \operatorname{Diag}(Y\otimes Y)v)^2 = \mathbb{E}\sum_{i,j=1}^{d}\langle Y,e_i\rangle^2\langle Y,e_j\rangle^2 v_i^2 v_j^2$$
$$= \sum_{i=1}^{d}\mathbb{E}\langle Y,e_i\rangle^4 v_i^4 + \sum_{i\neq j}^{d}\mathbb{E}\langle Y,e_i\rangle^2\langle Y,e_j\rangle^2 v_i^2 v_j^2 := (I)+(II).$$

Clearly, (I) is at most $p\kappa^4\sum_{i=1}^{d}\Sigma_{ii}^2 v_i^4 \le p\kappa^4\|\operatorname{Diag}(\Sigma)\|^2$. Next, by the arithmetic-geometric inequality

$$(II) \le \frac{p^2}{2}\sum_{i\neq j}\mathbb{E}(\langle X,e_i\rangle^4 v_i^2 v_j^2 + \langle X,e_j\rangle^4 v_i^2 v_j^2) \le p^2\kappa^4\|\operatorname{Diag}(\Sigma)\|^2.$$

For the off-diagonal term, we need to proceed carefully as the natural idea to decompose the off-diagonal matrix into the matrix itself minus the diagonal part leads to suboptimal dependence on $p$. We first expand it directly,

$$\mathbb{E}(v^T \operatorname{Off}(Y\otimes Y)v)^2 = \sum_{i\neq j;k\neq l}\mathbb{E}\langle Y,e_i\rangle\langle Y,e_j\rangle\langle Y,e_k\rangle\langle Y,e_l\rangle v_i v_j v_k v_l$$
$$\le p^2\sum_{i\neq j;k\neq l}\mathbb{E}\langle X,e_i\rangle\langle X,e_j\rangle\langle X,e_k\rangle\langle X,e_l\rangle v_i v_j v_k v_l = p^2\mathbb{E}(v^T \operatorname{Off}(X\otimes X)v)^2,$$

where the term $p^2$ comes from the fact that at least two indices are distinct in each summand. Now, we split the off-diagonal term $\mathbb{E}(v^T \operatorname{Off}(X \otimes X)v)^2$. More accurately,

$$\begin{aligned}&\mathbb{E}(v^T \operatorname{Off}(X \otimes X)v)^2 = \mathbb{E}(v^T(X \otimes X)v)^2 \\ &+ \mathbb{E}(v^T \operatorname{Diag}(X \otimes X)v)^2 - 2\mathbb{E}(v^T(X \otimes X)v)\mathbb{E}(v^T \operatorname{Diag}(X \otimes X)v) \\ &:= (a) + (b) + (c).\end{aligned}$$

The last term $(c)$ is negative because both matrices are positive semidefinite, so we can safely ignore it. The first term $(a)$ on is at most $\kappa^4(v^T\Sigma v)^2 \leq \kappa^4\|\Sigma\|^2$ by the moment equivalence assumption. Finally, the second term $(b)$ is at most $\kappa^4\|\Sigma\|^2$ by the same argument used above. □

Next, we proceed to prove Lemma 2.

*Proof.* Notice that $\psi(x) \leq \log(1 + x + x^2)$ holds trivially, and we add $x^2/6$ to make the latter function convex. It follows that

$$\psi(\mathbb{E}Z) \leq \min\{\log(1 + \mathbb{E}Z + \mathbb{E}Z^2) + \mathbb{E}Z^2/6, 1\}.$$

Now, we apply Jensen's inequality to conclude the proof of the first part. For the second part, notice that by Taylor series expansion, if $t \in [0, a]$ then we have the following inequality,

$$e^t \leq 1 + \frac{t}{a}\left(\sum_{i=1}^{\infty} \frac{a^i}{i!}\right) \leq 1 + \frac{t}{a}(e^a - 1),$$

therefore

$$\begin{aligned}&\mathbb{E}\log(1 + Z + Z^2) + a\mathbb{E}\min\{1, Z^2/6\} \\ &= \mathbb{E}\log\left(\left(1 + Z + Z^2\right)\exp(\min\{a, aZ^2/6\})\right) \\ &\leq \mathbb{E}\log\left(\left(1 + Z + Z^2\right)\left(1 + \min\{1, Z^2/6\}\left(e^a - 1\right)\right)\right).\end{aligned}$$

To get the inequality in the statement, we only need to split into the cases where $|Z|^2/6$ is smaller than one and where it is greater than one. □

## 3. Proof of Theorem 1

In the previous section, we showed in Proposition 1, that the proof of the main result boils down to estimate the trace of the covariance matrix, the operator norm, and the sparsifying parameter $p$. For the trace and operator norm, it is enough to estimate it with a multiplicative absolute constant. On the other hand, for the parameter $p$, we need a more accurate estimator. In fact, since we need to divide the estimator by $p$, an estimator $\widehat{p}$ that do not convergence to $p$ would insert a bias.

*Remark* 1. The best possible convergence rate is at least

$$\|\widehat{\Sigma} - \Sigma\| \leq \|\Sigma\| \left( \frac{1}{p} \sqrt{\frac{r(\Sigma) + \log(1/\delta)}{N}} \right).$$

The trivial estimator $\widehat{\Sigma} = 0$ satisfies $\|\widehat{\Sigma} - \Sigma\| \leq \|\Sigma\|$, so in order to have a meaningful result we need $\left( \frac{1}{p} \sqrt{\frac{r(\Sigma)+\log(1/\delta)}{N}} \right) < 1$. Therefore, without making any further comments, we may assume that

$$N \geq C \left( \frac{r(\Sigma) + \log(1/\delta)}{p^2} \right),$$

for some well-chosen $C > 0$.

### *3.1. Estimation of p*

The idea here is to explore the proportion of non-zeros entries in the observed data. In any standard data set, a missed value does not appear with zero; we set it to zero for convenience, as we have done throughout the entire manuscript until now. As it happens, when estimating the proportion of missing values, it could be the case that the distribution of the random vector $X$ has non-trivial mass at zero. Clearly, we can distinguish between the zero that comes from the distribution and the zero from the missing value. Equivalently, we may assume that the marginals of $X$, namely $\langle X, v \rangle$ (for every $v \in S^{d-1}$), do not have mass at zero.

The starting point is the following. We collect $Y_1, \ldots, Y_N$, and compute $Z_1, \ldots, Z_N$, where $Z_i(j) = 1$ if and only $Y_i(j) \neq 0$ and zero, otherwise. The goal is to estimate the mean of the random variable

$$R(Z) := \frac{1}{d} \|Z\|_{\ell_1},$$

as it is equal to $\mathbb{E}R(Z) = p$.

**Lemma 4.** *Let $Y_1, \ldots, Y_N$ be i.i.d copies of $X \odot \mathbf{p}$. There exists an estimator $\widehat{p}$ depending only on the sample and the confidence level $\delta$ satisfying that, with probability at least $1 - \delta$,*

$$|\widehat{p} - p| \leq Cp \sqrt{\frac{\log(1/\delta)}{N}}.$$

*As an immediate consequence, if $N \geq C \log(1/\delta)$, then (with the same probability guarantee)*

$$\frac{1}{2} p \leq \widehat{p} \leq \frac{3}{2} p.$$

Before we proceed to the proof, we remark that if $\widehat{p} > 1$, then we round it, $\widehat{p} = 1$.

*Proof.* Following the notation above, we collect $R(Z_1), \dots, R(Z_N)$ i.i.d copies of $R(Z)$. We invoke a standard sub-Gaussian mean estimator for $R(Z)$ (e.g trimmead mean estimator [19, Theorem 1]) together with the fact that $\operatorname{Var}(R(Z)) \leq p^2$, to obtain that, with probability at least $1 - \delta$,

$$|\widehat{R}(p) - \mathbb{E}R(Z)| \leq C\sqrt{\frac{\log(1/\delta)\operatorname{Var} R(Z)}{N}} \leq Cp\sqrt{\frac{\log(1/\delta)}{N}}.$$

□

### *3.2. Estimation of the trace*

To simplify the analysis, we can safely assume that $p$ is known because we can accurately estimate it using Lemma 4. Clearly,

$$p\operatorname{Tr}(\Sigma) = \mathbb{E}\sum_{i=1}^{d}\langle Y, e_i\rangle^2.$$

To invoke a mean estimator, we need to compute the standard deviation of the random variable in the right hand side. To this end, we have

$$\mathbb{E}\left(\sum_{i=1}^{d}\langle Y, e_i\rangle^2\right)^2 \leq p\sum_{i=1}^{d}\mathbb{E}\langle X, e_i\rangle^4 + p^2\sum_{i\neq j}\mathbb{E}\langle X, e_i\rangle\langle X, e_j\rangle^2 \lesssim p\kappa^4\operatorname{Tr}(\Sigma)^2.$$

The latter step follows from moment equivalence and Hölder's inequality (as we have been doing several times in this manuscript). Since $p$ is know, one may invoke Theorem 1 [19] to obtain an estimator $\widehat{\operatorname{Tr}}(\Sigma)$ satisfying that, with probability $1 - \delta$,

$$|\widehat{\operatorname{Tr}}(\Sigma) - p\operatorname{Tr}(\Sigma)| \leq C\kappa^2 p\operatorname{Tr}(\Sigma)\sqrt{\frac{\log(1/\delta)}{N}}.$$

If the sample size $N$ satisfies that $N \geq C\kappa^2\log(1/\delta)$ then for sufficiently large $C$, we have

$$|\widehat{\operatorname{Tr}}(\Sigma) - p\operatorname{Tr}(\Sigma)| \leq \frac{p\operatorname{Tr}(\Sigma)}{2},$$

and consequently

$$\frac{1}{2}\operatorname{Tr}(\Sigma) \leq p^{-1}\widehat{\operatorname{Tr}}(\Sigma) \leq \frac{3}{2}\operatorname{Tr}(\Sigma). \tag{3.1}$$

### *3.3. Estimation of the operator norm*

The most delicate part of this section is the estimation of the operator norm. The main lemma is the following

**Lemma 5.** *Let $Y_1, \dots, Y_N$ be i.i.d copies of $X \odot p$. There exist an absolute constant $C_N$ and an estimator $\widehat{\|\Sigma\|}$ depending only on the samples and $\kappa$ satisfying that, with probability at least $1-\delta$,*

$$c_2(\kappa)\|\Sigma\| \leq \widehat{\|\Sigma\|} \leq c_1(\kappa)\|\Sigma\|,$$

*provided that $N \geq C_N p^{-2}(\log(1/\delta) + r(\Sigma))$. Here $c_1, c_2 > 0$ are two absolute constants depending only on $\kappa$.*

The key idea is to repeat the same analysis as before for each part with an additional parameter $\alpha$, and show that if certain inequalities are satisfied then $\alpha$ needs to be of same order as the operator norm. Along the proof $C_1 > 0$ is an explicit constant that can be computed by just keeping track of the constants in the proofs of Section 2.

*Proof.* **Diagonal Part:** As before, we set

$$\Theta = \mathbb{R}^d \times \mathbb{R}^d.$$

Now, we slightly change the choice of measures. More accurately, we choose the measure $\mu$ to be a product of two zero mean multivariate Gaussians with mean zero and covariance $\beta^{-1} I_d$. For $v \in S^{d-1}$, let $\rho_v$ be a product of two multivariate Gaussian distribution with mean $\alpha v$ and covariance $\beta^{-1} I_d$. The $\mathcal{KL}$-divergence becomes

$$\mathcal{KL}(\rho_v, \mu) = \alpha^2 \beta.$$

To simplify the notation, we write $\rho_{v,\alpha} = \rho_v$. Following the same lines for the proof of the diagonal part, we have with probability at least $1 - 3\delta$,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \psi(\alpha^2 v^T \operatorname{Diag}(Y_i \otimes Y_i) v) \leq{}& \alpha^2 p v^T \operatorname{Diag}(\Sigma) v \\ &+ C_1 p \|\operatorname{Diag}(\Sigma)\|^2 \kappa^4 (\alpha^4 + \beta^{-1}\alpha^2) \\ &+ (C_1 \beta^{-2} p \kappa^4) \operatorname{Tr}(\Sigma)^2 + \frac{2\log(1/\delta)}{N} + \frac{\alpha^2 \beta}{N}.\end{aligned}$$

Next, we choose $\beta = c_\beta \operatorname{Tr}(\Sigma)$ where $c_\beta > 0$ is an absolute constant to be chosen later. By the Remark 1, we may define a constant $C_N > 0$ for which $N \geq C_N p^{-2} \max\{r(\Sigma), \log(1/\delta)\}$, and then

$$\begin{aligned}\frac{1}{pN}\sum_{i=1}^{N} \psi(\alpha^2 v^T \operatorname{Diag}(Y_i \otimes Y_i) v) \leq{}& \alpha^2 v^T \operatorname{Diag}(\Sigma) v + C_1 \|\Sigma\|^2 \kappa^4 \alpha^4 \\ &+ C_1 \kappa^4 c_\beta^{-1} \alpha^2 \|\Sigma\| + C_1 c_\beta^{-2} \kappa^4 + 2 C_N^{-1} \\ &+ \alpha^2 \|\Sigma\| c_\beta C_N^{-1}.\end{aligned}$$

**Off-Diagonal Part:** We use the same choice of the measures and proceed analogously. We obtain that, with probability at least $1 - 5\delta$, the following

holds

$$\frac{1}{p^2 n}\sum_{i=1}^{n}\psi(\alpha^2 v^T \operatorname{Off}(Y_i \otimes Y_i)v) \leq \alpha^2 v^T \operatorname{Off}(\Sigma)v + C_1\alpha^4\kappa^4\|\Sigma\|^2 + C_1 c_\beta^{-2}\kappa^4 + C_1 c_\beta^{-1}\kappa^4\alpha^2\|\Sigma\| + 2C_N^{-1} + C_N^{-1}c_\beta\alpha^2\|\Sigma\|.$$

**Everything Together:** We define the function $g(\alpha) : \mathbb{R} \to \mathbb{R}$ to be equal to

$$\frac{1}{Np}\sup_{v\in S^{d-1}}\sum_{i=1}^{N}\psi(\alpha^2 v^T \operatorname{Diag}(Y_i \otimes Y_i)v) + \frac{1}{Np^2}\sup_{v\in S^{d-1}\cup 0}\sum_{i=1}^{N}\psi(\alpha^2 v^T \operatorname{Off}(Y_i \otimes Y_i)v).$$

From above, we obtain that, with probability at least $1 - 8\delta$,

$$g(\alpha) \leq C_1\|\Sigma\|^2\alpha^4\kappa^4 + \|\Sigma\|\alpha^2(\kappa^2 + \kappa^4 C_1 c_\beta^{-1} + c_\beta C_N^{-1}) + C_1\kappa^4 c_\beta^{-2} + 4c_n^{-1}.$$

Notice that the constants $C_1 c_\beta^{-1} + c_\beta C_N^{-1}$ and $C_1\kappa^4 c_\beta^{-2} + 4C_N^{-1}$ can be made arbitrarily small by increasing $C_N$. In particular, we choose $c_\beta$ and $C_N$ so that

$$g(\alpha) \leq C_1\|\Sigma\|^2\alpha^4\kappa^4 + \|\Sigma\|\alpha^2(1 + L_1) + L_2, \tag{3.2}$$

where $L_1, L_2 > 0$ are two absolute constants satisfying the following conditions:

$$1.1L_2 < 1 \quad \text{and} \quad (1 - L_1)^2 - 8.4\kappa^4 C_1 L_2 > 0 \tag{3.3}$$

The reason for such choice will become clear in what follows. Next, without loss of generality, we assume that $\mathbb{P}(Y_i = 0) = 0$. This is always possible by adding a small amount of Gaussian noise without changing the covariance too much. We construct a vector $w \in S^{d-1}$ such that $\min_{i\in[n]}\langle Y_i, w\rangle \neq 0$ by sampling the vector from an isotropic Gaussian distribution, and normalizing it to have Euclidean norm exactly one.

Notice that $g(0) = 0$ and $g$ is a continuous function. Moreover, if we show that $g$ assumes values greater or equal to one then by intermediate value theorem, the function $g$ assumes any value within this range. Since $w$ is a unit vector for which $\min_{i\in[n]}\langle w, Y_i\rangle \neq 0$, and for every $i \in [N]$,

$$\langle Y_i, w\rangle^2 = w^T Y_i \otimes Y_i w = w^T \operatorname{Diag}(Y_i \otimes Y_i)w + w^T \operatorname{Off}(Y_i \otimes Y_i)w,$$

it follows that at least one of the terms in the right-hand side is non-zero. In the case that both are non-zero, we evaluate $g$ at the point

$$\min\left\{\min_{i\in[n]|}|w^T \operatorname{Diag}(Y_i \otimes Y_i)w|, \min_{i\in[d]}|w^T \operatorname{Off}(Y_i \otimes Y_i)w\}|\right\}. \tag{3.4}$$

It is clear that $g$ is at least one at such point. Moreover, observe that the function $g$ is non-negative as we are allowed to take $v = 0$ in the supremum of the off-diagonal part. In the case that one term is zero, we just remove it

from (3.4). Finally, regardless the case, we choose $\widehat{\alpha}$ such that $g(\widehat{\alpha}) = 1.1L_2$. This is a valid choice. Indeed, recall from (3.3) that $1.1L_2$ is strictly smaller than one, therefore existence of such $\widehat{\alpha}$ is guaranteed by the intermediate value theorem as argued before.

Next, (3.2) implies that

$$C_1\|\Sigma\|^2\widehat{\alpha}^4\kappa^4 + \|\Sigma\|\widehat{\alpha}^2(1+L_1) - 0.1L_2 \geq 0.$$

The expression above can be interpreted as a parabola in the variable $x := \widehat{\alpha}^2\|\Sigma\|$ that has two real roots. One root is negative, and it does not play any role. The other one is a positive absolute constant implying that there exists a constant $c_{min}(\kappa)$ such that

$$\widehat{\alpha}^2\|\Sigma\| \geq c_{min}(\kappa). \tag{3.5}$$

This translates in a lower bound for $\|\Sigma\|$. We now need an upper bound for $\|\Sigma\|$ in terms of $\widehat{\alpha}$. We repeat the same argument above for the product measure $\rho_{2,v}$ between $\theta$ and $\nu$, where $\theta \sim N(\alpha v, \beta^{-1}I_d)$ and $\nu \sim N(-\alpha v, \beta^{-1}I_d)$. Therefore, if $v_1 \in S^{d-1}$ is the normalized eigenvector corresponding to the maximum eigenvalue of $\Sigma$, then

$$-g(\alpha) \leq -\frac{1}{np}\sum_{i=1}^{n}\psi(\alpha^2 v_1^T \operatorname{Diag}(Y_i \otimes Y_i)v_1) - \frac{1}{np^2}\sum_{i=1}^{n}\psi(\alpha^2 v_1^T \operatorname{Off}(Y_i \otimes Y_i)v_1).$$

Moreover, since $-g(\alpha)$ is non-increasing in the interval $[0, \widehat{\alpha}]$, we have

$$-1.1L_2 = -g(\widehat{\alpha}) \leq C_1\|\Sigma\|^2\widehat{\alpha}^4\kappa^4 - \|\Sigma\|\alpha^2(1-L_1) + L_2$$

Setting $x = \|\Sigma\|\alpha^2$, the inequality above holds for all $\alpha \in [0, \widehat{\alpha}]$. It follows that,

$$C_1\kappa^4 x^2 - (1-L_1)x + 2.1L_2 \geq 0.$$

The discriminant of the quadratic equation is $\Delta = (1-L_1)^2 - 8.4C_1\kappa^4 L_2$ which is (strictly) positive by (3.3). It follows that the inequality above is true if $x \leq x_1$ or $x \geq x_2$, where $0 < x_1 < x_2$ are the positive roots of the corresponding quadratic equation. We claim that $x \geq x_2$ cannot happen. Otherwise, since the inequality above holds for all $\alpha \in [0, \widehat{\alpha}]$, it must hold for $\alpha^*$ such that $\|\Sigma\|\alpha^* \in (x_1, x_2)$, but this contradicts the fact that the parabola assumes negative values between $(x_1, x_2)$. Therefore, we obtain that there exists a constant $c_{max}(\kappa) > 0$ such that $\widehat{\alpha}^2\|\Sigma\| \leq c_{max}(\kappa)$. Putting together with (3.5), we obtain that

$$c_{min}(\kappa) \leq \widehat{\alpha}^2\|\Sigma\| \leq c_{max}(\kappa).$$

We conclude the proof by setting $\widehat{\|\Sigma\|} := \widehat{\alpha}^{-2}$. □

### *3.4. Completion of the proof of Theorem 1*

The final construction of our estimator is the following:

1. Split the sample $Y_1, \dots, Y_N$ into four parts of size at least $\lfloor N/4 \rfloor$ each.
2. Estimate the parameter $p$ with the first quarter of the sample using Lemma 4.
3. Estimate the trace $\operatorname{Tr}(\Sigma)$ with the second quarter using (3.1) and the operator norm $\|\Sigma\|$ with the third quarter using Lemma 5.
4. For the last quarter of the sample, use the estimator from Proposition 1 to estimate the covariance matrix.

Before proceeding to the proof, we highlight some features about the data-splitting approach. From a theoretical perspective, it only affects the convergence rate by a constant. However, from a practical standpoint, it might be of interest to avoid wasting one quarter of the sample if there are only a few observations missing. This means that we use the complete data to estimate the covariance by setting $\widehat{p} = 1$.

Moreover, it might be more appropriate to use a smaller fraction of the data to estimate the trace, as it is a one-dimensional quantity, and its convergence rate is faster than the convergence rate of the estimator itself. Unfortunately, due to the intractability of our estimator, we are unable to implement these ideas in a real dataset.

We are now in position to prove our main result, Theorem 1.

*Proof.* As discussed in Section 2, the proof follows easily once we estimate the parameters of the truncation level. Indeed, the truncation levels in Proposition 1 only requires the knowledge of $\operatorname{Tr}(\Sigma)$, $\|\Sigma\|$ and $p$ up to an absolute constant. The error that we need to take in account is to use the estimated value of $p$ instead of the true value when we divide by $p$. This is the only reason why we have to estimate the precise value of the parameter $p$. To this end, by triangle inequality

$$\left\| \frac{1}{\widehat{p}} \widehat{\Sigma}_1 - \operatorname{Diag}(\Sigma) \right\| \leq \left\| \frac{1}{\widehat{p}} \left( \widehat{\Sigma}_1 - p \operatorname{Diag}(\Sigma) \right) \right\| + \left\| \frac{1}{\widehat{p}} \left( p \operatorname{Diag}(\Sigma) - \widehat{p} \operatorname{Diag}(\Sigma) \right) \right\|.$$

We apply Lemma 4 to estimate both terms. The first term in the right hand side is, with probability at least $1 - \delta$, at most

$$\frac{\sqrt{p}}{\widehat{p}} C \|\Sigma\| \sqrt{\frac{r(\Sigma) + \log(1/\delta)}{N}} \lesssim \frac{1}{\sqrt{p}} \|\Sigma\| \sqrt{\frac{r(\Sigma) + \log(1/\delta)}{N}}.$$

The second term also satisfies, with probability $1 - \delta$,

$$\left\| \frac{1}{\widehat{p}} \left( p \operatorname{Diag}(\Sigma) - \widehat{p} \operatorname{Diag}(\Sigma) \right) \right\| \leq \| \operatorname{Diag}(\Sigma) \| \frac{1}{\widehat{p}} |p - \widehat{p}| \lesssim \|\Sigma\| \sqrt{\frac{\log(1/\delta)}{N}}.$$

The same argument holds for the off-diagonal part as clearly $\| \operatorname{Off}(\Sigma) \| \leq 2\|\Sigma\|$. We omit it for the sake of simplicity. Finally, the desired probability guarantee holds by union bound a constant number of times. □

**Acknowledgments**

The author would like to thank Tanja Finger, Felix Kuchelmeister and Nikita Zhivotovskiy for helpful discussions.

## References

[1] P. Abdalla and N. Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *arXiv preprint arXiv:2205.08494*, 2022.

[2] R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010. MR2601042

[3] T. T. Cai and A. Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of multivariate analysis*, 150:55–74, 2016. MR3534902

[4] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. MR2483528

[5] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. MR3052407

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

[7] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[8] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *arXiv preprint arXiv:1802.04308*, 2018.

[9] A. Elsener and S. van de Geer. Sparse spectral estimation with missing and corrupted measurements. *Stat*, 8(1):e229, 2019. MR3978409

[10] O. Guédon, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. On the interval of fluctuation of the singular values of random matrices. *Journal of the European Mathematical Society*, 19(5):1469–1505, 2017. MR3635358

[11] R. Kannan, L. Lovász, and M. Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997. MR1608200

[12] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019. MR4017523

[13] Y. Klochkov and N. Zhivotovskiy. Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method. *Electronic Journal of Probability*, 25, 2020. MR4073683

[14] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. MR3556768

[15] J. Depersin and G. Lecué. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probability Theory and Related Fields*, 183(3-4): 997–1025, 2022. MR4453320

[16] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric Aspects of Functional Analysis*, pages 277–299. Springer, 2017. MR3645128

[17] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24, 2011. MR3015038

[18] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014. MR3217437

[19] G. Lugosi and S. Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021. MR4206683

[20] S. Mendelson and G. Paouris. On generic chaining and the smallest singular value of random matrices with heavy tails. *Journal of Functional Analysis*, 262(9):3775–3811, 2012. MR2899978

[21] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, 48(3):1648–1664, 2020. MR4124338

[22] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019. MR4474486

[23] R. I. Oliveira and Z. F. Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails. *arXiv preprint arXiv:2209.13485*, 2022.

[24] L. Pardo. *Statistical inference based on divergence measures*. CRC press, 2018. MR2183173

[25] S. Park, X. Wang, and J. Lim. Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics*, 15(2):4868–4915, 2021. MR4327332

[26] E. Pavez and A. Ortega. Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Transactions on Information Theory*, 67(2):1201–1215, 2020. MR4232009

[27] M. Rao, T. Javidi, Y. C. Eldar, and A. Goldsmith. Estimation in autoregressive processes with partial observations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4212–4216, 2017.

[28] N. Srivastava and R. Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111, 2013. MR3127875

[29] K. Tikhomirov. Sample covariance matrices of heavy-tailed distributions.

*International Mathematics Research Notices*, 2018(20):6254–6289, 2018. MR3872323

[30] R. Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012. MR2956207

[31] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019. MR3967104

[32] N. Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *arXiv preprint arXiv:2108.08198*, 2021. MR4693860