Statistica Sinica Preprint No: SS-2021-0388	
Title	Robust Estimation of Covariance Matrices: Adversarial Contamination and Beyond
Manuscript ID	SS-2021-0388
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0388
Complete List of Authors	Stanislav Minsker and Lang Wang
Corresponding Authors	Stanislav Minsker
E-mails	minsker@usc.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Robust Estimation of Covariance Matrices: Adversarial Contamination and Beyond

Stanislav Minsker and Lang Wang

University of Southern California

Abstract: We consider the problem of estimating the covariance structure of a random vector $Y \in \mathbb{R}^d$ from an i.i.d. sample $Y_1, \dots, Y_n$. We are interested in the situation when d is large compared to n but the covariance matrix Σ of interest has (exactly or approximately) low rank. We assume that the given sample is either (a) ε-adversarially corrupted, meaning that ε fraction of the observations could have been replaced by arbitrary vectors, or that (b) the sample is i.i.d. but the underlying distribution is heavy-tailed, meaning that the norm of Y possesses only finite fourth moments. We propose estimators that are adaptive to the potential low-rank structure of the covariance matrix as well as to the proportion of contaminated data, and admit tight deviation guarantees despite rather weak underlying assumptions. Finally, we show that the proposed construction leads to numerically efficient algorithms that require minimal tuning from the user, and demonstrate the performance of such methods under various models of contamination.

Key words and phrases: Adversarial contamination, covariance estimation, heavy-tailed distribution, low-rank recovery, U-statistics.

1. Introduction.

The focus of this paper is on the problem of covariance estimation under various types of contamination, with an emphasis on the practical methods that admit efficient implementation. Assume that we are given independent copies $Y_1, \dots, Y_n$ of a random vector $Y \in \mathbb{R}^d$ which follows an unknown distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean $\mu := \mathbb{E}[X]$ and covariance matrix $\Sigma := \mathbb{E}\big[(Y-\mu)(Y-\mu)^T\big]$. The observations $Y_1, \dots, Y_n$ are assumed to be either ε-*adversarially corrupted*, meaning that an "adversary" could replace an $\varepsilon < 0.5$ fraction of them by arbitrary (possibly random) vectors, or that (b) the underlying distribution $\mathcal{D}$ is heavy-tailed, meaning that the Euclidean norm $\|Y\|_2$ is assumed to possess only 4 finite moments. Our goal is to construct an estimator of the covariance matrix Σ that performs well in the present framework.

As attested by some early references such as the works by Tukey (1960); Huber (1964), robust estimation has a long history. During the past two decades, growing number of applications created a high demand for the practical tools to recover high-dimensional parameters of interest from corrupted measurements. Robust covariance estimators in particular have been studied extensively. Statistical properties of the sample covariance

matrix for the "light-tailed," for example sub-Gaussian distributions, are well-understood: results in this direction can be found in the works by Koltchinskii and Lounici (2016); Vershynin (2010); Cai et al. (2010, 2016), among many others. In the paper Srivastava and Vershynin (2013), performance of the sample covariance matrix was investigated under weaker moment assumptions. Some popular robust estimators of scatter, such as the Minimum Covariance Determinant (MCD) estimator and the Minimum Volume Ellipsoid (MVE) estimator, are discussed in Hubert et al. (2008). However, rigorous results for these estimators are available only for elliptically symmetric distributions as, in general, they are biased. For instance, Butler et al. (1993) discusses asymptotic results for the MCD, while Davies (1992) – for the MVE estimator. Other popular constructions, such as Maronna (1976) and Tyler (1987) estimators of scatter, are consistent only for the distributions possessing certain symmetry properties. Among recent works, Chen et al. (2018) demonstrate minimax optimality, with respect to the proportion of outliers, of the robust estimator based on the so-called "matrix depth" function inspired by the notion of Tukey's depth; unfortunately, this estimator is not computationally tractable. Covariance estimation for heavy-tailed distributions has attracted significant attention in the last few years; results in this direction appear in the papers by Catoni

(2016), Giulini (2015), Fan et al. (2016), Abdalla and Zhivotovskiy (2022), Oliveira and Rico (2022) as well as Minsker (2018); Minsker and Wei (2020). The survey by Ke et al. (2019) contains a more detailed overview of the recent progress. Contributions by theoretical computer scientists have introduced a range of new ideas leading to theoretically optimal estimators in the adversarial contamination framework; a small subsample of works includes Lai et al. (2016); Diakonikolas et al. (2021, 2019, 2017); Cheng et al. (2019) as well as the excellent survey Diakonikolas and Kane (2019); the very recent works by Abdalla and Zhivotovskiy (2022), Oliveira and Rico (2022) describe estimators that achieve sharpest possible bounds. While several proposed approaches, including the very recent works by Abdalla and Zhivotovskiy (2022), Oliveira and Rico (2022), result in optimal with respect to the contamination proportion and dependence on the dimension factors estimators, the corresponding algorithms are either not computationally feasible or not user-friendly, as they are often sensitive to the choice of "absolute constants" appearing in the tuning parameters, require preliminary robust mean estimation, or assume that (typically unknown) parameters such as the contamination proportion ε are given as an input; other works focus only on the bounds with respect to the Frobenius norm, while we are interested in the error measured in the operator norm as well. Finally, the

dependence of resulting probabilistic estimates on the deviation parameter controlling the probability of the desirable bound is often not made explicit.

The present paper continues the line of research on robust covariance estimation: we design a "Lasso-type" penalized estimator, and show that (a) it admits nearly optimal error bounds in the cases of practical interest, namely when the so-called "effective rank" of the covariance matrix Σ (to be rigorously defined later) is small, (b) that it requires minimal tuning and can be efficiently calculated using traditional numerical methods and (c) the dependence of resulting estimates on all parameters of interest is explicitly stated. Let us emphasize that theoretical guarantees for our estimator are not restricted to the case when the data are generated from an elliptically-symmetric distribution.

The rest of the paper is organized as follows: section 2 introduces the main notation and background material. Sections 3 and 4 display the main results for the cases of adversarially corrupted data and heavy-tailed data respectively. Section 5 presents the algorithms for numerical evaluation of the proposed estimators as well as results of numerical experiments. Finally, additional simulation results and proofs are contained in the supplementary material.

2. Preliminaries.

In this section, we introduce the main notation and recall several useful facts that we rely on in the subsequent exposition.

2.1 Notation.

Given two real numbers $a, b \in \mathbb{R}$, we define $a \vee b := \max\{a, b\}$, $a \wedge b := \min\{a, b\}$. For $x \in \mathbb{R}$, we will denote $\lfloor x \rfloor := \max\{n \in \mathbb{Z} : n \leq x\}$ to be the largest integer less than or equal to x. The absolute constants will typically be unspecified and will be denoted via $c, C, C_1, \tilde{C}$, etc., where the same constant letter might denote different absolute constants in different expressions. When the constant depends on certain parameters of the problem, it will be written as $C(x, y, \ldots)$. Remaining notation will be introduced on demand.

2.2 Matrix algebra.

Assume that $A \in \mathbb{R}^{d_1 \times d_2}$ is a $d_1 \times d_2$ matrix with real-valued entries. Let A^T denote the transpose of A and define $S^d(\mathbb{R}) := \left\{A \in \mathbb{R}^{d \times d} : A^T = A\right\}$ to be the set of all symmetric $d \times d$ matrices. The eigenvalues of A will be denoted $\lambda_1, \ldots, \lambda_d$, all of which are real numbers. Given a square matrix $A \in \mathbb{R}^{d \times d}$, the trace of A is $\mathrm{tr}\,(A) := \sum_{i=1}^{d} A_{i,i}$, where $A_{i,i}$ represents the element on

the i^{th} row and i^{th} column of A. For a rectangular matrix $A \in \mathbb{R}^{d_1 \times d_2}$ with singular values $\sigma_1(A) \geq \cdots \geq \sigma_{\text{rank}(A)}(A) \geq 0$, the operator or spectral norm is defined via $\|A\| := \sigma_1(A) = \sqrt{\lambda_{max}(A^T A)}$, the Frobenius norm – via $\|A\|_F := \sqrt{\sum_{i=1}^{\text{rank}(A)} \sigma_i^2(A)} = \sqrt{\text{tr}\,(A^T A)}$, and the nuclear norm – via $\|A\|_1 := \sum_{i=1}^{\text{rank}(A)} \sigma_i(A) = \text{tr}\,(\sqrt{A^T A})$. The inner product associated with the Frobenius norm is defined as $\langle A, B \rangle := \langle A, B \rangle_F = \text{tr}\,(A^T B) = \text{tr}\,(AB^T)$, where $A, B \in \mathbb{R}^{d_1 \times d_2}$. Finally, we introduce the functions of matrix-valued arguments.

Definition 1. Given a real-valued function f defined on an interval $\mathbb{T} \subseteq \mathbb{R}$ and a real symmetric matrix $A \in S^d(\mathbb{R})$ with the spectral decomposition $A = U \Lambda U^T$ such that $\lambda_j(A) \in \mathbb{T}$, $j = 1, \ldots, d$, define $f(A)$ as $f(A) = U f(\Lambda) U^T$, where

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

Finally, the effective rank of a matrix $A \in S^d(\mathbb{R}) \setminus \{0\}$ is defined as

$$\text{rk}(A) := \frac{\text{tr}\,(A)}{\|A\|}.$$

Note that $1 \leq \text{rk}(A) \leq \text{rank}(A)$ is always true, and it is possible that $\text{rk}(A) \ll \text{rank}(A)$ for "approximately low-rank" matrices A.

2.3 Sub-Gaussian distributions.

Given a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and a convex nondecreasing function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $\psi(0) = 0$, we define the ψ-norm of X, following Vershynin (2018, Section 2.7.1), as

$$\|X\|_\psi := \inf \left\{ C > 0 : \mathbb{E}\left[\psi\left(\frac{|X|}{C}\right)\right] \leq 1 \right\}.$$

Below, we will be interested in $\psi_1(u) := \exp\{u\} - 1, u \geq 0$ and $\psi_2(u) := \exp\{u^2\} - 1, u \geq 0$, which correspond to the sub-exponential and sub-Gaussian norms respectively. We will say that a random variable X is sub-Gaussian (sub-exponential) if $\|X\|_{\psi_2} < \infty$ ($\|X\|_{\psi_1} < \infty$). Also, we define the L_2 norm of a random variable X via $\|X\|_{L_2} := (\mathbb{E}[|X|^2])^{1/2}$. The sub-Gaussian (or sub-exponential) random vector is defined as follows:

Definition 2. A random vector Z in $\mathbb{R}^d$ with mean $\mu = \mathbb{E}[Z]$ is called L-sub-Gaussian if for every $v \in \mathbb{R}^d$, there exists an absolute constant $L > 0$ such that

$$\|\langle Z - \mu, v\rangle\|_{\psi_2} \leq L \|\langle Z - \mu, v\rangle\|_{L_2}. \tag{2.1}$$

Moreover, Z is called L-sub-exponential if ψ_2-norm in (2.1) is replaced by ψ_1-norm.

3. Problem formulation and main results.

Let $Z_1, \dots, Z_n \in \mathbb{R}^d$ be i.i.d. copies of an L-sub-Gaussian random vector Z such that $\mathbb{E}[Z] = \mu$ and $\mathbb{E}\big[(Z-\mu)(Z-\mu)^T\big] = \Sigma$. Assume that we observe a sequence

$$Y_j = Z_j + V_j, \ j = 1, \dots, n, \tag{3.2}$$

where V_j's are arbitrary (possibly random) vectors such that only a small portion of them are not equal to zero. Namely, we assume that there exists a set of indices $J \subseteq \{1, \dots, n\}$ such that $|J| \ll n$ and $V_j = 0$ for $j \notin J$. In what follows, the sample points with $j \in J$ will be called *outliers* and $\varepsilon := |J|/n$ will denote the proportion of such points. In this case,

$$Y_j Y_j^T = Z_j Z_j^T + \underbrace{V_j V_j^T + V_j Z_j^T + Z_j V_j^T}_{:=\sqrt{n}U_j^*} := X_j + \sqrt{n}U_j^*,$$

where $\operatorname{rank}(U_j^*) \le 2$ and the $\sqrt{n}$ normalization factor is added for the technical convenience. Our main goal is to construct an estimator for the covariance matrix Σ in the presence of outliers V_j. In practice, we usually do not know the true mean μ of Z. Our next immediate goal is to recall the well-known fact that explicit estimation of μ can be avoided if one is only interested in Σ. To this end, we recall the definition of U-statistics.

Definition 3 (Hoeffding (1948))**.** Let $Y_1, \dots, Y_n$ $(n \ge 2)$ be a sequence of random variables taking values in a measurable space $(\mathcal{S}, \mathcal{B})$. Assume that

$H : \mathcal{S}^m \mapsto \mathbb{S}^d(\mathbb{R})$ $(2 \le m \le n)$ is an $\mathcal{S}^m$-measurable permutation-symmetric kernel, i.e. $H(y_1, \ldots, y_m) = H(y_{\pi_1}, \ldots, y_{\pi_m})$ for any $(y_1, \ldots, y_m) \in \mathcal{S}^m$ and any permutation π. The U-statistic with kernel H is defined as

$$U_n := \frac{(n-m)!}{n!} \sum_{(i_1,\ldots,i_m)\in I_n^m} H(Y_{i_1}, \ldots, Y_{i_m}),$$

where $I_n^m := \{(i_1, \ldots, i_m) : 1 \le i_j \le n, i_j \ne i_k \text{ if } j \ne k\}$.

A particular example of a U-statistic is the sample covariance matrix

$$\widetilde{\Sigma}_s := \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})(Y_j - \bar{Y})^T, \tag{3.3}$$

where $\bar{Y} := \frac{1}{n}\sum_{j=1}^n Y_j$. Indeed, it is easy to verify that

$$\widetilde{\Sigma}_s = \frac{1}{n(n-1)} \sum_{(i,j)\in I_n^2} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}, \tag{3.4}$$

hence the sample covariance matrix is a U-statistic with kernel

$$H(x, y) := \frac{(x-y)(x-y)^T}{2} \text{ for any } x, y \in \mathbb{R}^d.$$

Note that $\mathbb{E}\big[(Y_i - Y_j)/\sqrt{2}\big] = 0$ and $\mathbb{E}\Big[(Y_i - Y_j)(Y_i - Y_j)^T/2\Big] = \Sigma$ for all $(i, j) \in I_n^2$. Namely, by expressing the sample covariance matrix as a U-statistic in (3.4), the explicit estimation of the unknown mean μ can be avoided. Therefore, we consider the following settings:

$$\widetilde{Y}_{i,j} := \frac{Y_i - Y_j}{\sqrt{2}}, \quad \widetilde{Z}_{i,j} := \frac{Z_i - Z_j}{\sqrt{2}}, \quad \widetilde{V}_{i,j} := \frac{V_i - V_j}{\sqrt{2}}, \text{ for all } (i, j) \in I_n^2.$$

Then

$$\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T = \widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T + \underbrace{\widetilde{V}_{i,j}\widetilde{V}_{i,j}^T + \widetilde{V}_{i,j}\widetilde{Z}_{i,j}^T + \widetilde{Z}_{i,j}\widetilde{V}_{i,j}^T}_{:=\sqrt{n(n-1)}\widetilde{U}_{i,j}^*} := \widetilde{X}_{i,j} + \sqrt{n(n-1)}\widetilde{U}_{i,j}^*,$$

where the $n(n-1) = |I_n^2|$ factor equals the total number of $\widetilde{Y}_{i,j}$'s, and is added for technical convenience. The followings facts can be easily verified:

(1) $\widetilde{Y}_{i,j} = \widetilde{Z}_{i,j} + \widetilde{V}_{i,j}$, with $\mathbb{E}\Big[\widetilde{Z}_{i,j}\Big] = 0$ and $\mathbb{E}\Big[\widetilde{Z}_{i,j}\widetilde{Z}_{i,j}^T\Big] = \Sigma$, for any $(i,j) \in I_n^2$. Moreover, $\widetilde{Z}_{i,j}, (i,j) \in I_n^2$ has sub-Gaussian distribution according to Corollary 2.

(2) $\widetilde{Z}_{i,j}$'s are identically distributed, but not independent.

(3) Denote $\widetilde{J} = \Big\{(i,j) \in I_n^2 : \widetilde{V}_{i,j} \neq 0\Big\}$ to be the set of indices such that $\widetilde{V}_{i,j} = 0$, $\forall (i,j) \notin \widetilde{J}$. Then $|\widetilde{J}|$ represents the number of outliers in $\Big\{\widetilde{Y}_{i,j} : (i,j) \in I_n^2\Big\}$, and we have that

$$|\widetilde{J}| = 2|J|(n - |J|) + |J|(|J| - 1) = |J|(2n - |J| - 1). \tag{3.5}$$

(4) $\text{Rank}(\widetilde{U}_{i,j}^*) \leq 2$. This follows from the fact that for any vector $v \in \mathbb{R}^d$, $\widetilde{U}_{i,j}^* v \in \text{span}\Big\{\widetilde{V}_{i,j}, \widetilde{Z}_{i,j}\Big\}$.

In the sequel, we will let $\mathbf{U_{I_n^2}} := (U_{1,2}, \ldots, U_{n,n-1})$ represent the $n(n-1)$-dimensional sequence with subscripts valued in I_n^2. Similarly, the notation $(S, \mathbf{U_{I_n^2}})$ will represent the $(n^2 - n + 1)$-dimensional sequence $(S, U_{1,2}, \ldots,$

$U_{n,n-1}$). Now we are ready to define our estimator. Given $\lambda_1, \lambda_2 > 0$, set

$$(\widehat{S}_\lambda, \widehat{\mathbf{U}}_{\mathbf{I}_\mathbf{n}^\mathbf{2}}) = \underset{S, U_{1,2}, \ldots, U_{n,n-1}}{\operatorname{argmin}} \left[\frac{1}{n(n-1)} \sum_{i \neq j} \left\| \widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S - \sqrt{n(n-1)} U_{i,j} \right\|_{\mathrm{F}}^2 + \lambda_1 \|S\|_1 + \lambda_2 \sum_{i \neq j} \|U_{i,j}\|_1 \right], \quad (3.6)$$

where the minimization is over $S, U_{i,j} \in S^d(\mathbb{R})$, $\forall (i,j) \in I_n^2$.

Remark 1. The double penalized least-squares estimator defined in (3.6) is in fact a solution to nuclear-norm penalized Huber's loss minimization problem. In the context of robust linear regression, this fact has been previously observed by several authors, including Sardy et al. (2001); Gannaz (2007); McCann and Welsch (2007); She and Owen (2011); Donoho and Montanari (2016). In the setting of robust Principal Component Analysis, similar connections have been established by She et al. (2016); the approach of this work is similar in spirit to the present one however it is tailored for the problem of estimating the leading principal components when the number of these principal components is known. To see the connection between (3.6) and penalized Huber's loss minimization in our framework, let us express the estimator as

$$(\widehat{S}_\lambda, \widehat{\mathbf{U}}_{\mathbf{I}_\mathbf{n}^\mathbf{2}}) = \arg\min_{S} \min_{\mathbf{U}_{\mathbf{I}_\mathbf{n}^\mathbf{2}}} \left[\frac{1}{n(n-1)} \operatorname{tr} \left[\sum_{i \neq j} \left(\widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S - \sqrt{n(n-1)} U_{i,j} \right)^2 \right] + \lambda_1 \|S\|_1 + \lambda_2 \sum_{i \neq j} \|U_{i,j}\|_1 \right], \quad (3.7)$$

and observe that the minimization with respect to $\mathbf{U}_{\mathbf{I}_\mathbf{n}^\mathbf{2}}$ in (3.7) can be carried out explicitly. It yields that

$$\widehat{S}_\lambda = \underset{S}{\operatorname{argmin}} \left\{ \frac{2}{n(n-1)} \operatorname{tr} \left[\sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j} \widetilde{Y}_{i,j}^T - S) \right] + \lambda_1 \|S\|_1 \right\}, \quad (3.8)$$

where

$$\rho_\lambda(u) := \begin{cases} \frac{u^2}{2}, & |u| \leq \lambda \\ \lambda |u| - \frac{\lambda^2}{2}, & |u| > \lambda \end{cases} \quad \text{, for all } u \in \mathbb{R}, \lambda \in \mathbb{R}^+ \quad (3.9)$$

is the Huber's loss function. For the reader's convenience, details of the derivation are given in section B.1 of the supplementary material.

3.1 Performance guarantees for adversarial contamination.

We are ready to state our main results, the error bounds for the estimator defined in (3.6). We will compare performance of our estimator to the sample covariance matrix $\widetilde{\Sigma}_s$ defined in (3.3). When there are no outliers, it is well-known that $\widetilde{\Sigma}_s$ is a consistent estimator of Σ with expected error at most $\mathcal{O}(d/\sqrt{n})$ in the Frobenius norm, namely, $\mathbb{E}\left[\left\|\widetilde{\Sigma}_s - \Sigma\right\|_F\right] \leq Cd/\sqrt{n}$ for some absolute constant $C > 0$ (see for example, Cai et al. (2010)). However, in the presence of outliers, the error for $\widetilde{\Sigma}_s$ can be large (see section D in the supplementary material for some concrete examples). Recall that $\widetilde{X}_{i,j} = \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T$. The following bound characterizes the performance of the estimator (3.6).

Theorem 1. *Fix $\delta > 0$, assume that $n \geq 2$ and that $|J| \leq c_1(\delta)n$, where $c_1(\delta)$ is a constant depending only on δ. Then on the event*

$$\mathcal{E} = \left\{ \lambda_1 \geq \frac{140\,\|\Sigma\|}{\sqrt{n(n-1)}}\sqrt{\mathrm{rk}(\Sigma)} + 4\left\|\frac{1}{n(n-1)}\sum_{(i,j)\in I_n^2}\widetilde{X}_{i,j} - \Sigma\right\|, \right.$$
$$\left. \lambda_2 \geq \frac{140\,\|\Sigma\|}{n(n-1)}\sqrt{\mathrm{rk}(\Sigma)} + \frac{4}{\sqrt{n(n-1)}}\max_{(i,j)\in I_n^2}\left\|\widetilde{X}_{i,j} - \Sigma\right\| \right\},$$

the following inequality holds:

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_{\mathrm{F}}^2 \leq \inf_{S:\,\mathrm{rank}(S)\leq\frac{c_2 n^2\lambda_2^2}{\lambda_1^2}} \left\{(1+\delta)\,\|S-\Sigma\|_{\mathrm{F}}^2 + c(\delta)\left(\lambda_1^2\,\mathrm{rank}(S) + \lambda_2^2|J|^2\right)\right\}.$$

Detailed proof of Theorem 1 is presented in section A.2 of the supplementary material.

Remark 2. The bound in Theorem 1 contains two terms:

(1) The first term, $(1+\delta)\,\|S-\Sigma\|_{\mathrm{F}}^2 + c(\delta)\lambda_1^2\,\mathrm{rank}(S)$, does not depend on the number of outliers. When there are no outliers, i.e. $|J| = 0$, the bound will only contain this term. In such a scenario, Lounici (2014) proved that the optimal bound has the form

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_{\mathrm{F}}^2 \leq \inf_S \left\{ \|\Sigma - S\|_F^2 + C\,\|\Sigma\|^2\,\frac{(\mathrm{rk}(\Sigma)+t)}{n}\,\mathrm{rank}(S)\right\}$$

that holds with probability at least $1 - e^{-t}$. By making the smallest valid choice of λ_1 specified in (3.10), one sees that the first term of our bound coincides with this optimal bound.

(2) The second term, $c(\delta)\lambda_2^2|J|^2$, controls the worst possible effect due to the presence of outliers. When more conditions on the outliers are imposed (for example, independence), this bound can be improved; see the discussion following equation (4.15). Moreover, Diakonikolas et al. (2017) proved that when Z is centered Gaussian, there exists an estimator $\widehat{\Sigma}$ achieving theoretically optimal, with respect to ε, bound $\left\|\widehat{\Sigma}-\Sigma\right\|_F \leq \mathcal{O}(\varepsilon)\,\|\Sigma\|$, which is independent of the dimension d. In our case, by making the smallest possible choice of λ_2, we can show that the error bound scales like $\mathcal{O}\Big(\big(\log(n)+\mathrm{rk}(\Sigma)\big)\varepsilon\Big)\,\|\Sigma\|$. The additional factor $\big(\log(n)+\mathrm{rk}(\Sigma)\big)$ shows that our bound is sub-optimal in general. However, in the class of matrices with $\mathrm{rk}(\Sigma)$ bounded by a constant, our bound is nearly optimal up to a logarithmic factor.

Note that in Theorem 1 the regularization parameters λ_1, λ_2 should be chosen sufficiently large such that the event $\mathcal{E}$ happens with high probability. Under the assumption that $Z_j, j = 1, \ldots, n$ are independent, identically distributed L-sub-Gaussian vectors, we can prove the following result which gives an explicit lower bound on the choice of λ_1.

Proposition 1. *Assume that Z is L-sub-Gaussian with mean μ and covariance matrix Σ. Let $Z_1, \ldots, Z_n$ be independent copies of Z, and define $\widetilde{Z}_{i,j} := (Z_i - Z_j)/\sqrt{2}$ for all $(i,j) \in I_n^2$. Then $\widetilde{Z}_{i,j}, (i,j) \in I_n^2$ are mean*

zero L-sub-Gaussian random vectors with covariance Σ. *Moreover, for any* $t \geq 1$, *there exists* $c(L) > 0$ *depending only on* L *such that*

$$\left\| \frac{1}{n(n-1)} \sum_{i \neq j} \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma \right\| \leq c(L) \|\Sigma\| \left(\sqrt{\frac{\mathrm{rk}(\Sigma) + t}{n}} + \frac{\mathrm{rk}(\Sigma) + t}{n} \right)$$

with probability at least $1 - 2e^{-t}$.

Proposition 1 along with the definition of event $\mathcal{E}$ indicates that it suffices to choose λ_1 satisfying

$$\lambda_1 \geq c(L) \|\Sigma\| \sqrt{\frac{\mathrm{rk}(\Sigma) + t}{n}}, \tag{3.10}$$

given that $n \geq \mathrm{rk}(\Sigma) + t$. The next proposition provides a lower bound for the choice of λ_2.

Proposition 2. *Assume that* Z *is* L*-sub-Gaussian with mean zero and* $Z_1, \ldots, Z_n$ *are copies of* Z *(not necessarily independent). Then there exists* $c(L) > 0$ *depending only on* L, *such that for any* $t \geq 1$,

$$\max_{j=1,\ldots,n} \|Z_j Z_j^T - \Sigma\| \leq c(L) \|\Sigma\| \left(\mathrm{rk}(\Sigma) + \log(n) + t\right)$$

with probability at least $1 - e^{-t}$.

Since Proposition 2 does not require independence, it can be applied to the mean zero, L-sub-Gaussian vectors $\widetilde{Z}_{i,j}, (i,j) \in I_n^2$ to deduce that

$$\max_{i \neq j} \left\| \widetilde{Z}_{i,j} \widetilde{Z}_{i,j}^T - \Sigma \right\| \leq c(L) \|\Sigma\| \left[\mathrm{rk}(\Sigma) + \log(n(n-1)) + t\right]$$

with probability at least $1 - e^{-t}$. Combining this bound with the definition of event $\mathcal{E}$, we conclude that it suffices to choose λ_2 satisfying

$$\lambda_2 \geq c(L) \|\Sigma\| \frac{(\mathrm{rk}(\Sigma) + \log(n) + t)}{n}. \tag{3.11}$$

By choosing the smallest possible λ_1, λ_2 as indicated in (3.10), (3.11), we deduce the following corollary:

Corollary 1. *Let $\delta > 0$ be an absolute constant. Assume that $n \geq \mathrm{rk}(\Sigma) + \log(n)$ and $|J| \leq c_1(\delta)n$, where $c_1(\delta)$ is a constant depending only on δ. Then we have that*

$$\left\|\widehat{S}_\lambda - \Sigma\right\|_F^2 \leq \inf_{S:\, \mathrm{rank}(S) \leq c_2' n \left(\mathrm{rk}(\Sigma)+\log(n)\right)} \left\{ (1+\delta) \|S - \Sigma\|_{\mathrm{F}}^2 \right.$$
$$\left. + c(L,\delta) \|\Sigma\|^2 \left[\frac{\mathrm{rk}(\Sigma) + \log(n)}{n} \mathrm{rank}(S) + \frac{(\mathrm{rk}(\Sigma) + \log(n))^2}{n^2} |J|^2 \right] \right\} \tag{3.12}$$

with probability at least $1 - 3/n$.

Note that the term $\|\Sigma\|^2 \frac{(\mathrm{rk}(\Sigma)+\log(n))^2}{n^2}|J|^2$ in (3.12) can be equivalently written in terms of ε, the proportion of outliers, as $\|\Sigma\|^2 (\mathrm{rk}(\Sigma) + \log(n))^2 \varepsilon^2$.

4. Performance guarantees for the heavy-tailed distributions.

In this section, we consider the case to heavy-tailed data and compare this framework to the model of adversarial contamination. Let $Y \in \mathbb{R}^d$ be a random vector with mean $\mathbb{E}[Y] = \mu$, covariance matrix $\Sigma = \mathbb{E}\big[(Y-\mu)(Y-\mu)^T\big]$,

and such that $\mathbb{E}\left[\|Y-\mu\|_2^4\right] < \infty$. Assume that $Y_1, \ldots, Y_n$ are i.i.d copies of Y, and as before our goal is to estimate Σ. As before, we define as before (delete this duplicated "as before") $\widetilde{Y}_{i,j} = (Y_i - Y_j)/\sqrt{2}$, and denote $H_{i,j} := \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T$. It was previously shown that $\mathbb{E}\left[\widetilde{Y}_{i,j}\right] = 0$ and $\mathbb{E}[H_{i,j}] = \Sigma$. Given $\lambda_1, \lambda_2 > 0$, we propose the following estimator for Σ:

$$\widehat{S}_\lambda = \underset{S}{\operatorname{argmin}} \left\{ \frac{1}{n(n-1)} \operatorname{tr}\left[\sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) \right] + \frac{\lambda_1}{2} \|S\|_1 \right\}, \quad (4.13)$$

which is the minimizer of the penalized Huber's loss function

$$L(S) = \frac{1}{n(n-1)} \operatorname{tr}\left[\sum_{i \neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) \right] + \frac{\lambda_1}{2} \|S\|_1 . \quad (4.14)$$

Recall that the estimator $\widehat{S}_\lambda$ in (4.13) is equivalent to the double-penalized least-squares estimator in (3.6) (see section B.1 of the supplementary material for the details). The key idea behind the derivation of the error bounds for $\widehat{S}_\lambda$ is to decompose the heavy-tailed distribution into as a mixture of a "well-behaved" and a contamination components (similar approach was previously used by Prasad et al. (2019)). Such decomposition can be viewed as a "bridge" between the heavy-tailed model and the adversarial contamination model (3.2), allowing us to repeat parts of the reasoning used to obtain the inequalities in Section 3. Specifically, we consider the decompo-

sition

$$\widetilde{Y}_{i,j} = \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 \le R\right\}}_{:=\widetilde{Z}_{i,j}} + \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 > R\right\}}_{:=\widetilde{V}_{i,j}}, \tag{4.15}$$

where $R > 0$ is the truncation level that will be specified later. One can view $\widetilde{V}_{i,j}$ as "outliers". Note however that such outliers can not be too bad: in particular, they are identically distributed and mutually independent as long as the subscripts do not overlap, therefore one can expect many cancellations to occur in the sum $\sum_{i,j}\widetilde{V}_{i,j}$. This fact in turn translates into better performance bounds of the proposed estimators. In the following two subsections, we will show that the estimator $\widehat{S}_\lambda$ in (4.13) is close to Σ both in the operator and the Frobenius norms.

4.1 Bounds in the operator norm.

Our goal is to show that $\widehat{S}_\lambda$ is close to Σ in the operator norm with high probability. We will be interested in the effective rank of the "variance matrix" $\mathbb{E}[(H_{1,2}-\Sigma)^2]$, and denote it

$$r_H := \mathrm{rk}(\mathbb{E}\big[(H_{1,2}-\Sigma)^2\big]) = \frac{\mathrm{tr}\left(\mathbb{E}[(H_{1,2}-\Sigma)^2]\right)}{\|\mathbb{E}[(H_{1,2}-\Sigma)^2]\|}.$$

Minsker and Wei (2020, Lemma 4.1) suggest that under the bounded kurtosis assumption (to be specified later, see (4.16)), we can upper bound r_H by the effective rank of Σ, namely, $r_H \le C\mathrm{rk}(\Sigma)$ with some constant $C > 0$.

4. PERFORMANCE GUARANTEES FOR THE HEAVY-TAILED DISTRIBUTIONS.

Theorem 2. *Assume that* $t \geq 1$ *is such that* $r_H t \leq c_3 n$ *for some sufficiently small constant* c_3, $\sigma \geq \|\mathbb{E}[(H_{1,2} - \Sigma)^2]\|^{\frac{1}{2}}$, *and* $n \geq \max\left\{64ar_H t, 4bt^2 \|\Sigma\|^2/\sigma^2\right\}$ *for some sufficiently large constants* a, b. *Then for* $\lambda_1 \leq (\sigma/4)\sqrt{n/t}$ *and* $\lambda_2 \geq \sigma/\sqrt{(n-1)t}$, *we have that*

$$\left\|\widehat{S}_\lambda - \Sigma\right\| \leq \frac{20}{39}\lambda_1 + \frac{80}{39}\sigma\sqrt{\frac{t}{n}} + \frac{40}{39}\lambda_2 t$$

with probability at least $1 - (8r_H/3 + 1)\, e^{-t}$.

It is also easy to see that the bound still holds if $\lambda_1 > (\sigma/4)\sqrt{n/t}$: indeed, the following lemma is rather straightforward.

Lemma 1. *Assume that* $t \geq 0$, $\sigma \geq \|\mathbb{E}[(H_{1,2} - \Sigma)^2]\|^{\frac{1}{2}}$ *and*

$$n \geq \max\left\{64ar_H t, \frac{4bt^2 \|\Sigma\|^2}{\sigma^2}\right\},$$

where a,b *are sufficiently large positive constants. Then for any* $\lambda_1 > (\sigma/4)\sqrt{n/t}$, *we have that* $\operatorname{argmin}_S L(S) = 0$ *with probability at least* $1 - e^{-t}$.

In particular, under the conditions of the previous lemma, $\left\|\widehat{S}_\lambda - \Sigma\right\| = \|\Sigma\|$. The proofs of Lemma 1 and Theorem 2 are presented in section A.5 of the supplementary material.

Remark 3. According to Minsker and Wei (2020, Lemma 4.1), the "matrix variance" parameter σ^2 appearing in the statement of Theorem 2 can

be bounded by $\|\Sigma\| \operatorname{tr}(\Sigma) = \operatorname{rk}(\Sigma) \|\Sigma\|^2$ under the bounded kurtosis assumption (4.16) formally stated below. In this case, $\|\mathbb{E}[(H_{1,2} - \Sigma)^2]\| \lesssim \operatorname{rk}(\Sigma) \|\Sigma\|^2$ and σ can be chosen to be proportional to $\sqrt{\operatorname{rk}(\Sigma)} \|\Sigma\|$. Moreover, in this case the assumptions on n and t in Lemma 1 and Theorem 2 can be reduced to a single assumption that $r_H t \leq c_3' n$ for some sufficiently small constant c_3'. It is also worth noting that the magnitude of deviations suggested by Theorem 2 is controlled by $\|\Sigma\| \sqrt{\operatorname{rk}(\Sigma)}$ (indeed, the term involving the deviations parameter t has the form $\lambda_2 t$) while the optimal, sub-Gaussian type deviations are controlled by $\|\Sigma\|$ as shown by Mendelson and Zhivotovskiy (2020). Unfortunately, the estimator proposed by Mendelson and Zhivotovskiy (2020) that achieves such bounds is not computationally tractable.

4.2 Bounds in the Frobenius norm.

In this subsection we show that $\widehat{S}_\lambda$ is close to the covariance matrix of Y in the Frobenius norm with high probability, under a slightly stronger assumption on the fourth moment of Y.

Definition 4. A random vector $Y \in \mathbb{R}^d$ is said to satisfy an $L_4 - L_2$ norm equivalence with constant K (also referred to as the bounded kurtosis

assumption) if there exists a constant $K \geq 1$ such that

$$\left(\mathbb{E}\left[\langle Y - \mathbb{E}Y, v\rangle^4\right]\right)^{\frac{1}{4}} \leq K \left(\mathbb{E}\left[\langle Y - \mathbb{E}Y, v\rangle^2\right]\right)^{\frac{1}{2}} \tag{4.16}$$

for any $v \in \mathbb{R}^d$.

As previously discussed in Remark 3, condition (4.16) allows us to connect the matrix variance parameter σ^2 with $\mathrm{rk}(\Sigma_Y)$, the effective rank of the covariance matrix Σ_Y. We will assume that Y satisfies (4.16) with a constant K throughout this subsection. Recall the decomposition

$$\widetilde{Y}_{i,j} = \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 \leq R\right\}}_{:=\widetilde{Z}_{i,j}} + \underbrace{\widetilde{Y}_{i,j}\mathbb{1}\left\{\left\|\widetilde{Y}_{i,j}\right\|_2 > R\right\}}_{:=\widetilde{V}_{i,j}}, \tag{4.17}$$

where $R > 0$ is the truncation level to be specified later. Denote $\Sigma_Y := \mathbb{E}\left[\widetilde{Y}_{1,2}\widetilde{Y}_{1,2}^T\right]$, $\Sigma_Z := \mathbb{E}\left[\widetilde{Z}_{1,2}\widetilde{Z}_{1,2}^T\right]$ and recall that our goal is to estimate Σ_Y. Since $\left\|\widetilde{Z}_{i,j}\right\|_2 \leq R$ almost surely, the equation (4.17) represents $\widetilde{Y}_{i,j}$ as a sum of a bounded vector $\widetilde{Z}_{i,j}$ and a "contamination" component $\widetilde{V}_{i,j}$, which is similar to the model (3.2). On the other hand, we note that the truncation level R should be chosen to be neither too large (to get a better behaved truncated distribution) nor too small (to reduce the bias introduced by the truncation). Mendelson and Zhivotovskiy (2020) suggest that a reasonable choice is given by

$$R = \left(\frac{\mathrm{tr}\left(\Sigma_Y\right)\|\Sigma_Y\| n}{\log\left(\mathrm{rk}(\Sigma_Y)\right) + \log(n)}\right)^{\frac{1}{4}}. \tag{4.18}$$

4. PERFORMANCE GUARANTEES FOR THE HEAVY-TAILED DISTRIBUTIONS.

Denote $\widetilde{J} = \left\{(i,j) \in I_n^2 : \left\|\widetilde{Y}_{i,j}\right\|_2 > R\right\}$ to be the set of indices corresponding to the nonzero outliers (i.e. $\widetilde{V}_{i,j} \neq 0$), and $\varepsilon := |\widetilde{J}|/\big(n(n-1)\big)$ to be the proportion of such outliers. Under this setup, we have the following result which provides an upper bound on ε with high probability:

Lemma 2. *Assume that Y satisfies the $L_4 - L_2$ norm equivalence with constant K, and R is chosen as in* (4.18)*. Then*

$$\varepsilon \leq c(K)\frac{\mathrm{rk}(\Sigma_Y)\left[\log\big(\mathrm{rk}(\Sigma_Y)\big) + \log(n)\right]}{n}$$

with probability at least $1 - 1/n$.

The proof of Lemma 2 is presented in Section A.6 of the supplementary material. It is worth noting that the proportion of "outliers" (in a sense of the definition above) in the heavy-tailed model can be pretty small when the sample size n is large. The following inequality is the main result of this section.

Theorem 3. *Given $A \geq 1$, assume that $Y \in \mathbb{R}^d$ is a random vector with mean $\mathbb{E}[Y] = \mu$, covariance matrix $\Sigma_Y = \mathbb{E}\big[(Y-\mu)(Y-\mu)^T\big]$, and satisfying an $L_4 - L_2$ norm equivalence with constant K. Let $Y_1, \ldots, Y_n$ be i.i.d samples of Y, and let $\widetilde{Z}_{i,j}$ be defined as in* (4.17)*. Assume that $n \geq c_4(K)\mathrm{rk}(\Sigma_Y)\big(\log(\mathrm{rk}(\Sigma_Y)) + \log(n)\big)$, and $\mathrm{rank}(\Sigma_Y) \leq c_2(K)n$. Then*

for

$$\lambda_1 = c(K)\,\|\Sigma_Y\|\left[\mathrm{rk}(\Sigma_Y)\big(\log(\mathrm{rk}(\Sigma_Y)) + \log(n)\big)\right]^{1/2} n^{-1/2}$$

and

$$\lambda_2 = c(K)\,\|\Sigma_Y\|\left(\mathrm{rk}(\Sigma_Y)\log(n)\right)^{1/2}(An)^{-1/2}$$

we have that

$$\left\|\widehat{S}_\lambda - \Sigma_Y\right\|_F^2 \leq c(K)\,\|\Sigma_Y\|^2\left[\frac{\mathrm{rk}(\Sigma_Y)\Big(\log\big(\mathrm{rk}(\Sigma_Y)\big) + \log(n)\Big)}{n}\mathrm{rank}(\Sigma_Y) + \frac{A\mathrm{rk}(\Sigma_Y)^2\log(n)^3}{n}\right]$$

with probability at least $1 - (8r_H/3+1)n^{-A} - 4n^{-1}$.

The proof of Theorem 3 is given in section A.7 of the supplementary material.

Remark 4. Let us compare the result in Theorem 3 to the bound of Corollary 1:

1. The first term of the bound, $c(K)\,\|\Sigma_Y\|^2\,\frac{\mathrm{rk}(\Sigma_Y)\left(\log\left(\mathrm{rk}(\Sigma_Y)\right)+\log(n)\right)}{n}\mathrm{rank}(\Sigma_Y)$, has the same order as in Corollary 1 (up to a logarithmic factor), under the assumption that Σ_Y has low rank. This part of the bound is theoretically optimal according to Remark 2.

2. The second part of the bound, $c(K)\,\|\Sigma_Y\|^2\,\frac{\mathrm{rk}(\Sigma_Y)^2\log(n)^3}{n}$, controls the error introduced by the outliers. It is smaller than the corresponding

quantity in Corollary 1 which in the present setup would be order $c(K)\left\|\Sigma_Y\right\|^2 \frac{\mathrm{rk}(\Sigma_Y)^3\log(n)^3}{n}$ (note the additional $\mathrm{rk}(\Sigma_Y)$ factor). As we mentioned before, the improvement is mainly due to the special structure of the heavy-tailed data, namely, independence among the outliers $\widetilde{V}_{i,j}$ with non-overlapping subscripts; see the discussion following equation (4.15).

5. Numerical experiments.

In this section we discuss algorithms for evaluating the proposed estimators as well as the numerical experiments. Recall that the loss function is defined via

$$\widetilde{L}(S, \mathbf{U}_{\mathbf{I}_\mathbf{n}^\mathbf{2}}) = \frac{1}{n(n-1)} \sum_{i\neq j} \left\| \widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S - \sqrt{n(n-1)}U_{i,j} \right\|_F^2 + \lambda_1 \left\|S\right\|_1 + \lambda_2 \sum_{i\neq j} \left\|U_{i,j}\right\|_1 . \quad (5.19)$$

We aim at approximating $(\widehat{S}_\lambda, \widehat{\boldsymbol{U}}_{\boldsymbol{I}_\boldsymbol{n}^\mathbf{2}})$, the minimizer of (5.19), numerically. Since we are only interested in $\widehat{S}_\lambda$ while $\widehat{\boldsymbol{U}}_{\boldsymbol{I}_\boldsymbol{n}^\mathbf{2}}$ are the nuisance parameters, equation (3.8) suggests that it suffices to minimize the following function:

$$L(S) := \frac{1}{n(n-1)} \mathrm{tr} \sum_{i\neq j} \rho_{\frac{\sqrt{n(n-1)}\lambda_2}{2}} (\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S) + \frac{\lambda_1}{2} \left\|S\right\|_1 ,$$

where $\rho_\lambda(\cdot)$ is the Huber's loss function defined in (3.9).

5.1 Algorithm for computing the estimator.

Our computational approach, formally described in Algorithm 1 below, is based on minimizing the loss function $L(S)$ via the batch proximal gradient descent (PGD) method: suppose that we want to minimize the function $f(x) = g(x)+h(x)$, where (a) g is convex, differentiable, and (b) h is convex but not necessarily differentiable. The proximal gradient descent method for solving the problem starts from an initial point $x^{(0)}$, and performs updates

$$x^{(k)} = \text{prox}_{\alpha_k h}\left(x^{(k-1)} - \alpha_k \nabla g(x^{(k-1)})\right),$$

where $\alpha_k > 0$ are the step sizes and $\text{prox}_h(x)$, the proximal mapping of a convex function h at the point x, is defined via

$$\text{prox}_h(x) = \underset{u}{\text{argmin}} \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right).$$

When $g(x) = \frac{1}{n}\sum_{i=1}^n g_i(x)$, where $g_1, \ldots, g_n$ are convex functions, the update step of PGD will require the evaluation of n gradients, which is expensive for large values of n. A natural alternative is to consider the stochastic proximal gradient descent method (SPGD), where at each iteration $k = 1, 2, \ldots$, we pick an index i_k randomly from $\{1, 2, \ldots, n\}$, and make the following update: $x^{(k)} = \text{prox}_{\alpha_k h}\left(x^{(k-1)} - \alpha_k \nabla g_{i_k}(x^{(k-1)})\right)$. Batch SPGD method assumes that we pick a small random subset of indices at each iteration, balancing the computational cost and the variance introduced by random sampling. Additional facts about proximal gradient descent and

its variants are presented in section C.1 of the supplementary material.

Algorithm 1 Stochastic proximal gradient descent (SPGD)

Input: number of iterations T, step size η_t, batch size b, tuning parameters λ_1 and λ_2, initial estimation S^0, sample size n, dimension d.

1: **for** $t = 1, 2, \ldots, T$ **do**

2: (1) Randomly pick $i_t, j_t \in \{1, 2, \ldots, n\}$ without replacement.

3: (2) Compute $G_t = -\nabla g_{i,j}(S^t) = -\rho'_{\frac{\sqrt{n(n-1)}\lambda_2}{2}}(\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S^t)$.

4: (3) If $b > 1$, then repeat (1)(2) for b times and save the average gradient in G_t.

5: (4) (**gradient update**) $T^{t+1} = S^t - G_t$.

6: (5) (**proximal update**)

$$S^{t+1} = \underset{S}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| S - T^{t+1} \right\|_F^2 + \frac{\lambda_1}{2} \left\| S \right\|_1 \right\} = \gamma_{\frac{\lambda_1}{2}}(T^{t+1}),$$

where $\gamma_\lambda(u) = \operatorname{sign}(u)(|u| - \lambda)_+$.

7: **end for**

Output: S^{T+1}

5.1.1 Rank-one update of the spectral decomposition.

Note that at each iteration of Algorithm 1, we need to compute the spectral decomposition of the matrices $\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S^t$ which is computationally expensive. However, since $\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T$ is a matrix of rank 1, and the spectral decomposition of S^t was already performed on step $T - 1$, the problem of

computing the spectral decomposition of matrices $\widetilde{Y}_{i,j}\widetilde{Y}_{i,j}^T - S^t$ can be viewed as a rank-one update of the spectral decomposition, which has been extensively studied (for example, see Bunch et al. (1978) and Stange (2008)). It turns out that with the help of rank-one update methods, the complexity of spectral decomposition can be reduced from $\mathcal{O}(d^3)$ to $\mathcal{O}(d^2 \log^2 d)$. Detailed description of the required techniques is given in section C.2 of the supplementary material.

5.2 Simulation results.

As a proof of concept, consider the following setup: $d = 200$, $n = 100$, $|J| = 3$, $\mu = (0, \ldots, 0)^T$, $\Sigma = \mathrm{diag}(10, 1, 0.1, \ldots, 0.1)$. The inputs to the algorithm are generated as follows: we sample n independent realizations Z_j from the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, and then replace $|J|$ of them (randomly chosen) with $Z_j + V_j$, where $V_j, j \in J$ are the outliers that are drawn independently from another Gaussian distribution $\mathcal{N}(\mu_V, \Sigma_V)$, with $\mu_V = (0, \ldots, 0)^T$, $\Sigma_V = \mathrm{diag}(100, \ldots, 100)$; results for other types of outliers are given in section D of the supplementary material. The sample $Y_1, \ldots, Y_j$ obtained in this manner is the input to the SPGD algorithm. Next, we calculate $\widetilde{Y}_{i,j} = (Y_i - Y_j)/\sqrt{2}, i \neq j$ and perform our algorithm with $K = 500$ steps and the diminishing step size $\alpha_k = 1/k$. The initial value S^0 is de-

termined by a one-step full gradient update, as explained in the last paragraph of section C.1 of the supplementary material. (C.78). To analyze the performance of estimators, we define $\mathrm{RelErr}(S, \mathrm{Frob}) := \|S - \Sigma\|_F / \|\Sigma\|_F$ to be the relative error of the estimator S in the Frobenius norm, and $\mathrm{RelErr}(S, \mathrm{op}) := \|S - \Sigma\| / \|\Sigma\|$ to be the relative error of the estimator S in the operator norm. We will compare the performance of the estimator S^* produced by our algorithm with the performance of the sample covariance matrix $\widetilde{\Sigma}_s$ introduced in (3.3). We performed 200 repetitions of the experiment with $\lambda_1 = 3$, $\lambda_2 = 1$, and recorded S^*, $\widetilde{\Sigma}_s$ for each run. Histograms illustrating the distributions of relative errors in the Frobenius norm are shown in figures 1 and 2. The average and maximum (over 200 repetitions) relative errors of S^* were 0.2842 and 0.6346 respectively, with the standard deviation of 0.1108. The corresponding values for $\widetilde{\Sigma}_s$ were 34.5880, 39.6758 and 2.1501. It is clear that estimator S^* performed considerably better than the sample covariance $\widetilde{\Sigma}_s$, as expected. Figures 3 and 4 show that S^* yields smaller relative errors in the operator norm as well. The average and maximum relative errors of S^* in the operator norm were 0.2676 and 0.6290 respectively, with the standard deviation of 0.1148. The corresponding values for $\widetilde{\Sigma}_s$ were 22.9255, 28.2328 and 1.8791.

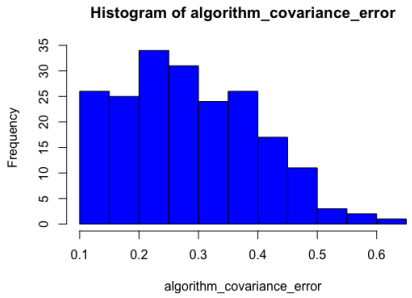

Figure 1: Distribution of RelErr(S^*, Frob).

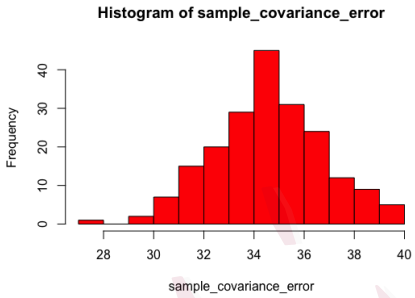

Figure 2: Distribution of RelErr($\widetilde{\Sigma}_s$, Frob).

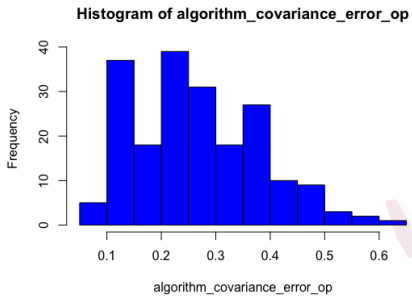

Figure 3: Distribution of RelErr(S^*, op).

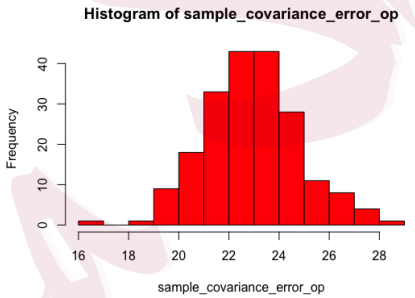

Figure 4: Distribution of RelErr($\widetilde{\Sigma}_s$, op).

Supplementary Materials

Supplementary materials, including detailed proofs and additional simulation results, will be provided separately as an online accessible file.

Acknowledgements

Authors acknowledge support by the National Science Foundation grants DMS CAREER-2045068 and CIF-1908905.

References

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

Aleksandrov, A. B. and V. V. Peller (2016). Operator Lipschitz functions. *Russian Mathematical Surveys 71*(4), 605.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bhatia, R. (2013). *Matrix analysis*, Volume 169. Springer Science & Business Media.

Bunch, J. R., C. P. Nielsen, and D. C. Sorensen (1978). Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik 31*(1), 31–48.

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

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

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance

matrix estimation. *The Annals of Statistics 38*(4), 2118–2144.

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

Chen, M., C. Gao, and Z. Ren (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics 46*(5), 1932–1960.

Cheng, Y., I. Diakonikolas, R. Ge, and D. P. Woodruff (2019). Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pp. 727–757. PMLR.

Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, 1828–1843.

Dekel, O., R. Gilad-Bachrach, O. Shamir, and L. Xiao (2012). Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research 13*(1).

Diakonikolas, I., G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing 48*(2), 742–864.

Diakonikolas, I., G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart (2017). Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pp. 999–1008. PMLR.

Diakonikolas, I., G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart (2021). Robustness

meets algorithms. *Communications of the ACM 64*(5), 107–115.

Diakonikolas, I. and D. M. Kane (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Donoho, D. and A. Montanari (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields 166*(3), 935–969.

Fan, J., W. Wang, and Y. Zhong (2016). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *arXiv preprint arXiv:1603.03516*.

Gandhi, R. and A. Rajgor (2017). Updating singular value decomposition for rank one matrix perturbation. *arXiv preprint arXiv:1707.08369*.

Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing 17*(4), 293–310.

Gimpel, K., D. Das, and N. A. Smith (2010). Distributed asynchronous online learning for natural language processing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 213–222.

Giulini, I. (2015). PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. *arXiv preprint arXiv:1511.06263*.

Golub, G. H. (1973). Some modified matrix eigenvalue problems. *Siam Review 15*(2), 318–334.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals*

of Mathematical Statistics, 293–325.

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

Hubert, M., P. J. Rousseeuw, and S. Van Aelst (2008). High-breakdown robust multivariate methods. *Statistical Science*, 92–119.

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

Khirirat, S., H. R. Feyzmahdavian, and M. Johansson (2017). Mini-batch gradient descent: Faster convergence under data sparsity. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 2880–2887. IEEE.

Kishore Kumar, N. and J. Schneider (2017). Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra 65*(11), 2212–2244.

Koltchinskii, V. and K. Lounici (2016). New asymptotic results in principal component analysis. *arXiv preprint arXiv:1601.01457*.

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

Lai, K. A., A. B. Rao, and S. Vempala (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 665–674. IEEE.

REFERENCES

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

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

McCann, L. and R. E. Welsch (2007). Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics and Data Analysis 52*(1), 249–257.

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

Minsker, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters 127*, 111–119.

Minsker, S. (2018). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics 46*(6A), 2871–2903.

Minsker, S. and X. Wei (2020). Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli 26*(1), 694–727.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, Volume 87. Springer Science & Business Media.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate O (1/kˆ 2). In *Dokl. akad. nauk Sssr*, Volume 269, pp. 543–547.

Nitanda, A. (2014). Stochastic proximal gradient descent with acceleration techniques. *Advances*

in Neural Information Processing Systems 27, 1574–1582.

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

Prasad, A., S. Balakrishnan, and P. Ravikumar (2019). A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*.

Sardy, S., P. Tseng, and A. Bruce (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing 49*(6), 1146–1152.

Shalev-Shwartz, S., Y. Singer, N. Srebro, and A. Cotter (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming 127*(1), 3–30.

She, Y., S. Li, and D. Wu (2016). Robust orthogonal complement principal component analysis. *Journal of the American Statistical Association 111*(514), 763–771.

She, Y. and A. B. Owen (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association 106*(494), 626–639.

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

Stange, P. (2008). On the efficient update of the singular value decomposition. In *PAMM: Proceedings in Applied Mathematics and Mechanics*, Volume 8, pp. 10827–10828. Wiley Online Library.

Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization.

submitted to SIAM Journal on Optimization 2(3).

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 448–485.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 234–251.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications 170*, 33–45.

Department of Mathematics, University of Southern California, Los Angeles, CA, 90089, U.S.A.

E-mail: minsker@usc.edu

Department of Mathematics, University of Southern California, Los Angeles, CA, 90089, U.S.A.

E-mail: langwang@usc.edu