Math. Stat. Learn. 8 (2025), 33–69
DOI 10.4171/MSL/48

Statistically optimal robust mean and covariance estimation for anisotropic Gaussians

Arshak Minasyan and Nikita Zhivotovskiy

Abstract. Assume that $X_1, \dots, X_N$ is an ε-contaminated sample of N independent Gaussian vectors in $\mathbb{R}^d$ with mean μ and covariance Σ. In the strong ε-contamination model, we assume that the adversary replaced an ε fraction of the vectors in the original Gaussian sample with arbitrary vectors. We show that there is an estimator $\widehat{\mu}$ of the mean satisfying, with probability at least $1-\delta$, a bound of the form

$$\|\widehat{\mu}-\mu\|_2 \leqslant c\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}}+\sqrt{\frac{\|\Sigma\|\log(1/\delta)}{N}}+\varepsilon\sqrt{\|\Sigma\|}\right),$$

where $c>0$ is an absolute constant and $\|\Sigma\|$ denotes the operator norm of Σ. In the same contaminated Gaussian setup, we construct an estimator $\widehat{\Sigma}$ of the covariance matrix Σ that satisfies, with probability at least $1-\delta$,

$$\|\widehat{\Sigma}-\Sigma\| \leqslant c\left(\sqrt{\frac{\|\Sigma\|\operatorname{Tr}(\Sigma)}{N}}+\|\Sigma\|\sqrt{\frac{\log(1/\delta)}{N}}+\varepsilon\|\Sigma\|\right).$$

Both results are optimal up to multiplicative constant factors. Several previously known results were either dimension-dependent and required Σ to be close to identity or had a sub-optimal dependence on the contamination level ε. As a part of the analysis, we derive sharp concentration inequalities for central order statistics of Gaussian, folded normal, and chi-squared distributions.

1. Robust multivariate mean estimation

The mean (or location parameter) estimation for contaminated Gaussian distributions is arguably one of the first questions rigorously studied in robust statistics [33]. A natural extension of this question is a problem of multivariate Gaussian mean estimation when the data are contaminated by a malicious adversary. When working with uncontaminated data, the celebrated Gaussian concentration inequality [7, 15] implies the

Mathematics Subject Classification 2020: 62F35 (primary); 62H12 (secondary).
Keywords: Gaussian mean and covariance, robust estimation, minimax optimality, PAC-Bayes, concentration for sample quantiles.

sharp non-asymptotic bound for the performance of the sample mean: if $X_1, \dots, X_N$ are independent Gaussian random vectors in $\mathbb{R}^d$ with mean μ and covariance Σ, then, with probability at least $1-\delta$,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} X_i - \mu \right\|_2 \leqslant \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{2\|\Sigma\| \log(1/\delta)}{N}}. \tag{1.1}$$

Our question is to estimate the mean μ of a Gaussian random vector when an ε-fraction of all observations is corrupted by a malicious adversary, who knows both the "clean" sample and our estimator. We focus on the multivariate case and make no assumptions on corrupted observations. This model, typically referred to as the strong contamination model [21, 23], originates from studies on replacement breakdown points [25, 29]. For an exact definition of the model, we use [16, Definition 1]. It is also known as the adversarial contamination model and includes various known contamination models, such as Huber's ε-contamination model [33]. Such a contaminated sample of Gaussian vectors will be referred to as the *ε-contaminated sample*. The sample mean can be compromised even if there is a single outlier, so we will be aiming to provide an analog of (1.1) for a different estimator. Despite recent progress in robust statistics (we refer to the surveys where both the statistical [43] and algorithmic [21] aspects are discussed in detail), there is still no sharp analog of the Gaussian bound (1.1) when the ε-strong adversarial contamination is allowed. Although the Gaussian case is historically the starting point in the theory of robust statistics, our question remains open even from an information-theoretic point of view, without considering computational aspects. Before stating our first bound, we need an additional definition. Given a covariance matrix Σ, its *effective rank* is defined as

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

where $\mathrm{Tr}(\Sigma)$ is the trace of matrix Σ. Obviously, $1 \leqslant \mathbf{r}(\Sigma) \leqslant d$ for a d-by-d covariance matrix Σ, but it can be much smaller if the distribution of the data is anisotropic and is defined by several principal directions. We are now ready to present our first bound.

Theorem 1 (Robust mean estimation in the Gaussian case). *There are absolute constants $c_1, c_2 > 0$ such that the following holds. Assume that $X_1, \dots, X_N$ is an ε-contaminated sample of Gaussian random vectors in $\mathbb{R}^d$ with mean μ and covariance Σ. Let $\varepsilon < c_1$; then, there is an estimator $\widehat{\mu} = \widehat{\mu}_{\beta,\varepsilon}(X_1, \dots, X_N)$ satisfying, with probability at least $1-\delta$,*

$$\|\widehat{\mu} - \mu\|_2 \leqslant c_2 \sqrt{\|\Sigma\|} \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \varepsilon \right).$$

Up to multiplicative constant factors, no estimator of the Gaussian mean can perform better.

Remark 1. The estimator $\widehat{\mu}$ in the above theorem depends on a tuning parameter β, described below, and the contamination level ε. While Theorem 1 does not impose any assumptions on the sample size, in Section 5, we discuss how one may select an integer β satisfying $\mathbf{r}(\Sigma)/10 \leqslant \beta \leqslant 10\mathbf{r}(\Sigma)$ based solely on the contaminated sample, under the mild condition that $N \geqslant c(\mathbf{r}(\Sigma) + \log(1/\delta))$. Notably, we avoid a sample-splitting approach when tuning this parameter.

Although the generality of our result is its main strength, we focus for a moment on the *isotropic case*, that is, $\Sigma = I_d$. In this case, if the adversary corrupts at most $O(\sqrt{dN})$ elements of the sample, we still get the optimal performance (1.1) of the sample mean in the setup where the data is not contaminated. This dependence stands in contrast with the more usual $\sqrt{\varepsilon}$-dependence on the contamination level that is known to be achievable under the two moments assumption [23, 45]. In particular, the latter results only allow $O(d)$ outliers to maintain the optimal performance. We also note that dependence on the contamination level can further be improved to $\varepsilon\sqrt{\log(1/\varepsilon)}$ under the sub-Gaussian assumption (see [45] and the discussion in what follows), but this still would not imply the desired bound even in the isotropic Gaussian case.

Another interesting aspect of our analysis is that we do not make any assumptions about the sample size. In comparison, the existing estimators that recover the optimal bound in the isotropic case [13, 21] require $N \geqslant c(d + \log(1/\delta))$, or $N \geqslant cd\varepsilon^{-2}$ as in [19], where $c > 0$ is some absolute constant. We will encounter some weaker assumptions in Section 5, but only when tuning a single real-valued parameter for our estimator.

In the context of mean estimation of anisotropic sub-Gaussian distributions, the sharpest known bound is due to Lugosi and Mendelson [45]. These authors proposed a multivariate version of a trimmed mean estimator that achieves the following error rate:

$$\|\widehat{\mu}_{\mathrm{LM}} - \mu\|_2 \leqslant c\sqrt{\|\Sigma\|}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \varepsilon\sqrt{\log(1/\varepsilon)}\right).$$

The same rate has also been provided by Dalalyan and the first author of this paper [16] for a different estimator under an additional assumption that Σ is known, through using a computationally efficient algorithm. One may think that the presence of an additional $\sqrt{\log(1/\varepsilon)}$ term is an artifact of the analysis in [16, 45]. This is in fact not true, and the presence of this term is known to be necessary for sub-Gaussian distributions [14, 45]. By trimming the observations as in [45], one can lose some of the specific properties of the Gaussian distribution. We provide an additional discussion

in Section 6, focusing on the fact that $\sqrt{\log(1/\varepsilon)}$ is inherent to the trimmed mean estimator, even in the favorable Gaussian case. Thus, our result asks for both a different estimator and a different analysis. The presence of an additional $\sqrt{\log(1/\varepsilon)}$ term is also interesting from the computational perspective. In particular, the authors of [24] argue that for any polynomial time Statistical Query algorithm the factor $\sqrt{\log(1/\varepsilon)}$ is unavoidable in the error bound (see also [32]). We note that, in the rich literature on robust statistics, a number of polynomial running time estimators were proposed with $\varepsilon\sqrt{\log(1/\varepsilon)}$ dependence on the contamination level [5, 16, 20, 21, 40].

It is worth mentioning that there are several results showing that the linear dependence on ε can be achieved in the isotropic case, where Σ is identity (or close to it). We refer to the analysis of the Tukey median, the direction-dependent median as well as the Stahel–Donoho median of means in, respectively, [13, 19, 21]. Apart from the fact that the isotropic assumption is quite restrictive, the presence of VC-type or sphere covering arguments is inherent to the analysis of existing dimension-dependent estimators. Unfortunately, these arguments cannot help us to prove the dimension-free bound of Theorem 1. It is well understood that the bound on the Gaussian complexity of ellipsoids, which corresponds to the $\sqrt{\|\Sigma\|\mathbf{r}(\Sigma)}$ term in our result, follows neither from the Dudley integral nor from VC-type arguments. This feature is especially pronounced in the Gaussian covariance estimation problem analyzed in Section 2.

The starting point of our analysis is the folklore property of the sample median of the Gaussian distribution in the one-dimensional case. We denote the sample median by $\mathrm{Med}(\cdot)$ in what follows. If $X_1, \dots, X_N$ is an ε-contaminated sample of independent standard Gaussians with mean μ and variance σ^2, then, with probability at least $1-\delta$,

$$|\mathrm{Med}(X_1, \dots, X_N) - \mu| \leqslant c\sigma\left(\sqrt{\frac{\log(1/\delta)}{N}} + \varepsilon\right),$$

whenever $N \geqslant c\log(1/\delta)$ and ε is smaller than some absolute constant. When going to higher dimensions, instead of working with Tukey's median, whose sharp analysis is only known in the isotropic case [13], or Stahel–Donoho-type estimators as in [19], we base our solution on what we call the *smoothed median estimator*.

Let $x_1, \dots, x_N$ be a set of vectors in $\mathbb{R}^d$, and let $\xi = (\xi_1, \dots, \xi_N)$ be a zero mean Gaussian random vector in $\mathbb{R}^N$ whose covariance matrix H is given by $H_{i,j} = \beta^{-1}\langle x_i, x_j\rangle$ for $i, j = 1, \dots, N$. That is, H is proportional to the Gram matrix of the original data. Here, $\beta > 0$ is any positive integer (chosen by the statistician) satisfying $\mathbf{r}(\Sigma)/10 \leqslant \beta \leqslant 10\mathbf{r}(\Sigma)$. For any direction $v \in S^{d-1}$, we are interested in the following quantity that we call the *smoothed median*:

$$\mathrm{SmoothMed}_v(x_1, \dots, x_N) = \mathbf{E}_\xi\, \mathrm{Med}(\langle x_1, v\rangle + \xi_1, \dots, \langle x_N, v\rangle + \xi_N).$$

Observe that $\mathrm{SmoothMed}_v(x_1,\dots,x_N)$ is a function of $x_1,\dots,x_N$ and v. The estimator of Theorem 1 has a simple form. Given an ε-contaminated sample $X_1,\dots,X_N$, we set

$$\widehat{\mu} = \operatorname*{arg\,min}_{\nu\in\mathbb{R}^d} \sup_{v\in S^{d-1}} |\mathrm{SmoothMed}_v(X_1,\dots,X_N) - \langle \nu, v\rangle|. \tag{1.2}$$

From the practical perspective, our estimator has complexity exponential in dimension. This is a typical limitation for all existing estimators that have a linear dependence on the contamination level ε in the Gaussian case[1]. However, in the case when the dimension d is small enough, one can replace the computation over the sphere S^{d-1} by an appropriate ε-net and approximate the smoothing integration uniformly over all the elements of this net using a Monte Carlo sampling technique.

The appearance of the smoothed median follows from the proof technique that guarantees a dimension-free nature of our bound. Our approach uses the so-called *PAC-Bayesian lemma*, whose applications were pioneered by O. Catoni and co-authors [3,11,12] in the context of mean/covariance estimation/linear regression in the heavy-tailed setup. Our application further develops these techniques but in the context of adversarial contamination, demonstrating their connections with the new class of *smoothed* empirical processes, wherein sample quantiles replace sample means. An additional discussion appears in Section 3.

Notation. Throughout the text, $c, c_1, c_2, \dots$ denote absolute constants that may change from line to line. For two positive semi-definite matrices A and B we write $A \preceq B$ if $B - A$ is positive semi-definite. The symbol $\|\cdot\|$ denotes the operator norm of a matrix or the Euclidean norm of a vector depending on the context. Let $\mathbb{S}_+^d$ denote the set of $d \times d$ positive semi-definite matrices. The symbol I_d denotes the identity $d \times d$ matrix. We denote the indicator of the event A by $\mathbb{1}(A)$. For any integer N, $[N]$ is the shortened notation of the set $\{1,\dots,N\}$. For a random variable Y and $\alpha \in [1,2]$, its ψ_α Orlicz norm is defined as follows:

$$\|Y\|_{\psi_\alpha} = \inf\{c > 0 : \mathbf{E}\exp(|Y|^\alpha/c^\alpha) \leqslant 2\}.$$

Using the standard convention, we say that $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm and $\|\cdot\|_{\psi_1}$ is the sub-exponential norm. Let

$$\mathcal{KL}(\rho,\gamma) = \int \log\Big(\frac{d\rho}{d\gamma}\Big)\, d\rho$$

[1] Recall that existing estimators of this kind lead to dimension-dependent bounds.

denote the Kullback–Leibler divergence between the two measures ρ and γ such that $\rho \ll \gamma$. The notation $\rho \ll \gamma$ means that the measure ρ is absolutely continuous with respect to the measure γ. We define the order statistics. Given a set of real numbers $x_1, \dots, x_n$, let $x_{(1)}, \dots, x_{(n)}$ denote their non-decreasing rearrangement. That is,

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}.$$

For $\alpha \in [0, 1]$, assuming that αn is an integer, we set

$$\mathrm{Quant}_\alpha(x_1, \dots, x_n) = x_{(\alpha n)}.$$

In particular, assuming for simplicity that n is odd, the sample median is given by

$$\mathrm{Med}(x_1, \dots, x_n) = x_{((n+1)/2)}.$$

Related literature. Robust statistics is a well-developed topic with several explanatory texts published recently. In our context, the most relevant are the surveys of Lugosi and Mendelson [43], and those of Diakonikolas and Kane [21]. Some classical references on robust statistics include the textbooks [30, 34, 51]. When discussing covariance estimation, we refer to the survey [35], where the focus is on heavy-tailed distributions. We also mention several recent papers on covariance estimation [1, 13, 48, 50], where the focus is on adversarial contamination.

Instead of working with the Euclidean norm as in Theorem 1, some authors focus on the Mahalanobis norm. That is, one aims to construct an estimator $\widehat{\mu}$ such that $(\widehat{\mu} - \mu)^\top \Sigma^{-1} (\widehat{\mu} - \mu)$ is small with high probability. It appears that the bounds with respect to this norm are necessarily dimension-dependent, and a simple VC-type/sphere covering argument is sufficient to obtain the optimal rates of convergence [19]. Similar observations are also valid for the covariance estimation problem. We focus on the operator norm, where the analysis allows for dimension-free bounds. This is also not the case for the (weighted) Frobenius norm commonly analyzed in the literature.

2. Covariance estimation

We now move to a more challenging problem of covariance estimation. For simplicity, we assume that our uncontaminated distribution is zero mean. We first need to present a sharp analog of inequality (1.1) in the case where no contamination is allowed. Such a result has been shown only recently by Koltchinskii and Lounici [39]. Their analysis is based on the generic chaining for quadratic processes. This non-trivial approach is motivated by the difficulty of replacing d with the effective rank $\mathbf{r}(\Sigma)$. Let us formulate their result. Assume that $Y_1, \dots, Y_N$ are independent zero mean

Gaussian vectors in $\mathbb{R}^d$ with covariance Σ. There are absolute constants $c_1, c_2 > 0$ such that, with probability at least $1 - \delta$,

$$\left\| \frac{1}{N} \sum_{i=1}^{N} Y_i Y_i^\top - \Sigma \right\| \leqslant c_1 \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} \right), \tag{2.1}$$

provided that $N \geqslant c_2(\mathbf{r}(\Sigma) + \log(1/\delta))$. When adversarial contamination is allowed, the sharpest known dimension-free result is implied by the bound of Abdalla and the second author of this paper [1]. See also the work of Oliveira and Rico [50]. These works suggest trimmed-mean-based estimators that achieve the rate

$$\|\widehat{\Sigma} - \Sigma\| \leqslant c_1 \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \varepsilon \log\left(\frac{1}{\varepsilon}\right) \right), \tag{2.2}$$

whenever $N \geqslant c_2(\mathbf{r}(\Sigma) + \log(1/\delta))$. The above bound is valid for any sub-Gaussian distribution and cannot be improved in general. Moreover, as discussed in Section 6, the estimators in [1, 50] do not account for the Gaussianity of the data, and the term $\varepsilon \log(1/\varepsilon)$ is inherent to these estimators even in the favorable Gaussian case. However, similarly to the case of mean estimation, we expect a better dependence on the contamination parameter ε in the Gaussian case. We also note that simpler versions of the bound (2.2) based on the median-of-means estimators [46] only lead to a $\sqrt{\varepsilon}$-dependence on the contamination level.

Remark 2. As a side note, when aiming for a sharp leading constant in (2.1) in the uncontaminated setup, an almost-optimal performance follows from the recent result of Han [31] combined with the second-order concentration inequality derived in [37]. A similar bound in the sub-Gaussian case with explicit constants can be found in [55].

In our analysis, we first make some additional assumptions. Our estimator depends on some parameters that could be pre-estimated based only on the observed ε-contaminated sample. A careful analysis of pre-estimation procedures is deferred to Section 5. For the rest of this section, we assume that we have access to the following quantities.

(1) There is an integer β and a real number ω satisfying, respectively,

$$\frac{1}{10}\mathbf{r}(\Sigma) \leqslant \beta \leqslant 10\mathbf{r}(\Sigma) \quad \text{and} \quad \frac{1}{10}\|\Sigma\| \leqslant \omega \leqslant 10\|\Sigma\|.$$

(2) Let H be a known positive semi-definite matrix. Assume that we know a real number $\alpha = \alpha(H)$ satisfying

$$|\alpha - \operatorname{Tr}(\Sigma H)| \leqslant c \operatorname{Tr}(\Sigma H) \left(\sqrt{\frac{\mathbf{r}(\Sigma) + \log(1/\delta)}{N}} + \varepsilon \right) \tag{2.3}$$

for some absolute constant $c > 0$.

(3) We have access to a positive semi-definite matrix G satisfying

$$\frac{1}{10}\Sigma \preceq G \quad \text{and} \quad \operatorname{Tr}(G) \leqslant 10\operatorname{Tr}(\Sigma). \tag{2.4}$$

Except for the matrix G, we only need to tune real-valued parameters. This can be usually done under the minimal assumption $N \geqslant c(\mathbf{r}(\Sigma) + \log(1/\delta))$ for some absolute constant $c > 0$. Observe that the assumption $\Sigma \preceq 10G$ does not imply that Σ is close to G in the operator norm. At the same time, this assumption requires some control over the smallest singular value of Σ. We show, in particular, that whenever $N \geqslant c(d + \log(1/\delta))$, we can always efficiently construct such a matrix G based on contaminated data, while still maintaining the dimension-free nature of our upper bound. Moreover, it appears that our guarantees are uniform with respect to the choice of the matrix G. One can rerun our estimator on the same data multiple times with any admissible G satisfying (2.4) without affecting the performance of our estimator. We discuss this formally in Section 5.

Theorem 2 (Robust covariance estimation in the Gaussian case). *There are absolute constants $c_1, c_2 > 0$ such that the following holds. Assume that $X_1, \dots, X_N$ is an ε-contaminated sample of zero mean Gaussian vectors in $\mathbb{R}^d$ with covariance Σ. Let $\varepsilon < c_1$. There is an estimator*

$$\widehat{\Sigma} = \widehat{\Sigma}_{\alpha,\beta,\omega,G,\varepsilon}(X_1, \dots, X_N)$$

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

$$\|\widehat{\Sigma} - \Sigma\| \leqslant c_2\|\Sigma\|\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \varepsilon\right).$$

Up to multiplicative constant factors, no estimator of the Gaussian covariance matrix performs better.

We are now ready to define our estimator.

We first construct the following distribution. For any $v \in S^{d-1}$—slightly abusing the notation—let ρ_v be a distribution in $\mathbb{R}^d$ whose density f_v is given by

$$f_v(x) = \frac{1}{p(2\pi\beta^{-1})^{d/2}} \exp\left(-\frac{\beta\|x - v\|^2}{2}\right)\mathbb{1}\big(\|G^{1/2}(x - v)\| \leqslant 100\sqrt{\omega}\big).$$

Here, $p > 0$ is a normalization factor. Assume that θ is a random vector distributed according to ρ_v. Let $H = \mathbf{E}_{\rho_v}(\theta - v)(\theta - v)^{\mathsf{T}}$ be a covariance matrix of ρ_v. (By the symmetry of ρ_v around v, the matrix H does not depend on v.)

For some specifically chosen absolute constant $c > 0$ and $\alpha = \alpha(H)$, define the set

$$\mathcal{H} = \Bigg\{\Gamma \in \mathbb{S}_+^d : |\operatorname{Tr}(\Gamma H) - \alpha| \leqslant c \operatorname{Tr}(\Gamma H)\left(\sqrt{\frac{\mathbf{r}(G) + \log(1/\delta)}{N}} + \varepsilon\right); \\ \Gamma \preceq 10G; \|\Gamma\| \leqslant 10\omega\Bigg\}. \tag{2.5}$$

For an ε-contaminated sample $X_1, \dots, X_N$, our estimator is defined as follows:

$$\widehat{\Sigma} = \operatorname*{arg\,min}_{\Gamma \in \mathcal{H}} \sup_{v \in S^{d-1}} \mathbf{E}_{\rho_v} \left|\operatorname{Med}(|\langle X_1, \theta\rangle|, \dots, |\langle X_N, \theta\rangle|) - \Phi^{-1}(3/4)\sqrt{\theta^{\mathsf{T}}\Gamma\theta}\right|.$$

This estimator is a more complex version of our smoothed median estimator. First, instead of working with the Gaussian smoothing measure, we restrict this distribution to an elliptic set $\{x \in \mathbb{R}^d : \|G^{1/2}(x - v)\| \leqslant 100\sqrt{\omega}\}$. Second, we need to restrict the eigenvalues of the output matrix and introduce the set $\mathcal{H}$. Our estimator is related to minimizing the so-called *median absolute deviation* (see [26] for related definitions). The proof of Theorem 2 exploits the fact that quantiles of $|\langle X, \theta\rangle|$ are tightly connected with corresponding variances. This is reflected in the term $\Phi^{-1}(3/4)$ appearing in the definition of our estimator.

3. Auxiliary results

The following section contains several technical results used throughout the paper. We start with a bound usually referred to as the *PAC-Bayesian lemma*, which is a direct consequence of the Donsker–Varadhan's variational formula for the relative entropy [27]. The proof of the next lemma and some of its applications can be found in [12, 55].

Lemma 1. *Assume that X is a random variable defined on some measurable space $\mathcal{X}$. Assume also that Θ (called the parameter space) is a subset of $\mathbb{R}^d$. Let γ be a distribution (called prior) on Θ, and let ρ be any distribution (called posterior) on Θ such that $\rho \ll \gamma$. Let $f : \mathcal{X} \times \Theta \to \mathbb{R}$ be such that $\mathbf{E}_X \exp(f(X, \theta))$ is finite γ-almost surely. Then, we have*

$$\Pr_X\big(\textit{for all } \rho \ll \gamma : \mathbf{E}_\rho f(X, \theta) \leqslant \mathbf{E}_\rho \log(\mathbf{E}_X \exp(f(X, \theta))) + \mathcal{KL}(\rho, \gamma) + t\big) \\ \geqslant 1 - e^{-t}.$$

One of the key arguments, used in several recent papers on mean and covariance estimation of heavy-tailed distributions [1, 12, 28, 50], is an application of this lemma, allowing to bypass the sphere covering and VC-type arguments. Lemma 1 will play the same key role in our analysis. However, previous applications of this lemma were based on sums of (truncated) random variables in place of $f(X, \theta)$ (in this case, X is essentially a vector of independent random variables $X_1, \dots, X_N$), while we are exploiting the interplay between Lemma 1 and sample quantiles of particular univariate distributions.

3.1. Analysis of the posterior distribution

Another technical aspect of our analysis is the introduction of truncated posterior distributions in the context of robust estimation. For a given positive semi-definite matrix G and $r \geqslant 0$, we truncate the multivariate Gaussian distribution with mean $v \in S^{d-1}$ and covariance $\beta^{-1} I_d$ as follows. Define the density function

$$f_v(x) = \frac{1}{p(2\pi\beta^{-1})^{d/2}} \exp\left(- \frac{\beta \|x - v\|^2}{2} \right) \mathbb{1}(\|G^{1/2}(x - v)\| \leqslant r), \tag{3.1}$$

where $p > 0$ is a normalization factor. We proceed with the following result.

Lemma 2 (Properties of the truncated posterior). *Let $r, \beta > 0$, and let G denote a positive semi-definite matrix in the definition (3.1). Let Σ be a covariance matrix of a zero mean random vector X in $\mathbb{R}^d$ satisfying*

$$\frac{1}{10}\Sigma \preceq G \quad \text{and} \quad \operatorname{Tr}(G) \leqslant 10 \operatorname{Tr}(\Sigma).$$

Let γ be a Gaussian measure in $\mathbb{R}^d$ with mean zero and covariance $\beta^{-1} I_d$. If, additionally,

$$r \geqslant \sqrt{20\beta^{-1} \operatorname{Tr}(\Sigma)},$$

then we have

$$\mathcal{KL}(\rho_v, \gamma) \leqslant \log(2) + \beta/2,$$

where, slightly abusing the notation, ρ_v is the distribution corresponding to the density function defined in (3.1). Furthermore, let θ be distributed according to ρ_v. Then, $\mathbf{E}_{\rho_v} \theta = v$, and almost surely with respect to the realization of θ, we have

$$\theta^{\mathsf{T}} \Sigma \theta \leqslant 2\|\Sigma\| + 20r^2.$$

Proof. We use that for θ distributed according to ρ_v it holds that $\mathbf{E}_{\rho_v} \theta = v$. This follows from the symmetry of the density around v. Let g denote the density of a Gaussian random vector with mean zero and covariance $\beta^{-1} I_d$. To control $\mathcal{KL}(\rho_v, \gamma)$, we

write

$$\begin{aligned}\int \log\left(\frac{f_v(x)}{g(x)}\right) f_v(x)dx &= \mathbf{E}_{\rho_v} \log\left(\frac{1}{p}\exp\left(\frac{-\beta\|\theta-v\|^2+\beta\|\theta\|^2}{2}\right)\right)\\ &= \log\left(\frac{1}{p}\right)+\mathbf{E}_{\rho_v}\left(\frac{-\beta\|v\|^2+2\beta\langle\theta,v\rangle}{2}\right)\\ &= \log\left(\frac{1}{p}\right)+\frac{\beta}{2}.\end{aligned}$$

To prove the desired inequality, we observe that

$$p = \Pr(\|G^{1/2}W\| \leqslant r),$$

where W is a zero mean Gaussian random vector with covariance $\beta^{-1}I_d$. Since $\operatorname{Tr}(G) \leqslant 10\operatorname{Tr}(\Sigma)$, a simple computation shows that

$$\Pr(\|G^{1/2}W\| \geqslant r) \leqslant \mathbf{E}W^{\mathsf{T}}GW/r^2 = \beta^{-1}\operatorname{Tr}(G)/r^2 \leqslant 10\beta^{-1}\operatorname{Tr}(\Sigma)/r^2 \leqslant 1/2,$$

as long as $r \geqslant \sqrt{20\beta^{-1}\operatorname{Tr}(\Sigma)}$. Under this assumption, $p \geqslant 1/2$. Therefore, we have

$$\log(1/p) \leqslant \log 2.$$

This proves the first inequality. Using the second property of the matrix G, we have

$$\begin{aligned}\theta^{\mathsf{T}}\Sigma\theta &\leqslant 2v^{\mathsf{T}}\Sigma v + 2(\theta-v)^{\mathsf{T}}\Sigma(\theta-v)\\ &\leqslant 2v^{\mathsf{T}}\Sigma v + 20(\theta-v)^{\mathsf{T}}G(\theta-v)\\ &\leqslant 2\|\Sigma\| + 20r^2.\end{aligned}$$

The claim follows. ∎

3.2. Concentration inequalities for sample quantiles

Mainly, for the purpose of completeness, we derive sub-Gaussian and sub-exponential concentration inequalities for the quantiles of i.i.d. observations sampled from several regular distributions. The analysis of sample quantiles is a standard question in statistics. The early work of Kolmogorov [38] focused on proving a central limit theorem for the sample median of some symmetric distributions. Subsequently, the focus was on explicit expansions for this limit law [10]. We additionally refer to the monograph of David and Nagaraja [17] on order statistics and to the monograph of De Haan and Ferreira [18] on the extreme value theory. Finally, many authors focused on the analysis of Bahadur's representation of sample quantiles (see, e.g., [4, 36], [52, Theorem 5.11]). Unfortunately, neither the exact expressions for the distribution of sample

quantiles nor various asymptotic expansions lead to the exact concentration inequalities we are interested in.

Less is known about concentration inequalities for sample quantiles. Some explicit non-asymptotic bounds appear in the monograph of Shao [52, Section 5.3]. Several related concentration inequalities appear in the work of Boucheron and Thomas [9], though their bounds are not sharp enough for our purposes. These authors provide a sub-exponential concentration inequality for the sample median of the Gaussian distribution, while our results will lead to sub-Gaussian concentration inequalities. Another line of results is due to Bobkov and Ledoux [6]. Their results provide sharp concentration inequalities for quantiles of log-concave distributions but only lead to sub-exponential tails due to their generality[2].

Our approach is straightforward, though, to the best of our knowledge, it is not explicit in the literature. When proving concentration inequalities for sample quantiles, we consider two regimes. For small deviations, we use the regularity of the density function and follow the reduction to a concentration of Bernoulli random variables as in [2, 13, 21, 52, 54], while for large deviations we use the sub-Gaussian/sub-exponential tails of our distribution. We discuss some straightforward extensions of our analysis in Section 6. Before providing our first concentration inequality, recall that the cumulative distribution function of a standard Gaussian is denoted by $\Phi(\cdot)$. Denote its inverse by $\Phi^{-1}(\cdot)$.

Lemma 3 (Concentration for Gaussian quantiles). *There are absolute constants c_1, $c_2 > 0$ such that the following holds. Let $\varepsilon \in [0, 1/4]$. Assume without loss of generality that $(1/2 \pm \varepsilon)N$ are integers. Let $Y_1, \dots, Y_N$ be a sample of independent standard Gaussian random variables. Then, for any $t \geqslant 0$,*

$$\Pr(|Y_{((1/2\pm\varepsilon)N)} - \Phi^{-1}(1/2 \pm \varepsilon)| \geqslant t) \leqslant 2\exp(-c_1 N t^2).$$

Equivalently,

$$\|Y_{((1/2\pm\varepsilon)N)} - \Phi^{-1}(1/2 \pm \varepsilon)\|_{\psi_2} \leqslant \frac{c_2}{\sqrt{N}}.$$

The proof of this result is deferred to Appendix. Our second result presents a similar concentration bound for the empirical quantiles of i.i.d. observations drawn from χ^2_1 distribution. This distribution coincides with the distribution of the squared standard Gaussian random variable. Denote the cumulative distribution function by $F_{\chi^2_1}(\cdot)$ and its inverse by $F^{-1}_{\chi^2_1}(\cdot)$. The key difference is that we only show a sub-exponential tail in this case. We remark that when considering the χ^2_k distribution

[2]For the special case of the uniform distribution in $[0, 1]$, Bobkov and Ledoux [6] provide a sub-Gaussian concentration inequality for all order statistics.

with $k \geqslant 2$ degrees of freedom, the desired concentration inequality follows from log-concavity and [6, Lemma 6.5].

Lemma 4 (Quantiles of the χ_1^2 distribution). *There is an absolute constant $c_1 > 0$ such that the following holds. Assume without loss of generality that $(1/2 \pm \varepsilon)N$ are integers. Let $Y_1, \dots, Y_N$ be a sample of independent χ_1^2 random variables and $\varepsilon \in [0, 1/4]$. Then,*

$$\|Y_{((1/2\pm\varepsilon)N)} - F_{\chi_1^2}^{-1}(1/2 \pm \varepsilon)\|_{\psi_1} \leqslant \frac{c_1}{\sqrt{N}}.$$

The proof of this result is also deferred to Appendix B. Our final result proves a similar bound for the (standard) half-normal distribution. Namely, we want to prove the concentration inequality for quantiles of the absolute values of standard Gaussian random variables. Denote the cumulative distribution function of this distribution by $\Phi_{\mathrm{H}}(\cdot)$ and its inverse by $\Phi_{\mathrm{H}}^{-1}(\cdot)$.

Lemma 5 (Quantiles of the half-normal distribution). *There is an absolute constant c_1 such that the following holds. Assume without loss of generality that $(1/2 \pm \varepsilon)N$ are integers. Let $Y_1, \dots, Y_N$ be a sample of independent half-normal random variables and $\varepsilon \in [0, 1/4]$. Then,*

$$\|Y_{((1/2\pm\varepsilon)N)} - \Phi_{\mathrm{H}}^{-1}(1/2 \pm \varepsilon)\|_{\psi_2} \leqslant \frac{c_1}{\sqrt{N}}.$$

The proof of this result repeats the same computations used in the proofs of Lemmas 3 and 4. We omit the details.

4. Proofs of the main results

We begin with the proof of our first main result that yields that the estimator defined in (1.2) achieves an optimal error bound for the robust mean estimation problem. We discuss the optimality of our results at the end of this section.

Proof of Theorem 1. First, immediately by the definition of our estimator and basic properties of the multivariate Gaussian distribution, we have

$$\widehat{\mu} = \operatorname*{arg\,min}_{v\in\mathbb{R}^d} \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v} \operatorname{Med}(\langle X_1, \theta\rangle, \dots, \langle X_N, \theta\rangle) - \langle v, v\rangle|,$$

where ρ_v is a multivariate Gaussian distribution in $\mathbb{R}^d$ with mean v and covariance $\beta^{-1} I_d$ (we are slightly abusing the notation, since ρ_v defined a clipped multivariate Gaussian distribution in the estimator of Theorem 2), and the expectation $\mathbf{E}_{\rho_v}$ is taken

with respect to θ just as in the formulation of Lemma 1. By the triangle inequality, the definition of our estimator, and the fact that $\mathbf{E}_{\rho_v}\theta = v$, we have

$$\begin{aligned}\|\widehat{\mu}-\mu\|_2 &= \sup_{v\in S^{d-1}} \langle \widehat{\mu}-\mu, v\rangle \\ &\leqslant \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v} \operatorname{Med}(\langle X_1,\theta\rangle,\dots,\langle X_N,\theta\rangle) - \langle \widehat{\mu}, v\rangle| \\ &\quad + \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v} \operatorname{Med}(\langle X_1,\theta\rangle,\dots,\langle X_N,\theta\rangle) - \langle \mu, v\rangle| \\ &\leqslant 2 \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v} \operatorname{Med}(\langle X_1-\mu,\theta\rangle,\dots,\langle X_N-\mu,\theta\rangle)|. \qquad (4.1)\end{aligned}$$

We need to bound the last quantity. From now on, we can assume without loss of generality that $\mu = 0$. Assume that $Y_1,\dots,Y_N$ is an uncontaminated sample of zero mean independent Gaussian vectors with covariance Σ. That is, at most εN elements among $X_1,\dots,X_N$ are different from their $Y_1,\dots,Y_N$ counterparts. Observe that the sample median of projections of the contaminated sample in any direction cannot be too far away from $1/2\pm\varepsilon$ quantiles of the corresponding projections for the uncontaminated sample. Formally, assuming that both the sample median and $1/2\pm\varepsilon$ sample quantiles are unique, we have ρ_v-almost surely

$$\begin{aligned}\operatorname{Quant}_{\frac12-\varepsilon}(\langle Y_1,\theta\rangle,\dots,\langle Y_N,\theta\rangle) &\leqslant \operatorname{Med}(\langle X_1,\theta\rangle,\dots,\langle X_N,\theta\rangle) \\ &\leqslant \operatorname{Quant}_{\frac12+\varepsilon}(\langle Y_1,\theta\rangle,\dots,\langle Y_N,\theta\rangle),\end{aligned}$$

and thus, taking the expectation with respect to ρ_v, we readily get the following bound, for any $v\in S^{d-1}$

$$\begin{aligned}|\mathbf{E}_{\rho_v} \operatorname{Med}(\langle X_1,\theta\rangle,\dots,\langle X_N,\theta\rangle)| &\leqslant |\mathbf{E}_{\rho_v} \operatorname{Quant}_{\frac12+\varepsilon}(\langle Y_1,\theta\rangle,\dots,\langle Y_N,\theta\rangle)| \\ &\quad + |\mathbf{E}_{\rho_v} \operatorname{Quant}_{\frac12-\varepsilon}(\langle Y_1,\theta\rangle,\dots,\langle Y_N,\theta\rangle)|.\end{aligned}$$

Both terms will be analyzed similarly. We only analyze the first one. Observe that, due to the spherical symmetry, we have that $S_N = \{\langle Y_1,\theta\rangle/\sqrt{\theta^\top\Sigma\theta},\dots,\langle Y_N,\theta\rangle/\sqrt{\theta^\top\Sigma\theta}\}$ (note that S_N depends on θ, but we omit the explicit dependence for brevity in the notation) consists of independent standard Gaussian random variables (in our case, $\theta\neq 0$ almost surely). We have

$$\begin{aligned}&|\mathbf{E}_{\rho_v} \operatorname{Quant}_{\frac12+\varepsilon}(\langle Y_1,\theta\rangle,\dots,\langle Y_N,\theta\rangle)| \\ &\quad\leqslant \left|\mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac12+\varepsilon}(S_N) - \mathbf{E}\operatorname{Quant}_{\frac12+\varepsilon}(S_N))\right| \\ &\qquad + \left|\mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}\cdot(\mathbf{E}\operatorname{Quant}_{\frac12+\varepsilon}(S_N) - \Phi^{-1}(1/2+\varepsilon))\right| \\ &\qquad + \mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}\cdot\Phi^{-1}(1/2+\varepsilon) \\ &\quad= \text{(I)} + \text{(II)} + \text{(III)},\end{aligned}$$

where $\mathbf{E}$ should be understood as the expectation with respect to both the observations $Y_1, \dots, Y_N$ and the distribution ρ_v, since θ is distributed according to ρ_v. We now bound each of the three terms (I), (II), (III) above separately.

First term. The first term $|\mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta}(\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \mathbf{E}\,\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N))|$ concerns the distance between the empirical quantile of standard Gaussians and its expectation. To bound this quantity, we apply Lemma 1. By Lemma 6 from Appendix A.1 we have

$$\left|\mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta}(\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \mathbf{E}\,\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N))\right| \leqslant c_3 \sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\| \log(1/\delta)}{N}}$$

with probability at least $1 - \delta$, uniformly over S^{d-1}, where c_3 is some positive absolute constant.

Second term. The second term $|\mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta} \cdot (\mathbf{E}\,\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \Phi^{-1}(1/2+\varepsilon))|$ is the distance between the expected quantile and its theoretical counterpart, which can be bounded using the observation that, for any scalar C, we have $\|C\|_{\psi_2} = |C|/\sqrt{\log 2}$. Using this observation, together with Jensen's inequality and Lemma 3, we have, for some $c_4 > 0$,

$$\begin{aligned}\left|\mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta} \cdot (\mathbf{E}\,\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \Phi^{-1}(1/2+\varepsilon))\right| &\leqslant \frac{c_4 \sqrt{\log 2} \cdot \mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta}}{\sqrt{N}} \\ &\leqslant \frac{c_4 \sqrt{\log 2} \cdot \sqrt{11 \|\Sigma\|}}{\sqrt{N}}.\end{aligned}$$

Third term. The third term is the magnitude of the standard Gaussian quantile computed around $1/2$. First, we notice that the function $\Phi^{-1}(\cdot)$ is locally Lipschitz on a closed interval $[1/2 - \varepsilon, 1/2 + \varepsilon]$ for $\varepsilon \in [0, 1/4]$. Hence, by bounding the Lipschitz constant, we arrive at the following inequality, $|\Phi^{-1}(1/2+\varepsilon)| = |\Phi^{-1}(1/2+\varepsilon) - \Phi^{-1}(1/2)| \leqslant 4\varepsilon$, for all $\varepsilon \in [0, 1/4]$. Therefore, we have

$$\mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta} \cdot \Phi^{-1}(1/2+\varepsilon) \leqslant 4\varepsilon \sqrt{11 \|\Sigma\|}.$$

Combining the results of the three terms bounded above, the relation between the median of the contaminated sample and the quantiles of an uncontaminated Gaussian sample, as well as the inequality (4.1), concludes the proof. ∎

We are now ready to prove our second main result.

Proof of Theorem 2. Slightly abusing the notation, we now use ρ_v to denote the measure in the definition of the covariance estimator of Theorem 2. Recall that H is a covariance matrix of ρ_v and does not depend on a direction $v \in S^{d-1}$. Observe

that $\mathbf{E}_{\rho_v}\theta^\mathsf{T}\Sigma\theta = v^\mathsf{T}\Sigma v + \operatorname{Tr}(\Sigma H)$. Moreover, since our choice of parameters implies $p \geqslant 1/2$ in (3.1), we have

$$H \preceq 2\beta^{-1} I_d \quad \text{and} \quad \|H\| \leqslant 2\beta^{-1}.$$

We also observe that $\mathbf{r}(G) = \operatorname{Tr}(G)/\|G\| \leqslant 100\mathbf{r}(\Sigma)$. Using the triangle inequality, as well as the definition of our estimator combined with the definition of the set $\mathscr{H}$ from (2.5), we have

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\| &= \sup_{v\in S^{d-1}} |v^\mathsf{T}\Sigma v - v^\mathsf{T}\widehat{\Sigma} v| \\
&\leqslant \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v}\theta^\mathsf{T}(\Sigma - \widehat{\Sigma})\theta| \\
&\quad + |(\operatorname{Tr}(\widehat{\Sigma}H) - \alpha(H)) - (\operatorname{Tr}(\Sigma H) - \alpha(H))| \\
&\leqslant \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v}\theta^\mathsf{T}(\Sigma - \widehat{\Sigma})\theta| \\
&\quad + c\operatorname{Tr}((\Sigma + \widehat{\Sigma})H)\left(\sqrt{\frac{\max\{\mathbf{r}(G), \mathbf{r}(\Sigma)\} + \log(1/\delta)}{N}} + \varepsilon\right) \\
&\leqslant \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v}\theta^\mathsf{T}(\Sigma - \widehat{\Sigma})\theta| \\
&\quad + 2\beta^{-1}c\operatorname{Tr}(\Sigma + 10G)\left(\sqrt{\frac{100\mathbf{r}(\Sigma) + \log(1/\delta)}{N}} + \varepsilon\right) \\
&\leqslant \sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v}\theta^\mathsf{T}(\Sigma - \widehat{\Sigma})\theta| \\
&\quad + 202\beta^{-1}c\operatorname{Tr}(\Sigma)\left(\sqrt{\frac{100\mathbf{r}(\Sigma) + \log(1/\delta)}{N}} + \varepsilon\right).
\end{aligned}$$

Since $\beta^{-1} \leqslant 10\mathbf{r}(\Sigma)$, the last term in the last inequality is not larger than the rate of convergence in the statement of Theorem 2. We now can focus only on bounding the first term in the last line of the inequalities from the previous display. We first need some auxiliary computations. Using the definition of the set $\mathscr{H}$, we have ρ_v-almost surely

$$\begin{aligned}
&\sqrt{\theta^\mathsf{T}\Sigma\theta} + \sqrt{\theta^\mathsf{T}\widehat{\Sigma}\theta} \\
&\quad \leqslant \sqrt{2v^\mathsf{T}\Sigma v + 2(\theta - v)^\mathsf{T}\Sigma(\theta - v)} + \sqrt{2v^\mathsf{T}\widehat{\Sigma} v + 2(\theta - v)^\mathsf{T}\widehat{\Sigma}(\theta - v)} \\
&\quad \leqslant \sqrt{2\|\Sigma\| + 20(\theta - v)^\mathsf{T}G(\theta - v)} + \sqrt{20\omega + 20(\theta - v)^\mathsf{T}G(\theta - v)} \\
&\quad \leqslant \sqrt{2\|\Sigma\| + 20r^2} + \sqrt{20\omega + 20r^2} \\
&\quad \leqslant c_1\sqrt{\|\Sigma\|},
\end{aligned}$$

where $c_1 > 0$ is some absolute constant. This implies the following lines:

$$
\begin{aligned}
&\sup_{v\in S^{d-1}} |\mathbf{E}_{\rho_v}\theta^\top(\Sigma-\widehat{\Sigma})\theta| \\
&\leqslant \sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\sqrt{\theta^\top\Sigma\theta}-\sqrt{\theta^\top\widehat{\Sigma}\theta}\big|\big|\sqrt{\theta^\top\Sigma\theta}+\sqrt{\theta^\top\widehat{\Sigma}\theta}\big| \\
&\leqslant c_1\sqrt{\|\Sigma\|}\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\sqrt{\theta^\top\Sigma\theta}-\sqrt{\theta^\top\widehat{\Sigma}\theta}\big| \\
&\leqslant c_1\sqrt{\|\Sigma\|}\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\mathrm{Med}(|\langle X_1,\theta\rangle|,\dots,|\langle X_N,\theta\rangle|)/(\Phi^{-1}(3/4))-\sqrt{\theta^\top\widehat{\Sigma}\theta}\big| \\
&\quad + c_1\sqrt{\|\Sigma\|}\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\mathrm{Med}(|\langle X_1,\theta\rangle|,\dots,|\langle X_N,\theta\rangle|)/(\Phi^{-1}(3/4))-\sqrt{\theta^\top\Sigma\theta}\big| \\
&\leqslant 2c_1\sqrt{\|\Sigma\|}\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\mathrm{Med}(|\langle X_1,\theta\rangle|,\dots,|\langle X_N,\theta\rangle|)/(\Phi^{-1}(3/4))-\sqrt{\theta^\top\Sigma\theta}\big|,
\end{aligned}
$$

where in the last line we used the definition of $\widehat{\Sigma}$ and that $\Sigma\in\mathcal{H}$. We focus on upper bounding the last expression. Let $Y_1,\dots,Y_N$ be the uncontaminated version of our ε-contaminated sample. Using the same argument as in the proof of Theorem 1, we have

$$
\begin{aligned}
&\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\mathrm{Med}(|\langle X_1,\theta\rangle|,\dots,|\langle X_N,\theta\rangle|)-\Phi^{-1}(3/4)\sqrt{\theta^\top\Sigma\theta}\big| \\
&\leqslant \sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\mathrm{Quant}_{\frac12+\varepsilon}(|\langle Y_1,\theta\rangle|,\dots,|\langle Y_N,\theta\rangle|)-\Phi^{-1}(3/4)\sqrt{\theta^\top\Sigma\theta}\big| \\
&\quad + \sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\big|\mathrm{Quant}_{\frac12-\varepsilon}(|\langle Y_1,\theta\rangle|,\dots,|\langle Y_N,\theta\rangle|)-\Phi^{-1}(3/4)\sqrt{\theta^\top\Sigma\theta}\big|.
\end{aligned}
$$

We only analyze the first term. Observe that, due to the spherical symmetry, we have that $S_N = \{|\langle Y_1,\theta\rangle|/\sqrt{\theta^\top\Sigma\theta},\dots,|\langle Y_N,\theta\rangle|/\sqrt{\theta^\top\Sigma\theta}\}$ (here we slightly abuse the notation and use the symbol S_N in a different context) consists of independent half-normal random variables (in our case, $\theta\neq 0$ almost surely). By the triangle inequality, we have

$$
\begin{aligned}
&\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}\,|\mathrm{Quant}_{\frac12+\varepsilon}(S_N)-\Phi^{-1}(3/4)| \\
&\quad\leqslant \sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta} \\
&\qquad\times\big(|\mathrm{Quant}_{\frac12+\varepsilon}(S_N)-\Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)|-\mathbf{E}|\mathrm{Quant}_{\frac12+\varepsilon}(S_N)-\Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)|\big) \\
&\qquad+\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\mathbf{E}\sqrt{\theta^\top\Sigma\theta}\,|\mathrm{Quant}_{\frac12+\varepsilon}(S_N)-\Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)| \\
&\qquad+\sup_{v\in S^{d-1}} \mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}\,|\Phi^{-1}(3/4)-\Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)| \\
&\quad= \mathrm{(I)}+\mathrm{(II)}+\mathrm{(III)}.
\end{aligned}
$$

We want to apply Lemma 1 to control (I). Fix $\lambda > 0$ and let γ be a multivariate Gaussian distribution in $\mathbb{R}^d$ with zero mean and covariance $\beta^{-1} I_d$. Lemma 2 implies that for our choice of parameters $\mathcal{KL}(\rho_v, \gamma) \leqslant \log(2) + \beta/2$. Denote

$$Q(S_N) = |\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)| - \mathbf{E}|\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)|.$$

Observe that, conditioned on θ, the random variable $Q(S_N)$ is a centered version of the random variable $|\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)|$ whose $\|\cdot\|_{\psi_2}$ is controlled by Lemma 5. Since centering multiplies the ψ_2-norm by at most an absolute constant factor, we have (conditioned on θ) that $\|Q(S_N)\|_{\psi_2} \leqslant \frac{c_2}{\sqrt{N}}$ for some absolute constant $c_2 > 0$. By Lemma 1 we have with probability at least $1-\delta$, simultaneously for all $v \in S^{d-1}$,

$$\lambda \mathbf{E}_{\rho_v} \sqrt{\theta^\top \Sigma \theta}\; Q(S_N) \leqslant \mathbf{E}_{\rho_v} \log \mathbf{E} \exp\left(\lambda \sqrt{\theta^\top \Sigma \theta}\; Q(S_N)\right) + \beta/2 + \log(2/\delta).$$

Thus, by [53, Proposition 2.5.2 (v)] (conditioned on θ, we take $\lambda\sqrt{\theta^\top\Sigma\theta}$ instead of λ in that result), repeating the lines of the proof of Theorem 1, we have, for some absolute constants $c_3, c_4 > 0$,

$$\mathbf{E}_{\rho_v} \log \mathbf{E} \exp\left(\lambda \sqrt{\theta^\top \Sigma \theta}\; Q(S_N)\right) \leqslant \mathbf{E}_{\rho_v} \frac{c_3 \lambda^2 \theta^\top \Sigma \theta}{N} \leqslant \frac{c_4 \lambda^2 \|\Sigma\|}{N}.$$

Combining the bounds and optimizing with respect to λ, we have simultaneously, for all $v \in S^{d-1}$, with probability at least $1-\delta$,

$$\text{(I)} \leqslant c_5 \sqrt{\|\Sigma\|}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right),$$

where $c_5 > 0$ is some absolute constant. We now bound the term (II). Similarly to the proof of Theorem 1, we use Lemma 5 to get, for some absolute constant $c_6 > 0$, the following bound:

$$\mathbf{E}_{\rho_v} \mathbf{E} \sqrt{\theta^\top \Sigma \theta} |\mathrm{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon)| \leqslant c_6 \sqrt{\frac{\|\Sigma\|}{N}}.$$

To bound (III) we first observe that $\Phi_{\mathrm{H}}^{-1}(1/2) = \Phi^{-1}(3/4)$. Now, we show that the difference $\Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon) - \Phi_{\mathrm{H}}^{-1}(1/2)$ is bounded by ε (up to multiplicative constant) for $\varepsilon \in [0, 1/4]$. Similarly to the arguments used in the proof of Theorem 1 for the quantile function of standard Gaussian distribution, we compute and bound the derivative of $\Phi_{\mathrm{H}}^{-1}(1/2+x)$ when $x \in [0, 1/4]$ as follows:

$$\begin{aligned}\frac{d}{dx}\Phi_{\mathrm{H}}^{-1}(1/2+x) &= \sqrt{\frac{\pi}{2}} \exp((\Phi_{\mathrm{H}}^{-1}(1/2+x))^2/2) \\ &\leqslant \sqrt{\frac{\pi}{2}} \exp((\Phi_{\mathrm{H}}^{-1}(3/4))^2/2) \leqslant 3.\end{aligned}$$

Therefore, we have, for some $c_7 > 0$,

$$\mathbf{E}_{\rho_v} \sqrt{\theta^\intercal \Sigma \theta} \cdot |\Phi_{\mathrm{H}}^{-1}(1/2 + \varepsilon) - \Phi_{\mathrm{H}}^{-1}(1/2)| \leqslant c_7 \varepsilon \sqrt{\|\Sigma\|}.$$

Combining the obtained bounds, we complete the proof. ∎

Statistical optimality of our estimators. We shortly discuss the claimed optimality of our bounds. The optimality results follow immediately from existing lower bounds. The bounds in [13, Theorems 2.2 and 3.2] show that Theorems 1 and 2 both have the optimal dependence on the contamination level with correct dimension-free parametric rate. For covariance estimation, the optimality of the remaining terms is described in detail in [1, Section 5]. Matching lower bounds for the mean estimation problem are shown in [44].

5. Tuning the unknown parameters

Our focus is now on tuning a few parameters used in our estimators. For the sake of simplicity, we assume that either ε is known exactly or a known upper bound ε_0 is available such that $\varepsilon \leqslant \varepsilon_0 < 1/2$. This is a standard assumption in the literature [45]. Observe that at least in mean estimation the value of ε is only used to tune the parameter β. The most standard approach to estimating other parameters is the sample-splitting idea. One splits the sample into several independent blocks of equal sizes. For each block, we can bound the number of contaminated points. This will allow us to state our result for any $\varepsilon \in [0, c]$, where c is some small enough absolute constant. An interesting aspect of our analysis is that we can tune different parameters on the same sample. We will now discuss this in detail.

Handling the dependencies. It is clear that in the strong contamination setup, the adversary can make the aforementioned blocks dependent. That is, the outliers in any sub-sample may depend on the entire sample. Some authors assume implicitly that the splitting of the sample results in independent subsamples. For example, the analysis of the trimmed-mean estimator in [45, Theorem 1] uses this independence, which holds, for example, in Huber's contamination model but is not true in the general strong contamination model. Taking care of the sample splitting step in the strong contamination model requires some additional stability-type analysis. We refer to [22, Section 6] as an example of this approach.

We now show that our approach allows one to tune the parameters on the same sample. Thus, our result is valid in the strong contamination model without additional assumptions. For clarity, we only focus on the mean estimation problem. Assume that we are given an ε-contaminated sample of size N. We denote it by S_N. Given S_N, we

first find an integer $\beta = \beta(S_N)$ satisfying, with probability at least $1 - \delta/2$,

$$\mathbf{r}(\Sigma)/10 \leqslant \beta(S_N) \leqslant 10\mathbf{r}(\Sigma). \tag{5.1}$$

We then compute our estimator defined in (1.2) on the same sample S_N with $\beta = \beta(S_N)$. Denote the event where (5.1) holds by E. We show that, due to the nature of Lemma 1, this dependence does not lead to additional technical issues. First, observe that since β is an integer, we can use the union bound over at most $10\mathbf{r}(\Sigma)$ prior Gaussian distributions γ to handle potential dependence of β on S_N. One can verify that this application of the union bound does not change the bound of Theorem 1. Importantly, the result of Lemma 1 is uniform with respect to the posterior distribution ρ_v and allows β to depend on the sample as long as $\mathcal{KL}(\rho_v, \gamma) = \beta(S_N)/2 \leqslant 5\mathbf{r}(\Sigma)$, which holds on the event E. Finally, one can easily verify that, on the same event E, the desired upper bound on the term

$$\mathbf{E}_{\rho_v} \log \mathbf{E} \exp\left(\lambda \sqrt{\theta^\top \Sigma \theta}(\mathrm{Quant}_{\frac{1}{2}\pm\varepsilon}(S_N) - \mathbf{E}\,\mathrm{Quant}_{\frac{1}{2}\pm\varepsilon}(S_N))\right),$$

appearing in the proof of Theorem 1, is not affected by the fact that $\beta = \beta(S_N)$. This argument allows us to use $\beta(S_N)$ in our estimator.

Similar ideas can also be applied in the covariance estimation setup. To avoid unnecessary technicalities, we assume that for covariance estimation we can indeed split the sample into several blocks and the adversary is not allowed to create dependencies between these blocks. This covers many standard contamination models, including Huber's ε-contamination model.

Estimating $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. This step follows from existing results. In particular, in the Gaussian case, [1, Proposition 6] provides an estimator ω satisfying $\|\Sigma\|/4 \leqslant \omega \leqslant 4\|\Sigma\|$ whenever $N \geqslant c(\mathbf{r}(\Sigma) + \log(1/\delta))$, where $c > 0$ is an absolute constant. We also need to estimate $\mathrm{Tr}(\Sigma)$. This problem reduces to mean estimation. The linear dependence on ε will not play any role since we only need to know $\mathrm{Tr}(\Sigma)$ up to a multiplicative constant factor. In particular, one can use any *sub-Gaussian mean estimator* in $\mathbb{R}$ (see [43] for the exact definition) that is tolerant to strong contamination and gives a $\sqrt{\varepsilon}$-dependence on the contamination level to find τ satisfying $\mathrm{Tr}(\Sigma)/2 \leqslant \tau \leqslant 2\,\mathrm{Tr}(\Sigma)$, whenever ε is small enough and $N \geqslant c\log(1/\delta)$. This allows us to find an integer β satisfying (5.1).

Constructing the matrix $\boldsymbol{G}$. We discuss how to construct a positive semi-definite matrix G, satisfying

$$\Sigma \preceq 10G \quad \text{and} \quad \mathrm{Tr}(G) \leqslant 10\,\mathrm{Tr}(\Sigma). \tag{5.2}$$

The following result allows us to construct such a matrix efficiently whenever $N \geqslant c(d + \log(1/\delta))$, where $c > 0$ is some absolute constant.

Proposition 1. *There are absolute constants $c, c_1 > 0$ such that the following holds. Assume that X is a zero mean Gaussian vector in $\mathbb{R}^d$ with covariance Σ. Let $X_1, \dots, X_N$ be an ε-contaminated set of independent copies of X. Fix $\delta \in (0,1)$. Assume that $\varepsilon \leqslant c$ and $N \geqslant c_1(d + \log(1/\delta))$. Then, with probability at least $1-\delta$, simultaneously for all $I' \subseteq [N]$ such that $|I'| = cN$, we have*

$$\Sigma \preceq \frac{10}{N} \sum_{i \in [N] \setminus I'} X_i X_i^\top.$$

Moreover, on the same event, there exists $I \subseteq [N]$ such that $|I| = cN$, and

$$\frac{1}{N} \sum_{i \in [N] \setminus I} \|X_i\|^2 \leqslant 10 \operatorname{Tr}(\Sigma).$$

This result implies immediately that the matrix

$$G = \frac{1}{N} \sum_{i \in [N] \setminus I} X_i X_i^\top$$

satisfies the desired property (5.2). In order to find this set, one only needs to find a set I of size εN such that $\sum_{i \in [N] \setminus I} \|X_i\|^2 \leqslant 10N \operatorname{Tr}(\Sigma)$. This can be done simply by removing the εN vectors with the largest norms.

Proof. Without loss of generality, we assume that cN is an integer. Fix any $I \subset [N]$ of size $2cN$. Let $Y_1, \dots, Y_N$ denote an uncontaminated sample. The total number of such subsets is upper bounded by $\binom{N}{2cN} \leqslant (2e/c)^{2cN}$. By the bound of Oliveira [49, Theorem 4.1 with h = 3] and the union bound over all sets I of size $2cN$, we have

$$\Sigma \left(1 - 27 \sqrt{\frac{d + 4cN \log(2e/c) + 2\log(2/\delta)}{N - 2cN}} \right) \preceq \frac{1}{N - 2cN} \sum_{i \in [N] \setminus I} Y_i Y_i^\top.$$

When c is small enough and $N \geqslant c_1(d + \log(1/\delta))$ for large enough $c_1 > 0$, on the same event, we have

$$\Sigma \preceq \frac{10}{N} \sum_{i \in [N] \setminus I} Y_i Y_i^\top.$$

Observe that since each term $Y_i Y_i^\top$ is a positive semi-definite matrix and $\varepsilon \leqslant c$, we have that, for any I' of size cN, there is a set I of size $2cN$ such that

$$\sum_{i \in [N] \setminus I} Y_i Y_i^\top \preceq \sum_{i \in [N] \setminus I'} X_i X_i^\top.$$

Indeed, to build such a set I, we consider the union of the set of contaminated points with the set I'. (We can add any additional elements if the cardinality of this union is

less than $2cN$.) This implies that, under our assumption for all $I' \subset [N]$ of size cN, with probability at least $1-\delta$,

$$\Sigma \preceq \frac{10}{N} \sum_{i \in [N] \setminus I'} X_i X_i^{\mathsf{T}}.$$

We now consider the second part of the statement. Combining the Gaussian concentration inequality [8, Example 5.7] and [53, Proposition 2.5.2], we get that there is an absolute constant $c_2 > 0$ such that

$$\| \|X\| - \mathbf{E}\|X\| \|_{\psi_2} \leqslant c_2 \sqrt{\|\Sigma\|}.$$

It is now standard to verify that $\| \|X\|^2 - \mathbf{E}\|X\|^2 \|_{\psi_1} \leqslant c_3 \|\Sigma\|$, where $c_3 > 0$ is an absolute constant. By the Bernstein inequality [53, Theorem 2.8.1] and the union bound, simultaneously for all $I \subset [N]$, $|I| = cN$, with probability at least $1-\delta$, it holds for some absolute constant $c_4 > 0$ that

$$\begin{aligned}
&\sum_{i \in [N] \setminus I} \|Y_i\|^2 \\
&\quad \leqslant N \operatorname{Tr}(\Sigma) + c_4 \|\Sigma\| \Big(\sqrt{N(\log(1/\delta) + cN \log(e/c))} + \log(1/\delta) + cN \log(e/c) \Big) \\
&\quad \leqslant 10N \operatorname{Tr}(\Sigma).
\end{aligned}$$

The last inequality holds provided that c is small enough and c_1 is large enough. We choose I to be the set corresponding to the set of contaminated points. For this set I, on the same event, we have

$$\frac{1}{N} \sum_{i \in [N] \setminus I} \|X_i\|^2 = \frac{1}{N} \sum_{i \in [N] \setminus I} \|Y_i\|^2 \leqslant 10 \operatorname{Tr}(\Sigma).$$

The claim follows by the union bound. ∎

Estimating $\alpha(H)$. We conclude by the analysis of a real-valued parameter $\alpha = \alpha(H)$, defined in (2.3). In what follows, H is a known positive semi-definite matrix. When allowing slightly sub-optimal dependence on ε, we can use the analysis of the trimmed mean estimator in $\mathbb{R}$ (see [45, Theorem 1]). Unfortunately, the analysis becomes more complicated when the linear dependence on the contamination level is of interest. Recall that we are interested in finding $\alpha = \alpha(H)$ such that, with probability at least $1-\delta$,

$$|\alpha - \operatorname{Tr}(\Sigma H)| \leqslant c \operatorname{Tr}(\Sigma H) \left(\sqrt{\frac{\mathbf{r}(\Sigma) + \log(1/\delta)}{N}} + \varepsilon \right).$$

We present an estimator that achieves this error rate in almost any interesting regime. More precisely, we will either make an additional assumption that $\delta \geqslant \exp(-\sqrt{\mathbf{r}(\Sigma)})$, or that $\log d \leqslant \mathbf{r}(\Sigma)$. In what follows, $e_1, \ldots, e_d$ denote the standard basis in $\mathbb{R}^d$.

Proposition 2. *There are absolute constants $c, c_1, c_2 > 0$ such that the following holds. Assume that X is a Gaussian zero mean vector in $\mathbb{R}^d$ with covariance Σ. Let $X_1, \dots, X_N$ be an ε-contaminated set of independent copies of X. Fix $\delta \in (0,1)$. Assume that $\varepsilon \leqslant c$ and $N \geqslant c_1 \log(1/\delta)$. Then, with probability at least $1-\delta$, it holds*

$$\left| (\Phi^{-1}(3/4))^{-2} \sum_{i=1}^{d} \operatorname{Med}(\langle e_i, H^{1/2} X_1\rangle^2, \dots, \langle e_i, H^{1/2} X_N\rangle^2) - \operatorname{Tr}(\Sigma H) \right| \leqslant c_1 \operatorname{Tr}(\Sigma H) \left(\frac{\log(1/\delta)}{\sqrt{N}} + \varepsilon \right). \tag{5.3}$$

If, additionally, $N \geqslant c_2(\log d + \log(1/\delta))$, then, on the same event, it holds

$$\left| (\Phi^{-1}(3/4))^{-2} \sum_{i=1}^{d} \operatorname{Med}(\langle e_i, H^{1/2} X_1\rangle^2, \dots, \langle e_i, H^{1/2} X_N\rangle^2) - \operatorname{Tr}(\Sigma H) \right| \leqslant c_1 \operatorname{Tr}(\Sigma H) \left(\sqrt{\frac{\log d + \log(1/\delta)}{N}} + \varepsilon \right).$$

Proof. Let $Y_1, \dots, Y_N$ denote the uncontaminated sample, and let Y be a zero mean Gaussian in $\mathbb{R}^d$ with covariance Σ. Since $\operatorname{Tr}(\Sigma H) = \operatorname{Tr}(H^{1/2}\Sigma H^{1/2})$, by triangle inequality and the arguments of the proof of Theorem 1, we have

$$\begin{aligned}
&\left| \sum_{i=1}^{d} \operatorname{Med}(\langle e_i, H^{1/2} X_1\rangle^2, \dots, \langle e_i, H^{1/2} X_N\rangle^2) - (\Phi^{-1}(3/4))^2 \operatorname{Tr}(\Sigma H) \right| \\
&\quad \leqslant \sum_{i=1}^{d} |\operatorname{Med}(\langle e_i, H^{1/2} X_1\rangle^2, \dots, \langle e_i, H^{1/2} X_N\rangle^2) \\
&\qquad\qquad - (\Phi^{-1}(3/4))^2 \|\Sigma^{1/2} H^{1/2} e_i\|^2| \\
&\quad \leqslant \sum_{i=1}^{d} |\operatorname{Quant}_{1/2+\varepsilon}(\langle e_i, H^{1/2} Y_1\rangle^2, \dots, \langle e_i, H^{1/2} Y_N\rangle^2) \\
&\qquad\qquad - (\Phi^{-1}(3/4))^2 \|\Sigma^{1/2} H^{1/2} e_i\|^2| \\
&\qquad + \sum_{i=1}^{d} |\operatorname{Quant}_{1/2-\varepsilon}(\langle e_i, H^{1/2} Y_1\rangle^2, \dots, \langle e_i, H^{1/2} Y_N\rangle^2) \\
&\qquad\qquad\qquad - (\Phi^{-1}(3/4))^2 \|\Sigma^{1/2} H^{1/2} e_i\|^2|.
\end{aligned}$$

We only consider the first sum, as the second sum is analyzed similarly. Observe that by the spherical symmetry the random variable $\langle e_i, H^{1/2} Y\rangle^2 / \|\Sigma^{1/2} H^{1/2} e_i\|^2$ is distributed according to the χ_1^2 distribution. Denote

$$S_{N,i} = \{\langle e_i, H^{1/2} Y_1\rangle^2 / \|\Sigma^{1/2} H^{1/2} e_i\|^2, \dots, \langle e_i, H^{1/2} Y_N\rangle^2 / \|\Sigma^{1/2} H^{1/2} e_i\|^2\}.$$

Using the notation from the previous display, triangle inequality, and the fact that $F^{-1}_{\chi^2_1}(1/2) = (\Phi^{-1}(3/4))^2$, we arrive at

$$\begin{aligned}
&|\mathrm{Quant}_{1/2+\varepsilon}(\langle e_i, H^{1/2}Y_1\rangle^2, \dots, \langle e_i, H^{1/2}Y_N\rangle^2) - (\Phi^{-1}(3/4))^2 \|\Sigma^{1/2}H^{1/2}e_i\|^2| \\
&\quad \leqslant \|\Sigma^{1/2}H^{1/2}e_i\|^2 \cdot |\mathrm{Quant}_{1/2+\varepsilon}(S_{N,i}) - \mathbf{E}\,\mathrm{Quant}_{1/2+\varepsilon}(S_{N,i})| \\
&\qquad + \|\Sigma^{1/2}H^{1/2}e_i\|^2 \cdot |\mathbf{E}\,\mathrm{Quant}_{1/2+\varepsilon}(S_{N,i}) - F^{-1}_{\chi^2_1}(1/2+\varepsilon)| \\
&\qquad + \|\Sigma^{1/2}H^{1/2}e_i\|^2 \cdot |F^{-1}_{\chi^2_1}(1/2+\varepsilon) - F^{-1}_{\chi^2_1}(1/2)| \\
&\quad = (\mathrm{I})_i + (\mathrm{II})_i + (\mathrm{III})_i.
\end{aligned}$$

By Lemma 4 we have, for some $c_1 > 0$,

$$\|(\mathrm{I})_i + (\mathrm{II})_i\|_{\psi_1} \leqslant \frac{c_1 \|\Sigma^{1/2}H^{1/2}e_i\|^2}{\sqrt{N}},$$

and therefore,

$$\left\| \sum_{i=1}^d ((\mathrm{I})_i + (\mathrm{II})_i) \right\|_{\psi_1} \leqslant \frac{c_1 \operatorname{Tr}(\Sigma H)}{\sqrt{N}},$$

where the last expression follows from the triangle inequality. Using the exact form of the inverse cumulative distribution function of the χ^2_1 distribution and the same technique used to bound the difference of quantiles of half-normal distribution, one can verify that for any $\varepsilon \leqslant 1/4$ we have $|F^{-1}_{\chi^2_1}(1/2 \pm \varepsilon) - F^{-1}_{\chi^2_1}(1/2)| \leqslant c_2\varepsilon$, where $c_2 > 0$ is an absolute constant. This readily yields

$$(\mathrm{III})_i = \|\Sigma^{1/2}H^{1/2}e_i\|^2 \cdot |F^{-1}_{\chi^2_1}(1/2+\varepsilon) - F^{-1}_{\chi^2_1}(1/2)| \leqslant c_2 \|\Sigma^{1/2}H^{1/2}e_i\|^2 \varepsilon.$$

Therefore, for $\varepsilon \leqslant 1/4$, we have $\sum_{i=1}^d (\mathrm{III})_i \leqslant c_2\varepsilon \operatorname{Tr}(\Sigma H)$. Combining the above computations and using the tail bound of [53, Proposition 2.7.1], we prove the inequality (5.3).

To prove the second part of the bound, we propose a slightly different analysis for the term $(\mathrm{I})_i$. Denote

$$S'_{N,i} = \{|\langle e_i, H^{1/2}Y_1\rangle|/\|\Sigma^{1/2}H^{1/2}e_i\|, \dots, |\langle e_i, H^{1/2}Y_N\rangle|/\|\Sigma^{1/2}H^{1/2}e_i\|\},$$

and observe that $S'_{N,i}$ consists of independent half-normal random variables. We have

$$\begin{aligned}
(\mathrm{I})_i = \|\Sigma^{1/2}H^{1/2}e_i\|^2 \cdot &\left|\mathrm{Quant}_{1/2+\varepsilon}(S'_{N,i}) - \sqrt{\mathbf{E}\,\mathrm{Quant}_{1/2+\varepsilon}(S_{N,i})}\right| \\
&\times \left|\mathrm{Quant}_{1/2+\varepsilon}(S'_{N,i}) + \sqrt{\mathbf{E}\,\mathrm{Quant}_{1/2+\varepsilon}(S_{N,i})}\right|.
\end{aligned}$$

We first bound the second multiplier of the expression from the last display. By Lemma 5, with probability at least $1-\delta$, we have

$$\operatorname{Quant}_{1/2+\varepsilon}(S'_{N,i}) + \sqrt{\mathbf{E}\operatorname{Quant}_{1/2+\varepsilon}(S_{N,i})} \leqslant 2\Phi_{\mathrm{H}}^{-1}(1/2+\varepsilon) + c_3\sqrt{\frac{\log(1/\delta)}{N}}.$$

Now, observe that the last expression from the previous line is bounded by some absolute constant given that $\varepsilon \in [0, 1/4]$ and $N \geqslant c_4 \log(1/\delta)$. Using Lemma 5 once again together with union bound, we bound the term $(\mathrm{I})_i$, with probability at least $1-\delta$, as follows:

$$(\mathrm{I})_i \leqslant c_4 \|\Sigma^{1/2} H^{1/2} e_i\|^2 \sqrt{\frac{\log(1/\delta)}{N}}.$$

By the union bound, for all $i \in [d]$, we have, with probability at least $1-\delta$,

$$(\mathrm{I})_i \leqslant c_3 \|\Sigma^{1/2} H^{1/2} e_i\|^2 \sqrt{\frac{\log d + \log(1/\delta)}{N}},$$

whenever

$$N \geqslant c_2(\log d + \log(1/\delta)).$$

Taking the sum over all $i \in [d]$ concludes the proof. ∎

6. Additional comparisons and further extensions

The previously existing trimmed-mean-based estimators for the mean [45] and covariance [1, 50] have a remarkable property: they work similarly for classes of distributions starting from heavy-tailed to sub-Gaussian. The bounds for these estimators automatically adjust their dependence on the contamination level ε. However, it appears that these estimators are not adaptive to the favorable Gaussian case and give additional $\sqrt{\log(1/\varepsilon)}$ and $\log(1/\varepsilon)$ in mean and covariance estimation, respectively. Assume that we are estimating the mean of a standard Gaussian random variable in $\mathbb{R}$. We focus on the regime where the confidence level δ is fixed and the sample size N is approaching infinity, emphasizing the dependence on ε. For large enough N, the trimmed-mean-based estimator operates as follows (see [45, Section 2]): it drops the fraction of observations proportional to ε from both the smallest and largest ends and averages the remaining observations to form the mean estimator. However, for small enough ε, it is known that, in the standard Gaussian case, even the population ε and $1-\varepsilon$ quantiles are of order $\sqrt{\log(1/\varepsilon)}$ away from zero. Thus, if a malicious adversary places εN observations close to one of these quantiles, the trimmed mean estimator,

while averaging over εN observations each scaling as $\sqrt{\log(1/\varepsilon)}$, will be away by order of $\varepsilon\sqrt{\log(1/\varepsilon)}$ from the true mean, which is zero.

The same problem exists for robust covariance estimators. For example, consider the covariance estimator by Oliveira and Rico [50]. Assume that $X_1, \dots, X_N$ is an ε-contaminated sample of standard Gaussians in $\mathbb{R}$. In this particular case, the estimator described in [50] drops the k largest values among $X_1^2, \dots, X_N^2$ and averages the remaining $N - k$ observations. However, as with the trimmed mean, for large enough N, the "optimal" choice of k (termed k_0 in their paper), proportional to εN, corresponds to the $1 - \varepsilon$ quantiles of the squared Gaussian random variable, which are known to be of order $\log(1/\varepsilon)$ away from 1. As before, an adversary can exploit this to incur a loss of $\varepsilon \log(1/\varepsilon)$. In contrast, our estimators, based on different principles, avoid averaging and bypass these additional logarithmic factors.

While trimmed-mean-based estimators do not adaptively capture the specific behavior of the Gaussian distribution, this does not preclude other estimators from achieving similar dimension-free bounds with optimal dependence on ε. Since the release of first version of this paper, the authors have explored whether other estimators can match the optimal bounds of Theorems 1 and 2. The approach of Minsker [47], which concentrates on minimizing Huber's loss, seems to be the most promising. Although not explicit in his text, [47, Corollary 3.1] suggests a dimension-free bound, showing the same optimal ε dependence as in Theorem 1. The key difference lies in the treatment of the quantity $G_{f_v}(n, \Delta)$ appearing in this bound. While that paper opts for upper bounding this quantity in terms of the number of moments of the distribution, one can instead demonstrate that $G_{f_v}(n, \Delta)$ equals zero in the Gaussian case. This indicates no error in the Berry–Esseen approximation in this case. However, it remains unclear if the techniques in [47] extend beyond Gaussian mean estimation to recover the bound of Theorem 2.

Several natural questions arise from our results. The first is on the existence of computationally efficient estimators achieving our bounds. It is known that getting a polynomial time algorithm with a linear dependence on ε in the strong contamination model matching the bound of Theorem 1 is a challenging problem, even when the covariance matrix is identity. Covariance estimation is an even harder problem from the computational perspective. To the best of our knowledge, it is unknown if there is a polynomial time algorithm achieving the statistical performance of Theorem 2 even with the much weaker $\sqrt{\varepsilon}$-dependence on the contamination level.

One simpler question is if our bounds can be generalized beyond the Gaussian case. The answer is yes, and we opted for explicit Gaussian computations only to make our proofs more reader-friendly. In particular, the proof of Theorem 1 only uses the following properties of the distribution.

(1) The distribution of $X - \mu$ is symmetric around the origin.

(2) The distribution is spherically symmetric. That is, for any $v \in S^{d-1}$, the distribution of $\langle X - \mu, v\rangle / \sqrt{v^\top \Sigma v}$ does not depend on v. Denote the density function of this distribution by f.

(3) The density function f is separated from zero by some absolute constant for all $x \in [F^{-1}(1/2 - \varepsilon), F^{-1}(1/2 + \varepsilon)]$. This, in particular, implies

$$|F^{-1}(1/2 \pm \varepsilon) - F^{-1}(1/2)| \leqslant c\varepsilon$$

for some $c > 0$ and small enough ε.

(4) The distribution corresponding to the density function f is sub-Gaussian. That is, for Y distributed according to this distribution, we have $\|Y\|_{\psi_2} \leqslant c$ for some $c > 0$.

Following the lines of our proof almost verbatim, one can analyze these more general distributions. It will be interesting to understand if the sub-Gaussian tails assumption (Property 5) can be avoided. In fact, assuming Properties 1–4, and additionally that $\Sigma = I_d$, combining our techniques and the analysis in [21, Proposition 1.3], one can build an estimator $\widehat{\mu}$ satisfying, with probability at least $1 - \delta$,

$$\|\widehat{\mu} - \mu\| \leqslant c\left(\sqrt{\frac{d + \log(1/\delta)}{N}} + \varepsilon\right),$$

whenever $N \geqslant c_1(d + \log(1/\delta))$. Here, $c, c_1 > 0$ are some absolute constants. A similar bound without the sub-Gaussian assumption is also given by Chen, Gao, and Ren [13, Section 4]. In our case, the sub-Gaussian assumption (Property 5) is needed to control the moment-generating function when applying Lemma 1, while the proof in [21, Proposition 1.3] is based on the union bound over the ε-net for which we do not need sub-Gaussian tails in the "large deviation" regime.

Finally, some of the parameters of Theorem 2 are rather hard to estimate without making additional assumptions on the sample size and confidence level. One can adapt other approaches, such as, for example, Lepskii's method [42]. This could provide an alternative way of tuning these parameters.

A. Details for the proof of Theorems 1 and 2

This appendix contains lemmas that are used to prove Theorems 1 and 2.

A.1. Quantiles of Gaussian projections

Lemma 6. *Let $Y_1, \dots, Y_N$ be a sample of Gaussian vectors with covariance Σ. Denote $S_N = \{\langle Y_1, \theta\rangle / \sqrt{\theta^\top \Sigma \theta}, \dots, \langle Y_N, \theta\rangle / \sqrt{\theta^\top \Sigma \theta}\}$, where θ is a random vector*

with distribution ρ_v. Then, with probability at least $1-\delta$, for all $v \in S^{d-1}$, we have

$$\left|\mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)-\mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N))\right| \leqslant c_3\sqrt{\frac{\operatorname{Tr}(\Sigma)+\|\Sigma\|\log(1/\delta)}{N}}$$

for some absolute constant $c_3 > 0$.

Proof. Fix $\lambda > 0$ and let γ be a multivariate Gaussian distribution in $\mathbb{R}^d$ with zero mean and covariance $\beta^{-1}I_d$. The standard formula for the KL divergence between Gaussian vectors implies $\mathcal{KL}(\rho_v,\gamma) = \beta/2$. Thus, by Lemma 1, we have with probability at least $1-\delta$, simultaneously for all $v \in S^{d-1}$,

$$\begin{aligned}
&\lambda\mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)-\mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N))\\
&\quad\leqslant \mathbf{E}_{\rho_v}\log\mathbf{E}\exp\left(\lambda\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)-\mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N))\right)\\
&\qquad+\beta/2+\log(1/\delta).
\end{aligned}$$

Since centering multiplies the ψ_2-norm by at most an absolute constant factor (see, e.g., [53, Lemma 2.6.8]), we have by Lemma 3, for some absolute constant $c_1 > 0$,

$$\|\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)-\mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)\|_{\psi_2} \leqslant \frac{c_1}{\sqrt{N}}.$$

Thus, by [53, Proposition 2.5.2 (v)] (conditioned on θ, we take $\lambda\sqrt{\theta^\top\Sigma\theta}$ instead of λ in that result), we have, for some absolute constant $c_2 > 0$,

$$\begin{aligned}
&\mathbf{E}_{\rho_v}\log\mathbf{E}\exp\left(\lambda\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)-\mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N))\right)\\
&\quad\leqslant\frac{\mathbf{E}_{\rho_v}c_2\lambda^2\theta^\top\Sigma\theta}{N}\\
&\quad=\frac{c_2\lambda^2(v^\top\Sigma v+\beta^{-1}\operatorname{Tr}(\Sigma))}{N}\\
&\quad\leqslant\frac{11c_2\lambda^2\|\Sigma\|}{N},
\end{aligned}$$

where the last lines are based on a direct computation and our choice of β (we have $\beta^{-1} \leqslant 10\|\Sigma\|/\operatorname{Tr}(\Sigma)$). Optimizing the bound on $\lambda\mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N) - \mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N))$ with respect to $\lambda > 0$ and since $\beta \leqslant 10\operatorname{Tr}(\Sigma)/\|\Sigma\|$, we obtain that uniformly over S^{d-1},

$$\mathbf{E}_{\rho_v}\sqrt{\theta^\top\Sigma\theta}(\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)-\mathbf{E}\operatorname{Quant}_{\frac{1}{2}+\varepsilon}(S_N)) \leqslant c_3\sqrt{\frac{\operatorname{Tr}(\Sigma)+\|\Sigma\|\log(1/\delta)}{N}},$$

where $c_3 > 0$ is an absolute constant. Repeating the proof for $\lambda < 0$ and using the union bound, we get the desired upper bound. ∎

B. Proofs of concentration inequalities for sample quantiles

Proof of Lemma 3

We only analyze the quantile $Y_{((1/2+\varepsilon)N)}$, since the analysis for $Y_{((1/2-\varepsilon)N)}$ is the same. We analyze two parts of the tail separately. First, we show that

$$\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \geqslant t) \leqslant \exp(-c_1 N t^2).$$

This analysis is also split into two regimes. For some absolute constant $C > 0$, we first show the above inequality for $0 \leqslant t \leqslant C$ and then proceed with the case $t \geqslant C$. In the first regime, we follow the standard reduction to binomial tails (see similar computations in [52, Theorem 5.9] and [54]). Let $Z_1, \dots, Z_N$ be independent Bernoulli random variables with expectation

$$p = 1 - \Phi(s)$$

for some fixed $s \in \mathbb{R}$. We have

$$\Pr(Y_{((1/2+\varepsilon)N)} \geqslant s) = \Pr\left(\sum_{i=1}^{N} Z_i > (1/2-\varepsilon)N\right).$$

We set $s = \Phi^{-1}(1/2+\varepsilon) + t$ and obtain, using that $\mathbf{E}Z_i = 1 - \Phi(\Phi^{-1}(1/2+\varepsilon)+t)$,

$$\begin{aligned}&\Pr(Y_{((1/2+\varepsilon)N)} \geqslant \Phi^{-1}(1/2+\varepsilon)+t)\\ &\quad= \Pr\left(\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbf{E}Z_i > \Phi(\Phi^{-1}(1/2+\varepsilon)+t) - \frac{1}{2} - \varepsilon\right).\end{aligned}$$

Denoting $\varphi_+(t) = \Phi(\Phi^{-1}(1/2+\varepsilon)+t) - 1/2 - \varepsilon$, we have by Hoeffding's inequality applied to independent Bernoulli random variables, whenever $\varphi_+(t) \geqslant 0$,

$$\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \geqslant t) \leqslant \exp(-2N(\varphi_+(t))^2). \tag{B.1}$$

Let us lower bound the function φ_+. Using the density formula, we have

$$\begin{aligned}\varphi_+(t) &= \Phi(\Phi^{-1}(1/2+\varepsilon)+t) - \Phi(\Phi^{-1}(1/2+\varepsilon))\\ &\geqslant \frac{t}{\sqrt{2\pi}} \exp(-(\Phi^{-1}(1/2+\varepsilon)+t)^2/2)\\ &\geqslant \frac{t}{\sqrt{2\pi}} \exp(-(\Phi^{-1}(3/4)+t)^2/2).\end{aligned}$$

We combine these computations with the tail for large values of t. Since $\varepsilon \leqslant 1/4$, we need that at least $N/4$ (assume that it is an integer without loss of generality) of all

observations are above $\Phi^{-1}(1/2+\varepsilon)+t$. This can be controlled as follows:

$$\begin{aligned}\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \geqslant t) &\leqslant \binom{N}{N/4}(\Pr(Y_1 \geqslant \Phi^{-1}(1/2+\varepsilon)+t))^{N/4} \\ &\leqslant 2^N(\Pr(Y_1 \geqslant t))^{N/4} \\ &\leqslant 2^N \exp(-Nt^2/8) \\ &= \exp(N\log(2) - Nt^2/8) \\ &\leqslant \exp(-Nt^2/16),\end{aligned}$$

whenever $t \geqslant 4\sqrt{\log(2)}$. The inequality (B.1) and the lower bound on $\varphi_+(t)$ give us

$$\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \geqslant t) \leqslant \exp\left(\frac{-2Nt^2}{2\pi\exp((\Phi^{-1}(3/4)+4\sqrt{\log(2)})^2)}\right),$$

whenever $0 \leqslant t \leqslant 4\sqrt{\log(2)}$. Combining two regimes and adjusting the absolute constant, we prove an upper tail. Let us prove the lower tail bound. The proof is similar, though the computations are slightly different. We want to show that, for any $t \geqslant 0$,

$$\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \leqslant -t) \leqslant \exp(-c_1 Nt^2).$$

We have for any $s \in \mathbb{R}$ and $Z_1, \ldots, Z_N$ as above

$$\Pr(Y_{((1/2+\varepsilon)N)} \leqslant s) = \Pr\left(\sum_{i=1}^N Z_i \leqslant (1/2-\varepsilon)N\right).$$

Define

$$\varphi_-(t) = \Phi(\Phi^{-1}(1/2+\varepsilon) - t) - 1/2 - \varepsilon.$$

We have similarly

$$\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \leqslant -t) \leqslant \exp(-2N(\varphi_-(t))^2).$$

Now, we lower bound the quantity $|\varphi_-(t)|$ as follows:

$$\begin{aligned}|\varphi_-(t)| &= \Phi(\Phi^{-1}(1/2+\varepsilon)) - \Phi(\Phi^{-1}(1/2+\varepsilon) - t) \\ &\geqslant \frac{t}{\sqrt{2\pi}}\exp(-\max\{(\Phi^{-1}(1/2+\varepsilon))^2, (\Phi^{-1}(1/2+\varepsilon)-t)^2\}/2) \\ &\geqslant \frac{t}{\sqrt{2\pi}}\exp(-\max\{(\Phi^{-1}(3/4))^2, t^2\}/2).\end{aligned}$$

As above, we need to get the tail for large values of t. We need that at least $N/4$ of all observations are below $\Phi^{-1}(1/2+\varepsilon) - t$. In what follows, we assume

$$t \geqslant 2\Phi^{-1}(3/4).$$

We have

$$\begin{aligned}
&\Pr(Y_{((1/2+\varepsilon)N)} - \Phi^{-1}(1/2+\varepsilon) \leqslant -t) \\
&\quad = \Pr(-Y_{((1/2+\varepsilon)N)} \geqslant t - \Phi^{-1}(1/2+\varepsilon)) \\
&\quad \leqslant \binom{N}{N/4} (\Pr(-Y_1 \geqslant t - \Phi^{-1}(1/2+\varepsilon)))^{N/4} \\
&\quad \leqslant 2^N (\Pr(Y_1 \geqslant t - \Phi^{-1}(3/4)))^{N/4} \\
&\quad \leqslant 2^N (\Pr(Y_1 \geqslant t/2))^{N/4} \\
&\quad \leqslant 2^N \exp(-Nt^2/32) \\
&\quad \leqslant \exp(-Nt^2/64),
\end{aligned}$$

whenever $t \geqslant \max\{2\Phi^{-1}(3/4), 8\sqrt{\log(2)}\}$. The union bound concludes the proof. Finally, our bound on the ψ_2-norm follows from [53, Proposition 2.5.2]. ∎

Our next auxiliary result converts a mixed sub-Gaussian/sub-exponential tail bound into a bound on the $\|\cdot\|_{\psi_1}$-norm. We present this standard computation for the sake of completeness.

Lemma 7. *Assume that a random variable X satisfies, for all $t \geqslant 0$,*

$$\Pr(|X| \geqslant t) \leqslant 2\exp(-K\min\{t^2, t\}),$$

where $K > 1$ is some constant. Then, there is an absolute constant $c > 0$ such that

$$\|X\|_{\psi_1} \leqslant \frac{c}{\sqrt{K}}.$$

Proof. We can simply compute the moments of X. For fixed $p \geqslant 1$, we have

$$\begin{aligned}
\mathbf{E}|X|^p &= \int_0^\infty \Pr(|X|^p \geqslant t)\, dt \\
&= \int_0^\infty \Pr(|X| \geqslant t) p t^{p-1}\, dt \leqslant 2\int_0^\infty \exp(-K\min\{t^2, t\}) p t^{p-1} dt \\
&\leqslant 2\int_0^\infty \exp(-Kt^2) p t^{p-1} dt + 2\int_0^\infty \exp(-Kt) p t^{p-1} dt \\
&= \frac{1}{K^{p/2}} p\Gamma(p/2) + \frac{1}{K^p} 2p\Gamma(p) \\
&\leqslant \frac{3p(p/2)^{p/2}}{K^{p/2}} + \frac{2p^p}{K^p} \leqslant \frac{3p^p}{K^{p/2}} + \frac{2p^p}{K^p} \leqslant \frac{5p^p}{K^{p/2}},
\end{aligned}$$

where $\Gamma(\cdot)$ stands for the gamma function, and we used $\Gamma(x) \leqslant 3x^x$ for all $x \geqslant 1/2$ together with $p\Gamma(p) = \Gamma(p+1) \leqslant p^p$. Finally, [53, Proposition 2.7.1 (b)] implies the desired bound. ∎

Proof of Lemma 4

First, the density g of χ_1^2 is given by

$$g(x) = \frac{\exp(-x/2)}{\sqrt{2\pi x}}, \quad x > 0. \tag{B.2}$$

It is also easy to show that $F^{-1}_{\chi_1^2}(1/2) = (\Phi^{-1}(3/4))^2$. Using the same notation again, we denote $\varphi_+(t) = F_{\chi_1^2}(F^{-1}_{\chi_1^2}(1/2+\varepsilon)+t) - 1/2 - \varepsilon$. Repeating the lines of the proof of Lemma 3, whenever $\varphi_+(t) \geqslant 0$, we have

$$\Pr(Y_{((1/2+\varepsilon)N)} - F^{-1}_{\chi_1^2}(1/2+\varepsilon) \geqslant t) \leqslant \exp(-2N(\varphi_+(t))^2).$$

Using (B.2), we have

$$\begin{aligned}\varphi_+(t) &= F_{\chi_1^2}(F^{-1}_{\chi_1^2}(1/2+\varepsilon)+t) - F_{\chi_1^2}(F^{-1}_{\chi_1^2}(1/2+\varepsilon)) \\ &\geqslant \frac{t\exp(-F_{\chi_1^2}(F^{-1}_{\chi_1^2}(1/2+\varepsilon)+t))}{\sqrt{2\pi F_{\chi_1^2}(F^{-1}_{\chi_1^2}(1/2+\varepsilon)+t)}}.\end{aligned}$$

This gives us a sub-Gaussian tail for as long as $\varepsilon \leqslant 1/4$ and $t \geqslant 0$ is bounded by some absolute constant. By the concentration of the χ^2 distribution [41, Lemma 1], we have, for $t \geqslant 0$,

$$\begin{aligned}&\Pr(Y_1 \geqslant 1 + \sqrt{2t} + t) \leqslant \exp(-t), \\ &\text{and thus, } \Pr(Y_1 \geqslant t) \leqslant \exp(-t/2), \text{ whenever } t \geqslant 4 + 2\sqrt{3}.\end{aligned}$$

Since $\varepsilon \leqslant 1/4$, we need that at least $N/4$ of all observations are above $F^{-1}_{\chi_1^2}(1/2+\varepsilon)+t$. Therefore, whenever $t \geqslant 4+2\sqrt{3}$, we have

$$\begin{aligned}\Pr(Y_{((1/2+\varepsilon)N)} - F^{-1}_{\chi_1^2}(1/2+\varepsilon) \geqslant t) &\leqslant \binom{N}{N/4}(\Pr(Y_1 \geqslant \Phi^{-1}(1/2+\varepsilon)+t))^{N/4} \\ &\leqslant \binom{N}{N/4}(\Pr(Y_1 \geqslant t))^{N/4} \\ &\leqslant 2^N \exp(-Nt/8) \\ &= \exp(N\log(2) - Nt/8) \\ &\leqslant \exp(-Nt/16),\end{aligned}$$

where the last inequality requires additionally $t \geqslant 16\log(2)$. Combining the above bounds and adjusting the absolute constant $c_1 \geqslant 0$, we show that

$$\Pr(Y_{((1/2+\varepsilon)N)} - F^{-1}_{\chi_1^2}(1/2+\varepsilon) \geqslant t) \leqslant \exp(-c_1 N \min\{t, t^2\}).$$

We continue with the bound on the lower tail. We want to show, for any $t \geqslant 0$,

$$\Pr(Y_{((1/2+\varepsilon)N)} - F^{-1}_{\chi^2_1}(1/2+\varepsilon) \leqslant -t) \leqslant \exp(-c_2 N t^2),$$

where c_2 is an absolute constant. For $0 \leqslant t < F^{-1}_{\chi^2_1}(1/2+\varepsilon)$, we define

$$\varphi_-(t) = F^{-1}_{\chi^2_1}(F^{-1}_{\chi^2_1}(1/2+\varepsilon) - t) - 1/2 - \varepsilon.$$

Using the same argument, we show

$$\Pr(Y_{((1/2+\varepsilon)N)} - F^{-1}_{\chi^2_1}(1/2+\varepsilon) \leqslant -t) \leqslant \exp(-2N(\varphi_-(t))^2).$$

We have

$$\begin{aligned} |\varphi_-(t)| &= F_{\chi^2_1}(F^{-1}_{\chi^2_1}(1/2+\varepsilon)) - F_{\chi^2_1}(F^{-1}_{\chi^2_1}(1/2+\varepsilon) - t) \\ &\geqslant \frac{t \exp(-F^{-1}_{\chi^2_1}(1/2+\varepsilon))}{\sqrt{2\pi F^{-1}_{\chi^2_1}(1/2+\varepsilon)}}. \end{aligned}$$

Since $\varepsilon \leqslant 1/4$, we conclude the proof in the regime $t \leqslant F^{-1}_{\chi^2_1}(1/2+\varepsilon)$. Observe that due to the non-negativity of $Y_{((1/2+\varepsilon)N)}$ we can extend this bound to all $t > F^{-1}_{\chi^2_1}(1/2+\varepsilon)$. Lemma 7 in Section 3 concludes the proof. The analysis of the $(1/2-\varepsilon)$-th quantile repeats the same lines. ∎

Acknowledgments. The authors wish to thank Ankit Pensia for insightful discussions on the differences between contamination models, Vladimir V. Ulyanov for remarks on asymptotic laws for sample quantiles, and Arnak Dalalyan, Gábor Lugosi, and Stanislav Minsker for their useful comments.

Funding. The work of AM was supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and by the FAST Advance Grant.

References

[1] P. Abdalla and N. Zhivotovskiy, Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *J. Eur. Math. Soc.* (2024), DOI 10.4171/JEMS/1505

[2] J. Altschuler, V.-E. Brunel, and A. Malek, Best arm identification for contaminated bandits. *J. Mach. Learn. Res.* **20** (2019), article no. 91 Zbl 1434.68391 MR 3960945

[3] J.-Y. Audibert and O. Catoni, Robust linear least squares regression. *Ann. Statist.* **39** (2011), no. 5, 2766–2794 Zbl 1231.62126 MR 2906886

[4] R. R. Bahadur, A note on quantiles in large samples. *Ann. Math. Statist.* **37** (1966), 577–580 Zbl 0147.18805 MR 0189095

[5] A.-H. Bateni, A. Minasyan, and A. S. Dalalyan, Nearly minimax robust estimator of the mean vector by iterative spectral dimension reduction. 2022, arXiv:2204.02323v1

[6] S. Bobkov and M. Ledoux, One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Mem. Amer. Math. Soc.* **261** (2019), no. 1259, v+126 Zbl 1454.60007 MR 4028181

[7] C. Borell, The Brunn–Minkowski inequality in Gauss space. *Invent. Math.* **30** (1975), no. 2, 207–216 Zbl 0292.60004 MR 0399402

[8] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013 Zbl 1279.60005 MR 3185193

[9] S. Boucheron and M. Thomas, Concentration inequalities for order statistics. *Electron. Commun. Probab.* **17** (2012), article no. 51 Zbl 1349.60021 MR 2994876

[10] M. V. Burnashev, Asymptotic expansions for the median estimate of a parameter. *Theory of Probability & Its Applications* **41** (1997), no. 4, 632–645 Zbl 0901.62024

[11] O. Catoni, PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. 2016, arXiv:1603.05229v1

[12] O. Catoni and I. Giulini, Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. 2017, arXiv:1712.02747v2

[13] M. Chen, C. Gao, and Z. Ren, Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** (2018), no. 5, 1932–1960 Zbl 1408.62104 MR 3845006

[14] Y. Cheng, I. Diakonikolas, and R. Ge, High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2755–2771, SIAM, Philadelphia, PA, 2019 Zbl 1432.68615 MR 3909640

[15] B. S. Cirel'son, I. A. Ibragimov, and V. N. Sudakov, Norms of Gaussian sample functions. In *Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975)*, pp. 20–41, Lecture Notes in Math. 550, Springer, Berlin, 1976 Zbl 0359.60019 MR 0458556

[16] A. S. Dalalyan and A. Minasyan, All-in-one robust estimator of the Gaussian mean. *Ann. Statist.* **50** (2022), no. 2, 1193–1219 Zbl 1486.62159 MR 4404933

[17] H. A. David and H. N. Nagaraja, *Order statistics*. 3rd edn., Wiley Ser. Probab. Stat., Wiley-Interscience John Wiley & Sons, Hoboken, NJ, 2003 Zbl 1053.62060 MR 1994955

[18] L. de Haan and A. Ferreira, *Extreme value theory. An introduction*. Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2006 Zbl 1101.62002 MR 2234156

[19] J. Depersin and G. Lecué, On the robustness to adversarial corruption and to heavy-tailed data of the Stahel–Donoho median of means. *Inf. Inference* **12** (2023), no. 2, 814–850 Zbl 1528.62019 MR 4565752

[20] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, Being robust (in high dimensions) can be practical. In *Proceedings of the 34th international conference on machine learning, ICML, 70*, pp. 999–1008, 2017

[21] I. Diakonikolas and D. M. Kane, Recent advances in algorithmic high-dimensional robust statistics. 2019, arXiv:1911.05911v1

[22] I. Diakonikolas, D. M. Kane, J. C. H. Lee, and A. Pensia, Outlier-robust sparse mean estimation for heavy-tailed distributions. 2022, arXiv:2211.16333v1

[23] I. Diakonikolas, D. M. Kane, and A. Pensia, Outlier robust mean estimation with subgaussian rates via stability. In *Advances in neural information processing systems, 33*, pp. 1830–1840, 2020

[24] I. Diakonikolas, D. M. Kane, and A. Stewart, Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures (extended abstract). In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pp. 73–84, IEEE Computer Society, Los Alamitos, CA, 2017 MR 3734219

[25] D. Donoho and P. J. Huber, The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, pp. 157–184, Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA, 1983 Zbl 0523.62032 MR 0689745

[26] D. L. Donoho and M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** (1992), no. 4, 1803–1827 Zbl 0776.62031 MR 1193313

[27] M. D. Donsker and S. R. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. I. *Comm. Pure Appl. Math.* **28** (1975), 1–47 Zbl 0323.60069 MR 0386024

[28] I. Giulini, Robust dimension-free Gram operator estimates. *Bernoulli* **24** (2018), no. 4B, 3864–3923 Zbl 1415.62012 MR 3788191

[29] F. R. Hampel, A general qualitative definition of robustness. *Ann. Math. Statist.* **42** (1971), 1887–1896 Zbl 0229.62041 MR 0301858

[30] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics. The approach based on influence functions*. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley & Sons, New York, 1986 Zbl 0593.62027 MR 0829458

[31] Q. Han, Exact bounds for some quadratic empirical processes with applications. [v1] 2022, [v3] 2024, arXiv:2207.13594v3

[32] S. B. Hopkins and J. Li, How hard is robust mean estimation? In *Conference on learning theory*, pp. 1649–1682, 99, PMLR, 2019

[33] P. J. Huber, Robust estimation of a location parameter. *Ann. Math. Statist.* **35** (1964), 73–101 Zbl 0136.39805 MR 0161415

[34] P. J. Huber, *Robust statistics*. Wiley Ser. Probab. Math. Statist., John Wiley & Sons, New York, 1981 Zbl 0536.62025 MR 0606374

[35] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou, User-friendly covariance estimation for heavy-tailed distributions. *Statist. Sci.* **34** (2019), no. 3, 454–471 Zbl 1429.62312 MR 4017523

[36] J. Kiefer, On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* **38** (1967), 1323–1342 Zbl 0158.37005 MR 0217844

[37] Y. Klochkov and N. Zhivotovskiy, Uniform Hanson–Wright type concentration inequalities for unbounded entries via the entropy method. *Electron. J. Probab.* **25** (2020), article no. 22 Zbl 1445.60019 MR 4073683

[38] A. Kolmogorov, La méthode de la mediane dans la théorie des erreurs. *Rec. Math. Moscou* **38** (1931), no. 3-4, 47–50 Zbl 0006.06803

[39] V. Koltchinskii and K. Lounici, Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** (2017), no. 1, 110–133 Zbl 1366.60057 MR 3556768

[40] K. A. Lai, A. B. Rao, and S. Vempala, Agnostic estimation of mean and covariance. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pp. 665–674, IEEE Computer Society, Los Alamitos, CA, 2016 MR 3631029

[41] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (2000), no. 5, 1302–1338 Zbl 1105.62328 MR 1805785

[42] O. V. Lepskiĭ, On a problem of adaptive estimation in Gaussian white noise (in Russian). *Teor. Veroyatnost. i Primenen.* **35** (1990), no. 3, 459–470. English translation. *Theory Probab. Appl.* **35** (1991), no. 3, 454–466 Zbl 0745.62083 Zbl 0725.62075

[43] G. Lugosi and S. Mendelson, Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** (2019), no. 5, 1145–1190 Zbl 1431.62123 MR 4017683

[44] G. Lugosi and S. Mendelson, Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields* **175** (2019), no. 3-4, 957–973 Zbl 1431.62234 MR 4026610

[45] G. Lugosi and S. Mendelson, Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** (2021), no. 1, 393–410 Zbl 1461.62069 MR 4206683

[46] S. Mendelson and N. Zhivotovskiy, Robust covariance estimation under L_4-L_2 norm equivalence. *Ann. Statist.* **48** (2020), no. 3, 1648–1664 Zbl 1451.62084 MR 4124338

[47] S. Minsker, Uniform bounds for robust mean estimators. [v1] 2018, [v4] 2019, arXiv:1812.03523v4

[48] S. Minsker and L. Wang, Robust estimation of covariance matrices: Adversarial contamination and beyond. *Statist. Sinica* **34** (2024), no. 3, 1565–1583 Zbl 07901854 MR 4764687

[49] R. I. Oliveira, The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields* **166** (2016), no. 3-4, 1175–1194 Zbl 1360.60075 MR 3568047

[50] R. I. Oliveira and Z. F. Rico, Improved covariance estimation: Optimal robustness and sub-Gaussian guarantees under heavy tails. *Ann. Statist.* **52** (2024), no. 5, 1953–1977 Zbl 07961544 MR 4829476

[51] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. Wiley Ser. Probab. Math. Stat., John Wiley & Sons, New York, 1987 Zbl 0711.62030 MR 0914792

[52] J. Shao, *Mathematical statistics*. 2nd edn., Springer Texts Statist., Springer, New York, 2003 Zbl 1018.62001 MR 2002723

[53] R. Vershynin, *High-dimensional probability. An introduction with applications in data science*. Camb. Ser. Stat. Probab. Math. 47, Cambridge University Press, Cambridge, 2018 Zbl 1430.60005 MR 3837109

[54] D. Xia, Non-asymptotic bounds for percentiles of independent non-identical random variables. *Statist. Probab. Lett.* **152** (2019), 111–120 Zbl 1459.60055 MR 3953052

[55] N. Zhivotovskiy, Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electron. J. Probab.* **29** (2024), article no. 13 Zbl 1531.60019 MR 4693860

Received 16 November 2023; revised 4 February 2025.

Arshak Minasyan
Department of Statistics, Centrale-Supélec, Université Paris-Saclay, Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France; arshak.minasyan@centralesupelec.fr

Nikita Zhivotovskiy
Department of Statistics, University of California, Berkeley, 315 Evans Hall, Berkeley, CA 94720-3860, USA; zhivotovskiy@berkeley.edu