Doctoral Thesis

Optimal statistical estimation: sub-Gaussian properties, heavy-tailed data, and robustness

Candidate: Zoraida Fernandez-Rico
Advisor: Roberto Imbuzeiro Oliveira

Instituto de Matemática Pura e Aplicada

Rio de Janeiro, March, 2022.

Abstract

This thesis focus on optimal statistical estimation for finite samples from the perspectives of robustness and heavy-tailed data. We study two problems: properties of the trimmed mean, and covariance matrix estimation, both from a nonasymptotic perspective.

Regarding the trimmed mean, our main result is that the trimmed mean achieves sub-Gaussian performance, up to constant factors, when the trimming parameter $k \approx \log(1/\alpha)$ under suitable moment conditions. We also show that a different tuning of the trimming parameter gives minimax-optimal results with respect to adversarial data contamination, where a fraction ϵ of sample points can be modified arbitrarily. Furthermore, for more generality, we provide a way of choosing the trimming parameter based on Lepskii's.

Concerning the covariance matrix estimation, this thesis provides a sub-Gaussian optimal covariance estimator under heavy tails. Our main result improves the current state-of-art regarding high probability bounds given by Mendelson and Zhivotovskiy.

Keywords: sub-Gaussian estimators, trimmed mean, robustness, covariance.

A todas las que sabían la respuesta y no levantaron la mano.

Acknowledgements

Começo por agradecer aos meus pais por sempre me darem a janela nos voos, mesmo quando não havia avião, e agradeço a minha mãe em especial, pela esperança.

Gostaria de expressar minha mais profunda gratidão a Roberto Imbuzeiro Oliveira, meu orientador, por tornar esta tese possível, pelos conselhos e por tanta generosidade. Ele me apresentou ao tema desta tese e me orientou do começo ao fim. Agradeço também por todo o conhecimento e tempo que ele compartilhou comigo. Sinto-me muito afortunada de ter podido trabalhar com um matemático tão incrível e com um coração tão grande. Além disso, a perspectiva dele sobre a matemática e a vida cotidiana moldou um pouquinho a minha, o que é muito para mim.

Gostaria de agradecer também aos membros da minha banca de tese: Augusto Texeira, Florencia Leonardi, Hubert Lacoin, Marco Avella, Paulo Orenstein e Yannick Baraud. Seus comentários e sugestões sobre esta tese foram muito significativos. Também gostaria de expressar meu agradecimento a todos os professores do IMPA pelos seus cursos. Em especial, sou muito grata ao grupo de probabilidades do IMPA pelo ambiente amigável criado, e ao Jorge Zubelli por todo o apoio e a confiança.

Estou muito agradecida a Mariel Orellana, Sandra Vargas, Mariana Mendez e Lucia Toledo pelo amor. A Emily "Pi" Quesada-Herrera por sempre cruzar os dedos por mim. A Cristian González porque a magia do cinema era assistir aos filmes com ele. A Daniel Yukimura porque não era a cerveja o que me fazia sorrir, e sim sua companhia.

Agradeço especialmente a Miguel Ibieta e Roberto Viveros por me ensinarem truques de mágica na hora do recreio. Também sou muito grata ao Enzo pelos cafés com leite e por compartilhar comigo o céu que ele via, e ao Ticuliru (Pedro Campos) pela música e o verso livre.

Sou muito grata a Mauricio Loures, Lorena Bulhosa e Ian Pereira por me ensinar português e a misturar samba com chuva. Agradeço também a Daniela Cuesta, Daniela Paiva, Lorena

Duarte e Diana Pulido pela amizade e por me ajudar a mover os móveis da sala para poder dançar melhor.

Reservo um lugar especial nos meus agradecimentos para meus amigos. Agradeço a Claudia Lopez, Cristina Gareca, Maria Martha Sarabia, Danny Flores, Tainara Gobetti, Mateus Sousa, Cynthia Bortolotto, Maria Clara Mendes, Walner Mendoça, Santiago Achig, Javier Gargiulo, Juan Carlos Arroyave, Alejandro Vicente, Jonathan Trejos, Diego Andia, Olivier Thom, Roberto Villaflor, Valentino Sichinel, Adriana Laurindo, Clarice Netto, Jennifer Loria, Yaya Tall e Victor Perez pelo carinho, pela saudade e pelos reencontros. Estou muito agradecida com todas as meninas do Seminário de Mulheres porque uma conversa com elas mudava a minha tarde. Agradeço também aos meus colegas que pararam no corredor do IMPA para dizer "que bonita a sua ideia", "essa apresentação ficou muito boa" ou "você é muito corajosa por fazer o que faz".

Gostaria de agradecer a toda a equipe administrativa do IMPA por toda a ajuda com prazos e processos administrativos. A Flavia por manter minha sala sempre limpa e cuidar as minhas coisas da chuva. Para terminar, reconheço o apoio financeiro do CNPq.

Contents

Chapter 1

Introduction

1.1 A historical summary

1.1.1 Optimal estimation. In this thesis, we study a subject of Mathematical Statistics: optimal statistical estimation for finite samples. There has been a recent surge of interest on study this topic. In this line of research, one of the most important goals is to design estimators with sub-Gaussian guarantees under minimal assumptions.

Estimating the mean of a real random variable based on observations of a finite sample is the most fundamental problem in statistics. We start this chapter with the so-called sub-Gaussian case for this mean estimation problem. Here, we are interested in finding, for each sample size n and confidence level $1-\alpha$, an estimator $\widehat{E}_{n,\alpha} : \mathbb{R}^n \to \mathbb{R}$ with the following property. Let $X_1, \ldots, X_n$ be an independent and identically distributed sample from an unknown distribution with mean μ and finite variance σ^2. Then:

$$\mathbb{P}\left[|\widehat{E}_{n,\alpha}(X_1, \ldots, X_n) - \mu| \leq C\sigma \sqrt{\frac{\log(1/\alpha)}{n}} \right] \geq 1 - \alpha, \tag{1.1}$$

where $C > 0$ is uniform in n and α. The crucial point here is that, while the estimator may depend on α as well on n, the above bound should hold *uniformly* over all distributions with finite second moments, irrespective of how heavy their tails are.

Asymptotically (when $n \to \infty$) this problem is direct. Indeed, the Central Limit Theorem gives Gaussian guarantees for the *standard sample mean* (empirical mean). However, it is well known, that the sample mean is not optimal for finite samples unless the sample points are Gaussian. In [Catoni, 2012], the author showed that Chebyshev's inequality is essentially tight

for some data distribution: if $X_1, \ldots, X_n$ are i.i.d. random variable on $\mathbb{R}$ with mean μ and variance $\sigma^2 < +\infty$, the following holds for $1 - \alpha \in (0, 1)$ where c is a positive constant.

$$c\alpha \leq \mathbb{P}\left(\left| n^{-1} \sum_{i=1}^{n} X_i - \mu \right| \leq \sigma \sqrt{\frac{1}{\alpha n}} \right) \leq \alpha.$$

There exist other estimators satisfying optimal sub-Gaussian rates. For instance, Catoni in [Catoni, 2012] improves the empirical mean estimator achieving a bound in the fashion of (1.1) under the assumption of finite known variance, and from nonasymptotic point of view. For more examples, see [Devroye et al., 2016, Lee and Valiant, 2020]. § 2.2.2 provides information of this theme in more detail. In [Devroye et al., 2016], the authors prove bounds as in (1.1) for the median-of-means estimator. In addition, they prove that the assumption of finite variance is necessary. The next chapter is devoted to the analysis of the *trimmed-mean* (or truncated-mean) defined by removing a the k largest and smallest points for some parameter of the observations, and then averaging over the rest of the sample points. One of our most important results is that this estimator achieves minimax-optimal performance.

Statistical estimation in the presence heavy-tailed situations and outliers has recently attracted much attention. Following we give a short and partial review of some notable contributions on the subject matter.

1.1.2 Heavy tails.

Much recent work has been devoted to understanding sub-Gaussian estimation under weak assumptions. Following, we present a brief review of this topic for the mean of vector case and the covariance matrix estimation. For a thorough review see [Lugosi and Mendelson, 2019a]. In that work, Lugosi and Mendelson survey the progress in mean estimation and regression function estimation in the presence of heavy-tailed data.

In regards to the mean estimation for a vector, some important references are [Lugosi and Mendelson, 2019b, Minsker, 2015]. In [Minsker, 2015], Minsker provided a general estimator in Banach spaces with tight concentration bounds, although not optimal. Based on the idea of the multivariate median, in [Lugosi and Mendelson, 2019b], Lugosi and Mendelson present a sub-Gaussian estimator under the sole hypothesis of second moment. Similar rates were obtained years after in [Lugosi and Mendelson, 2021]. In the last paper, the authors show that the high-dimensional trimmed mean estimator has optimal performance under adversarial contamination and weak tail assumptions. Other important references in the multidimensional case are [Joly et al., 2017, Lugosi and Mendelson, 2019b] that propose successive improvements of the median-of-means approach to get an estimator with non-asymptotic sub-

Gaussian performance and dimension-free tail bound.

Let us mention that efficiently computable estimators for the mean of vectors have also been investigated. As opposed to the works cited above, Catoni and Giulini presented a computable estimator in [Catoni and Giulini, 2017]. In that article, the authors construct an estimator reached from the empirical mean under the unique hypothesis of the existence of a finite covariance matrix. Their estimator is straightforward to compute. However, it does not achieve sub-Gaussian bounds. Major results in this topic are given in [Hopkins, 2020, Cherapanamjeri et al., 2019]. Hopkins [Hopkins, 2020] provides the first polynomial-time algorithm to estimate the vector mean with sub-Gaussian performance; under finite mean and covariance hypotheses. Cherapanamjeri et al. obtained similar performance in [Cherapanamjeri et al., 2019], but they improve appreciably the run-time achieving optimal statistical efficiency. See also [Depersin and Lecué, 2022] for an estimator that is robust to outliers and heavy-tailed data, and also it has a nearly linear running time.

Concerning the covariance case, there has been wide interest in this problem. Important results in this topic are the following, [Koltchinskii and Lounici, 2014, Lounici, 2012, Minsker, 2018, Minsker and Wei, 2018, Ostrovskii and Rudi, 2019, Mendelson and Zhivotovskiy, 2019, Vershynin, 2011]. Let us mention that Minsker in [Minsker, 2018] provides a mean estimator for a random matrix assuming only finite second moment on the entries of such matrix. His estimator achieves sub-Gaussian or sub-exponential performance. In 2018, Minsker and Wei [Minsker and Wei, 2018] design an estimator that admits tight deviation bounds for heavy-tailed data. Their estimator depends on the dimension of the space. Recently, Mendelson and Zhivotovskiy [Mendelson and Zhivotovskiy, 2019] have in interested in the problem. Their main result shows that there is an estimator for the covariance matrix of a high-dimensional random vector that almost has optimal performance and is free-dimensional, under $L4 - L2$ norm equivalence assumption. One of the main results of this thesis is to construct a minimax-optimal estimator for the covariance matrix under mild hypotheses. Chapter § 3 is entirely dedicated to this question. For complementary information and further references see § 3.1.

We end this section by giving further detail about the sample points that we are considering. As we are not only interested in estimators that are sub-Gaussian optimal under heavy-tailed data, but also robust to modifications of a small fraction of observations.

1.1.3 Robustness.

In 1964, Huber's breakthrough paper [Huber, 1964] introduced the basis of Robust Statistics. Since then, a body of work on the subject of robust estimation has emerged [Huber, 1972, Huber, 1972, Stigler, 2010]. In Hubert's contamination model the outliers are i.i.d. with an unknown probability distribution. In this work, we consider the model of adversarial contamination of the data. In this sense, an adversary can corrupt a fraction ϵ of the sample. This model has gained space in the literature recently since the work [Diakonikolas and Kane, 2019].

It was noticed by Tukey and McLaughlin in [Tukey and Mclaughlin, 1963] that the Winsorized and the so-called trimmed mean estimate the mean from outlier contaminated data. Other results on the theoretical properties of the trimmed mean and Winsorized estimator were obtained in [Bickel, 1965]. The two aforementioned works are from an asymptotic approach. For further background on the robust theory and its connections with trimmed mean see § 2.2. Actually, we prove in chapter § 2 that the trimmed mean estimator for the mean of a real variable has sub-Gaussian performance under heavy-tailed data and also adversarial corruption.

Following the theme above, Lerasle and Oliveira [Lerasle and Oliveira, 2011] develop the theory of "robust empirical mean estimators". They obtain that the median-of-means estimator is robust. That estimator was constructed based on methods developed by Nemirovski and Yudin [Blair, 1985]. The problem of mean robust estimation gained much attention in high dimension. As mentioned in § 1.1.2 Minsker [Minsker, 2015] constructs an alternative version of the median-of-means estimator using the geometric mean. His estimator is robust in the general context of Banach spaces.

Broadly progress has been made in the direction of computationally efficient robust estimation. The first efficient robust estimators for learning several fundamental classes of high-dimensional distributions were provided in [Diakonikolas et al., 2019]. Subsequent work [Hopkins et al., 2021] unified view on robust and heavy-tailed mean estimation in high dimensions. That result translates into algorithms for both cases: heavy-tailed data and robustness. Therefore, the work of Hopkins et al. provides an algorithm that has a run-time that matches the fastest known algorithms on both perspectives. Recently, Diakonikolas and Kane [Diakonikolas et al., 2020] provide the first computationally efficient algorithm with sub-Gaussian and robust guarantees for mean estimation under a finite covariance assumption. Finally, a recent survey [Diakonikolas and Kane, 2019] on this setting may be consulted for further review.

1.2 Our contribution in this thesis

This thesis provides sub-Gaussian estimators under heavy tails and adversarial contamination, from a nonasymptotic perspective. We study the problem in the following settings: (i) mean estimation of a real random variable from an i.i.d. sample; (ii) covariance matrix estimation of a high-dimensional random vector. Our contributions to these topics are explained below.

We will present in detail our collaboration concerning the first problem in § 2. The trimmed mean is a classical estimator for expectations and location parameters of distribution. This thesis presents new finite-sample results on this estimator. One result is that the trimmed mean achieves minimax-optimal results for prespecified confidence and contamination levels. In particular, it satisfies a sub-Gaussian bound under the sole assumption that the variance exists. We also build nonasymptotic confidence intervals for the trimmed mean under higher-moment conditions, or assuming symmetry of the data distribution. Finally, we present an adaptive procedure for choosing the trimming parameter which is based on Lepskii's method. These results were obtained in collaboration with Paulo Orenstein and Roberto I. Oliveira.

The chapter § 3 is dedicated to the second question. Our work provides a minmax-optimal estimator for the covariance matrix Σ of a d-dimensional random vector from an i.i.d. random sample. Under the only assumption of bounded kurtosis (or $L^4 - L^2$ equivalence) over its one-dimensional marginals. Our estimator has an error performance that matches with the case of the Gaussian setting presented in [Koltchinskii and Lounici, 2014]. This holds even though we allow for very general distributions that may not have moments of order > 4. Our result improves a recent theorem by Mendelson and Zhivotovskiy displayed in [Mendelson and Zhivotovskiy, 2019]. These results were obtained in collaboration with Roberto I. Oliveira.

1.3 Notation

In this section, we introduce some notation used throughout the thesis. We denote by c and C absolute positive constants whose value may change from line to line. The cardinality of a finite set A is denoted by $\#A$. For real numbers x and y, $x_+ := \max\{x, 0\}$ and $x_- := \max\{-x, 0\}$ denote its positive and negative parts, respectively; and $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$. Moreover, $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor and ceiling of x, respectively. The set of positive integers is $\mathbb{N} := \{1, 2, 3, \dots\}$. For $n \in \mathbb{N}$, $[n] := \{i \in \mathbb{N} : 1 \leq i \leq n\}$ is

the set of numbers from 1 to n. The unit sphere of the Euclidean norm in $\mathbb{R}^d$ is denoted by $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\| = 1\}$. For a matrix M, $\|M\|_{\text{op}}$ denotes the operator norm. The effective rank of a non-null positive semidefinite square matrix $M \in \mathbb{R}^{d\times d}$ is given by

$$r(M) = \frac{\text{tr}\,(M)}{\|M\|_{\text{op}}}.$$

We use the symbol $X \sim P$ to say that X is a random variable with distribution P. We also write $X \sim Y$ when X, Y are random variables with the same distribution.

Chapter 2

Trimmed-means results

2.1 Introduction

We consider the problem of estimating the expectation or a location parameter of a random variable from an i.i.d. random sample. The sample mean is the standard estimator for these tasks. However, it is very sensitive to outliers.

The trimmed mean is a more robust alternative to the sample mean. Let $X_1, \ldots, X_n$ be a random sample and denote by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ its order statistics. Given an integer $0 \leq k < n/2$, the k-trimmed-mean is given by:

$$\overline{X}_{n,k} := \frac{1}{n-2k} \sum_{i=k+1}^{n-k} X_{(i)}.$$

This is the arithmetic mean of sample points after the k largest and k smallest values of the sample are removed. $\overline{X}_{n,k}$ equals the standard sample mean for $k = 0$, whereas $k = \lceil n/2 \rceil - 1$ gives a sample median. Intermediate choices of k will in general lead to different trade-offs between bias and variance.

Starting in the late Sixties, the trimmed mean was analyzed in a number of papers. Important asymptotic results include [Stigler, 1973, Jaeckel, 1971, Hall, 1981, Léger and Romano, 1990, Jana Jurecková, 1994], and experiment-based analyses are given in [Hogg, 1974, Stigler, 1977, Rocke et al., 1982], among other references. Further discussion of this large literature is provided in §2.2.1 below.

In this thesis, we investigate the trimmed mean from a nonasymptotic perspective, obtaining

new results. Our main contributions are listed below.

- *Minimax-optimal trimming under moment conditions and contamination.* Given a confidence level $1 - \alpha$, we show that the trimmed mean with a distribution-independent choice $k \approx \log(1/\alpha)$ achieves minimax optimal error rates for estimating the population mean. In particular, we obtain a "sub-Gaussian" estimator in the finite-variance setting [Catoni, 2012, Devroye et al., 2016]. A different, but also distribution-independent choice of k gives optimal results to adversarial data contamination in the sense of [Diakonikolas and Kane, 2019].
- *Nonasymptotic confidence intervals.* In the contamination-free setting, we show how one can build confidence intervals in two settings. The first one is the classical setting where the parameter of interest is the median of the distribution, and data is symmetrical around the median. The other case is that of mean estimation under kurtosis-type assumptions.
- *Adaptive choice trimming.* The above results are for fixed choices of the trimming parameters. The problem of tuning the trimming has also been much studied at least since Jaeckel [Jaeckel, 1971]. We give here a proposal for selecting the trimming parameter based on Lepskii's adaptive method, which is related to recent work on Winsorized importance sampling [Orenstein, 2018].

Underlining all of these results is a Bernstein-type concentration inequality for the trimmed mean that holds under essentially no assumptions. Loosely speaking, this inequality holds conditionally on certan order statistics in the random sample, and gives concentration around a "population trimmed mean." This simple observation seems new, but turns out to be essential to our theory.

The remainder of the chapter is organized as follows. Related work on the trimmed mean, "sub-Gaussian" estimators and other topics is discussed in Section 2.2. Section 2.3 presents the main definitions and assumptions in the chapter. Our conditional concentration bound is presented in Section 2.4. The subsequent Section 2.5 presents bounds for general and optimal choices of k under moment conditions and contamination. Section 2.6 presents nonasymptotic confidence intervals for the trimmed mean. Finally, Section 2.7 discusses our method for choosing the trimming parameters from the data. An appendix contains additional proofs.

2.2 Related work

2.2.1 Background on the trimmed mean. The literature on the trimmed mean is quite large. In what follows, we present a brief and partial review. Readers interested in this topic are encouraged to consult the book of Huber and Ronchetti [Huber and Ronchetti, 2009], and also Stigler's historical overview [Stigler, 2010] for some background on the general field of robust estimators.

The idea of trimming a sample predates Statistics as a science; see [Huber, 1972] for early historical references. Tuckey's seminal paper [Tukey, 1962] explicitly proposes the trimmed mean and the Winsorized estimator:

$$\overline{X}^{w}_{n,k} = \frac{n-2k}{n}\,\overline{X}_{n,k} + \frac{k}{n}\left(X_{(k)} + X_{(n-k+1)}\right)$$

as ways to estimate location parameters from outlier-contaminated data. Tuckey also suggested the possibility or data-dependent choices of k.

With the advent of classical Robust Statistics, in the hands of Huber [Huber and Ronchetti, 2009, Huber, 1972] and others, the trimmed mean became a popular topic of study. An important result, due to Stigler [Stigler, 1973] gives the asymptotic distribution of $\overline{X}_{n,k}$ when $n \to +\infty$ and $k = \lfloor \eta n \rfloor$ with $\eta \in (0, 1/2)$. This distribution will depend on continuity properties of the cumulative distribution function. Moreover, the estimator is asymptotically unbiased when the data distribution is symmetric, but not in general. We note that other methods based on linear statistics of the ordered sample have also been proposed [Stigler, 1974] which can be "nicer" than the trimmed mean in some ways. We note that higher-dimensional versions of the trimmed mean have been considered [Maller, 1988].

Starting with Jaeckel [Jaeckel, 1971], a number of papers have looked at the problem of choosing k adaptively. The main result of [Jaeckel, 1971], sharpened in [Hall, 1981] and improved in [Jana Jurecková, 1994], shows that a certain way of choosing k that "minimizes the asymptotic variance" gives an asymptotically normal estimator. Other methods for choosing k include the bootstrap [Léger and Romano, 1990] and random weighting methods [Shi Jian, Zheng Zhongguo , 1996]. All of these papers make two strong assumptions. Firstly, the distribution of the data must be symmetric around the median. Secondly, the cumulative distribution function of the data-generating distribution must be "well-behaved" in some sense.

A line of works has studied trimmed means via experiments. Hogg [Hogg, 1974]

presents a number of results on adaptive robust estimators and makes concrete suggestions on trimmed means. The experiments comparing trimming and Winsorization in [Wilfrid J. Dixon, Karen K. Yuen, 1974] suggest trimming is usually better. Stigler [Stigler, 1977] compares different robust estimators over real datasets and shows that trimmed mean with $k = \lfloor 0.1n \rfloor$ is often one of the very best estimators. Further analysis by Rocke et al [Rocke et al., 1982] does not quite corroborate Stigler, but still indicates that the trimmed mean has good performance. Other work [Lee, 2004] proposes a method for choosing k with good practical performance, but no mathematical analysis. Finally, we note in passing that there are papers on high-dimensional versions of the trimmed mean [Maller, 1988].

2.2.2 Minimax optimality for fixed confidence, and adversarial contamination. We now turn to a topic of more recent interest: that of mean estimators with near-optimal finite-sample guarantees. As we discuss in § 1.1.1 it is not obvious that such sub-Gaussian estimators should exist. For instance, the sample mean is not sub-Gaussian in this sense, as Chebyshev's inequality is nearly tight for i.i.d. sums. Catoni's seminal work [Catoni, 2012] provides one such estimator in the case where σ^2 is known and $\log(1/\alpha) \ll n$. In fact, his estimator achieves $C = \sqrt{2} + o(1)$ for (1.1) which can be shown to be optimal. Experiments reveal that Catoni's estimator is more efficient than the sample mean even for very simple distributions P.

Reference [Devroye et al., 2016] explores the notion of sub-Gaussian estimators in greater depth. That paper shows that, in general, sub-Gaussian estimators must indeed depend on the desired confidence $1-\alpha$, and that some bound of the sort $\log(1/\alpha) \leq c\,n$ is needed. On the other hand, estimators that work across wide range of α are possible under slightly stronger conditions. That paper also noted that a simple estimator called "median-of-means" is sub-Gaussian even when the variance is unknown, albeit with a suboptimal constant $C > 0$. Recent work [Lee and Valiant, 2020] gives sub-Gaussian estimators with near optimal $C = \sqrt{2} + o(1)$ for the case of unknown variance. There has also been great interest in extending these results to higher dimensions: see [Lugosi and Mendelson, 2019b, Lugosi and Mendelson, 2021] and the survey [Lugosi and Mendelson, 2019a] for more details. Incidentally, the "high dimensional trimmed mean" in [Lugosi and Mendelson, 2021] was inspired by an early version of the present work. A further line of work considers what happens in the case of even heavier tails. Instead of assuming finite variance, assume now that $\mathbb{E}\left[|X_1 - \mu|^p\right] \leq \nu_p^p$ for some $1 < p < 2$. It

follows from [Bubeck et al., 2013] that the median of means estimator satisfies

$$\mathbb{P}\left[|\widehat{E}_{n,\alpha}(X_1,\ldots,X_n)-\mu| \le C\nu_p\left(\frac{\log(1/\alpha)}{n}\right)^{1-1/p}\right] \ge 1-\alpha, \tag{2.1}$$

for some universal $C > 0$. It is possible to show that this cannot be improved, up to the value of C [Devroye et al., 2016, Theorem 3.1]. The upshot is that the median-of-means estimator is optimal for any choice of $1 \le p \le 2$. As it turns out, a suitably tuned trimmed mean has similar performance (cf. Theorem 2.18 below).

Finally, we discuss the model of adversarial data contamination. Recall that the traditional contamination model in Robust Statistics is that of Huber [Huber, 1964], where there is a uncontaminated distribution P, but data comes from a contaminated law $(1-\epsilon)P + \epsilon Q$, with Q unknown. In the adversarial model we consider, an ϵ fraction of data points may be replaced *arbitrarily*. In particular, one may imagine that an adversary gets to see the uncontaminated random sample and then chooses which points to replace so as to foil the statistician. This model has become standard in recent work on algorithmic high-dimensional Statistics [Diakonikolas and Kane, 2019]. It places quite strong requirements on an estimator, and proving results about it can be easier for this very reason.

2.2.3 Adaptive estimators. There are many methods in the literature for adaptively selecting an estimator from a family of candidates. In the mean estimation setting, [Devroye et al., 2016] propose one such method based on confidence intervals, which require $p > 2$ moments of the data distribution. A recent paper by one of the authors [Orenstein, 2018] discusses adaptive Winsorization of the sample mean under weak assumptions. The main application that method is to importance sampling, where sample points receive weights that may vary significantly, thus leading to large variance. The main contribution of that paper was to show that a procedure derived from Lepskii's adaptation method [Lepskii, 1991] (see also [Mathé, 2006]) balances the bias and variance of Winsorized estimates. Here, we show that a method based on sample trimming has similar performance.

2.3 Assumptions

2.3.1 Basic assumptions on the data. Our most basic assumption will be that we have i.i.d. data.

Assumption 2.1 (i.i.d. data)**.** $X_1, \dots, X_n$ is a sample of size $n \in \mathbb{N}$ of independent and identically distributed random variables with common distribution P and cumulative distribution function

$$F(t) := P(-\infty, t] = \mathbb{P}[X_1 \leq t] \ \ (t \in \mathbb{R}).$$

We let F^{-1} be the quantile transform (or generalized inverse) of P.

We will also consider variants of this assumption. In the first one, we make moment assumptions on P.

Assumption 2.2 (i.i.d. data with a mean and higher moments)**.** Besides Assumption 2.1, we assume P has a well-defined mean

$$\mu := \int_{\mathbb{R}} x \, P(dx) = \mathbb{E}[X_1]$$

We also define the (possibly infinite) centered L^p norms:

$$\nu_p := \left(\int_{\mathbb{R}} |x - \mu|^p \, P(dx) \right)^{1/p} = \|X_1 - \mu\|_{L^p} \in [0, +\infty] \ \ (1 \leq p < +\infty)$$

and also write $\sigma^2 := \nu_2^2$ for the (possibly infinite) variance.

The second assumption makes no restrictions whatsoever on the moments of P, but requires symmetry; for instance, P could be a shifted Cauchy distribution.

Assumption 2.3 (i.i.d. symmetrical data)**.** Besides Assumption 2.1, we assume P is symmetrical about its median μ, i.e. that $X_1 \sim 2\mu - X_1$.

Assumption 2.4 (Adversarial contamination)**.** Let $X_1', \dots, X_n'$ a set of random variables defined over the same probability space as the X_i. We call this set an ϵ-contamination of $\{X_i\}_{i=1}^n$ if

$$\#\{i \in [n] \, : \, X_i' \neq X_i\} \leq \epsilon \, n.$$

2.3.2 The trimmed mean and related quantities.

The trimmed mean will be our main object of study in this chapter. The next definition adds a few other quantities that will be of interest when analyzing the trimmed mean.

Definition 2.5 ((k_1, k_2)-trimmed mean, variance and width)**.** Let $X_1, \ldots, X_n$ satisfy Assumption 2.1, and let

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

denote the increasing rearrangement of the sample (ie. its order statistics). Assume $k_1, k_2 \in \mathbb{N} \cup \{0\}$ satisfy $k_1 + k_2 < n$. The (k_1, k_2)-trimmed mean estimator is defined as

$$\overline{X}_{n,k_1,k_2} := \frac{1}{n - k_1 - k_2} \sum_{i=k_1+1}^{n-k_2} X_{(i)}$$

and the (k_1, k_2)-trimmed variance estimator is

$$\hat{\sigma}^2_{n,k_1,k_2} := \frac{1}{n - k_1 - k_2} \sum_{i=k_1+1}^{n-k_2} (X_{(i)} - \overline{X}_{n,k_1,k_2})^2.$$

The (k_1, k_2)-width is defined as $\Delta_{n,k_1,k_2} := X_{(n-k_2-1)} - X_{(k_1)}$. When $k_1 = k_2 = k$, we write $\overline{X}_{n,k}$ for $\overline{X}_{n,k,k}$, and similarly for the other quantities.

2.3.3 Contaminated data.

We will then consider the trimmed mean computed on the contaminated sample.

Definition 2.6. When $X'_1, \ldots, X'_n$ is a ϵ-contaminated sample satisfying Assumption 2.4, we similarly write

$$X'_{(1)} \leq X'_{(2)} \leq \cdots \leq X'_{(n)}$$

for its order statistics, and

$$\overline{X'}_{n,k_1,k_2} := \frac{1}{n - k_1 - k_2} \sum_{i=k_1+1}^{n-k_2} X'_{(i)} \text{ and } \hat{\sigma'}^2_{n,k_1,k_2} := \frac{1}{n - k_1 - k_2} \sum_{i=k_1+1}^{n-k_2} (X'_{(i)} - \overline{X'}_{n,k_1,k_2})^2$$

to denote trimmed mean and variance computed on the contaminated sample. We define $\hat{\sigma}'_{n,k_1,k_2}$ and Δ'_{n,k_1,k_2} in analogy with the above and with Definition 2.5.

The following facts will be useful when analyzing the contaminated trimmed mean.

Proposition 2.7. *Let $X'_1, \ldots, X'_n$ be an ϵ-contaminated sample and $X_1, \ldots, X_n$ be its uncontaminated version. Assume $k_1, k_2 \in \mathbb{N} \cup \{0\}$ satisfy $\min\{k_1, k_2\} \geq \lfloor \epsilon n \rfloor$ and $k_1 + k_2 < n$ Then:*

$$\overline{X}_{n,k_1-\lfloor \epsilon n \rfloor, k_2+\lfloor \epsilon n \rfloor} \leq \overline{X'}_{n,k_1,k_2} \leq \overline{X}_{n,k_1+\lfloor \epsilon n \rfloor, k_2-\lfloor \epsilon n \rfloor}.$$

Proof. It suffices to prove the following

$$\textbf{Claim: } \forall i \in \{k_1 + 1, \ldots, n - k_2\} \,:\, X_{(i-c)} \leq X'_{(i)} \leq X_{(i+c)}. \tag{2.2}$$

To prove this, notice that

$$X'_{(i)} = \inf\{t \in \mathbb{R} \,:\, \#\{j \in [n] \,:\, X'_j \leq t\} \geq i\}.$$

Now, if we take $t = X_{(i+c)}$ above, we see that $X_j \leq t$ for at least $i + c$ indices $j \in [n]$. Since $X_j = X'_j$ for all but at most c indices j, we conclude:

$$\#\{j \in [n] \,:\, X'_j \leq X_{(i+c)}\} \geq \#\{j \in [n] \,:\, X_j \leq X_{(i+c)}\} - c \geq i.$$

Therefore, $X'_{(i)} \leq X_{(i+c)}$. This proves the upper bound part of the claim, and the lower bound part is similar. □

2.4 General finite-sample bounds

In this section we study the trimmed mean for fixed trimming parameters k_1, k_2 and arbitrary distribution P, as in Assumption 2.1. We show that it is possible to prove a conditional finite-sample result that holds under no additional assumptions. This result – Theorem 2.10 below – says that the trimmed mean concentrates around a "randomly trimmed population mean". Under our stronger Assumptions 2.2 and 2.3, this trimmed population mean will be shown to be close to the parameter μ of interest.

2.4.1 Conditional concentration.

Before we can state our conditional concentration result, we need some preliminaries. Take $X_1, \ldots, X_n$ as in Assumption 2.1, and notice that:

$$(X_1, \ldots, X_n) \sim (F^{-1}(U_1), \ldots, F^{-1}(U_n))$$

where the U_i are i.i.d. uniform over $[0, 1]$. Additionally,

$$(X_{(1)}, \ldots, X_{(n)}) \sim (F^{-1}(U_{(1)}), \ldots, F^{-1}(U_{(n)}))$$

where $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ are the order statistics of the U_i. This leads us to the following fact.

Proposition 2.8 (Proof omitted)**.** *Under Assumption 2.1, one can define (on a richer probability space, if needed) random variables $U_1, \ldots, U_n$ that are i.i.d. uniform over $[0, 1]$, such that $F^{-1}(U_i) = X_i$ and $F^{-1}(U_{(i)}) = X_{(i)}$ almost surely for each $i \in [n]$.*

We now define conditional mean and variance parameters associated with the random variables $F^{-1}(U_i)$.

Definition 2.9 ((a,b)-trimmed population mean, variance and width)**.** Let F^{-1} be the cumulative distribution function quantile transform of P as in Assumption 2.1, Also let U be a uniform random variable over $[0,1]$. Given $0 < a < b < 1$, we define the (a,b)-trimmed population mean and variance as

$$\mu(a,b) := \mathbb{E}\left[F^{-1}(U) \mid a < U < b\right] = \frac{1}{b-a}\int_a^b F^{-1}(u)\,du$$

and

$$\sigma^2(a,b) := \mathbb{V}\left[F^{-1}(U) \mid a < U < b\right] = \frac{1}{b-a}\int_a^b (F^{-1}(u) - \mu(a,b))^2\,du.$$

We also define the (a,b)-widthz as $\Delta(a,b) := F^{-1}(b) - F^{-1}(a)$.

Our first finite-sample result shows that the sample trimmed mean concentrates around a trimmed population mean, and the same holds for the variance.

Theorem 2.10. *Let $X_1, \ldots, X_n$ satisfy Assumption 2.1. Assume additionally that we have defined i.i.d.* Uniform$[0,1]$ *random variables $U_1, \ldots, U_n$ over the same probability space, with $X_i = F^{-1}(U_i)$ for each $i \in [n]$ (as per Proposition 2.8). Let $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ denote the order statistics of the U_i. Choose parameters $k_1, k_2 > 0$. Then, almost surely, conditionally on $U_{(k_1)}$ and $U_{(n-k_2+1)}$:*

1. *given $\alpha_1 \in (0,1)$, the following inequality holds with (conditional) probability at least $1-\alpha_1$:*

$$\begin{aligned} &\overline{X}_{n,k_1,k_2} - \mu(U_{(k_1)}, U_{(n-k_2+1)}) \\ \le\ &\sigma(U_{(k_1)}, U_{(n-k_2+1)})\sqrt{\tfrac{2\log(1/\alpha_1)}{n-k_1-k_2}} + \Delta_{n,k_1,k_2}\tfrac{\log(1/\alpha_1)}{3(n-k_1-k_2)}; \end{aligned}$$

2. *given $\alpha_2 \in (0,1)$, the following two inequalities hold simultaneously with probability $\ge 1-\alpha_2$:*

$$\sigma(U_{(k_1)}, U_{(n-k_2+1)}) \le |\overline{X}_{n,k_1,k_2} - \mu(U_{(k_1)}, U_{(n-k_2+1)})| + \hat{\sigma}_{n,k_1,k_2} + 2\Delta_{n,k_1,k_2}\sqrt{\frac{\log(2/\alpha_2)}{(n-k_1-k_2)}}$$

and

$$\hat{\sigma}_{n,k_1,k_2} \le \sigma(U_{(k_1)}, U_{(n-k_2+1)}) + \Delta_{n,k_1,k_2}\sqrt{\frac{\log(2/\alpha_2)}{2(n-k_1-k_2)}};$$

3. *when the events described in items 1 and 2 both hold, and additionally* $2\log(1/\alpha_1) < n - k_1 - k_2$*, we also have:*

$$\overline{X}_{n,k_1,k_2} - \mu(U_{(k_1)}, U_{(n-k_2+1)}) \leq \frac{\hat{\sigma}_{n,k_1,k_2}\sqrt{\frac{2\log(1/\alpha_1)}{n-k_1-k_2}} + c_0\,\Delta_{n,k_1,k_2}\,\frac{\log(1/\alpha_{\min})}{(n-k_1-k_2)}}{1-\sqrt{\frac{2\log(1/\alpha_1)}{n-k_1-k_2}}},$$

where $\alpha_{\min} := \max\{\alpha_1, \alpha_2/2\}$ *and* $c_0 := 1/\sqrt{2} + \sqrt{5/6}$.

Finally, when P *is supported on* $[0, +\infty)$*, the inequalities above also work for* $k_1 = 0$ *if we set* $U_{(0)} = 0$.

One interesting aspect of this theorem is that it can be used to obtain a confidence interval around the trimmed population mean $\mu(U_{(k_1)}, U_{(n-k_2+1)})$. This is the case because both $\hat{\sigma}_{n,k_1,k_2}$ and Δ_{n,k_1,k_2} can be computed from data. However, $\mu(U_{(k_1)}, U_{(n-k_2+1)})$ is a random quantity because the pair $(U_{(k_1)}, U_{(n-k_2+1)})$ is random. Much of what we do in the rest of the chapter is to find conditions under which this random quantity is close to the mean or median of P. The results in the next subsection are a first step in this direction.

Proof of Theorem 2.10. We only consider the case where $k_1 > 0$, as the proof for nonnegative variables (and $k_1 = 0$) is entirely analogous.

For the remainder of the proof, we condition on $U_{(k_1)} = a < U_{(n-k_2+1)} = b$. Under this conditioning:

$$(U_{(i)})_{i=k_1+1}^{n-k_2} \,|_{U_{(k_1)}=a, U_{(n-k_2+1)}=b} \sim (V_{(j)})_{j=1}^{n-k_1-k_2}$$

where the $V_{(j)}$ are the order statistics of an i.i.d. random sample

$$(V_1, V_2, \ldots, V_{n-k_1-k_2}) \overset{i.i.d.}{\sim} \text{Uniform}[a, b].$$

Therefore, the conditional distribution of the trimmed mean,

$$\overline{X}_{n,k_1,k_2} = \frac{1}{n-k_1-k_2}\sum_{i=k_1+1}^{n-k_2} F^{-1}(U_{(i)}),$$

is

$$\overline{X}_{n,k_1,k_2} \,|_{U_{(k_1)}=a, U_{(n-k_2+1)}=b} \sim \frac{1}{n-k_1-k_2}\sum_{j=1}^{n-k_1-k_2} F^{-1}(V_{(j)}) = \frac{1}{n-k_1-k_2}\sum_{i=1}^{n-k_1-k_2} F^{-1}(V_j).$$

The RHS is a sum of i.i.d. terms with mean $\mu(a, b)$ and variance $\sigma^2(a, b)$. We also know that the $F^{-1}(V_i)$ take values in the interval $[F^{-1}(a), F^{-1}(b)]$, which has length $\Delta(a, b)$. Therefore, the

first statement in the Theorem is a direct application of Bernstein's inequality to the conditional distribution of X_{n,k_1,k_2}.

For the second part, we notice that

$$\hat{\sigma}^2_{n,k_1,k_2} = H^2 - (\mu(a,b) - \overline{X}_{n,k_1,k_2})^2, \text{ where } H^2 := \frac{1}{n-k_1-k_2}\sum_{i=k_1+1}^{n-k_2}(X_{(i)} - \mu(a,b))^2.$$

Under our conditioning, H is an i.i.d. sum:

$$H^2 \mid_{U_{(k_1)}=a, U_{(n-k_2+1)}=b} \sim \frac{1}{n-k_1-k_2}\sum_{i=1}^{n-k_1-k_2}(F^{-1}(V_j) - \mu(a,b))^2.$$

The terms in sum in the RHS have mean $\sigma^2(a,b)$, are bounded by $\Delta(a,b)^2$ in absolute value, and have variance:

$$\mathbb{V}\left[(F^{-1}(V_j) - \mu(a,b))^2\right] \leq \mathbb{E}\left[(F^{-1}(V_j) - \mu(a,b))^4\right] \leq \Delta(a,b)^2\,\sigma^2(a,b).$$

Therefore, Bernstein's inequality implies that, with probability $\geq 1-\alpha_2$,

$$\left|H^2 - \sigma^2(a,b)\right| \leq \Delta(a,b)\,\sigma(a,b)\sqrt{\frac{2\log(2/\alpha_2)}{n-k_1-k_2}} + \frac{\Delta(a,b)^2\log(2/\alpha_2)}{3\,(n-k_1-k_2)}. \tag{2.3}$$

When (2.3) holds, we have:

$$\hat{\sigma}^2_{n,k_1,k_2} \leq H^2 \leq \left(\sigma(a,b) + \Delta(a,b)\sqrt{\frac{\log(2/\alpha_2)}{2(n-k_1-k_2)}}\right)^2,$$

so that:

$$\hat{\sigma}_{n,k_1,k_2} \leq \sigma(a,b) + \Delta(a,b)\sqrt{\frac{\log(2/\alpha_2)}{2(n-k_1-k_2)}}. \tag{2.4}$$

Under (2.3), we also have that:

$$\left(\sigma(a,b) - \Delta(a,b)\sqrt{\frac{\log(2/\alpha_2)}{2(n-k_1-k_2)}}\right)^2 \leq H^2 + \frac{5\Delta(a,b)^2\log(2/\alpha_2)}{6(n-k_1-k_2)},$$

and we may use the formula for H^2 and sub-additivity of the square root function to obtain:

$$\sigma(a,b) \leq \hat{\sigma}_{n,k_1,k_2} + |\overline{X}_{n,k_1,k_2} - \mu(a,b)| + c_0\,\Delta(a,b)\sqrt{\frac{\log(2/\alpha_2)}{(n-k_1-k_2)}}, \tag{2.5}$$

where $c_0 = 1/\sqrt{2} + \sqrt{5/6} \leq 2$. Thus (2.4) and (2.5) – the two inequalities claimed in item 2 of the Theorem – follow from (2.3), which holds with probability $\geq 1-\alpha_2$. To prove item 3 in the theorem, we just observe that, if $X_{n,k_1,k_2} - \mu(a,b) < 0$, there is nothing to prove. Otherwise, one may plug (2.5) into the inequality in item 1 and obtain the desired result after some manipulations. □

2.4.2 Bounds on the trimmed population mean and variance.

We now discuss how one can control the two trimmed population parameters appearing in Theorem 2.10. We start with a result that holds under moment conditions.

Lemma 2.11. *Under Assumption 2.2 (whereby P has mean μ), for any $p > 1$ and $0 \le a < b \le 1$,*

$$|\mu(a,b) - \mu| \le \inf_{p>1} \nu_p \frac{\delta^{1-1/p}}{1-\delta},$$

where

$$\delta := 1 - (b-a).$$

Proof. Since $\mu = \int_0^1 F^{-1}(u)\,du$,

$$\mu(a,b) - \mu = \frac{\int_a^b (F^{-1}(u) - \mu)\,du}{1-\delta} = -\frac{\int_{[0,1]\setminus[a,b]} (F^{-1}(u) - \mu)\,du}{1-\delta}.$$

The integral in the RHS can be rewritten as

$$-\int_{[0,1]} (F^{-1}(u) - \mu)\mathbf{1}_A\,du, \text{ where } A := [0,1]\setminus[a,b].$$

Since the Lebesgue measure of A is δ, we obtain, for any $p > 1$:

$$|\mu(a,b) - \mu| \le \frac{\left(\int_{[0,1]} |F^{-1}(u) - \mu|^p\,du\right)^{\frac{1}{p}} \delta^{1-1/p}}{1-\delta} = \nu_p \frac{\delta^{1-\frac{1}{p}}}{1-\delta}.$$

from Hölder's inequality. Taking the infimum over p finishes the proof. □

We now consider what happens when P is symmetric about its median.

Lemma 2.12. *Under Assumption 2.3 (whereby P has median μ), for any $p > 1$ and $0 \le a < 1/2 < b \le 1$, we have:*

$$|\mu(a,b) - \mu| \le \Delta(a,b)\ \frac{\eta}{1-\delta},$$

where

$$\delta := 1 - (b-a) \text{ and } \eta = \max\{b - (1-a), a - (1-b)\}.$$

Proof. Assume without loss that $b \ge 1 - a$. By symmetry, we have $\mu = \mu(a, 1-a) \in$

$[F^{-1}(a), F^{-1}(b)]$. Therefore:

$$\begin{aligned}
\mu(a,b) &= \frac{1}{b-a}\int_a^b F^{-1}(u)\,du \\
&= \frac{1-2a}{b-a}\left(\frac{1}{1-2a}\int_a^{1-a} F^{-1}(u)\,du\right) \\
&\quad + \frac{b-1+a}{b-a}\left(\frac{1}{b-1+a}\int_{1-a}^b F^{-1}(u)\,du\right) \\
&= \left(\frac{1-2a}{b-a}\right)\mu + \left(\frac{b-1+a}{b-a}\right) R,
\end{aligned}$$

with $R \in [F^{-1}(a), F^{-1}(b)]$. So $\mu(a,b)$ is a convex combination of μ and R. Both of these values lie in the interval $[F^{-1}(a), F^{-1}(b)]$, which has length $\Delta(a,b)$. We deduce:

$$|\mu(a,b) - \mu| \leq \left(\frac{b-1+a}{b-a}\right)|R-\mu| \leq \Delta(a,b)\,\frac{\eta}{1-\delta}.$$

□

Finally, we compare $\sigma(a,b)$ to the population moments.

Lemma 2.13. *Make Assumption 2.2. Fix $0 \leq a < b \leq 1$ and set $\delta := 1-(b-a)$.*

$$\sigma(a,b) \leq \frac{\sigma}{1-\delta},$$

and more generally, for any $1 < q \leq 2$,

$$\sigma(a,b) \leq \frac{(\nu_p)^{\frac{q}{2}}\,\Delta(a,b)^{1-\frac{q}{2}}}{1-\delta}.$$

For $p > 2$, if $\nu_p < +\infty$ and $4\delta^{1-\frac{2}{p}}\,(\nu_p^2/\sigma^2) < 1$, then:

$$\sigma \leq \frac{(1-\delta)\,\sigma(a,b)}{\sqrt{1-4\delta^{1-\frac{2}{p}}\,(\nu_p^2/\sigma^2)}}.$$

Proof. The first statement follows from a simple chain of inequalities:

$$\begin{aligned}
\sigma^2(a,b) &= \frac{1}{2\,(b-a)^2}\int_a^b\int_a^b (F^{-1}(u)-F^{-1}(v))^2\,du\,dv \\
(\text{integrand is} \geq 0) &\leq \frac{1}{2\,(1-\delta)^2}\int_0^1\int_0^1 (F^{-1}(u)-F^{-1}(v))^2\,du\,dv \\
&= \frac{\sigma^2}{(1-\delta)^2}.
\end{aligned}$$

The second statement is similar, as

$$\forall (u,v) \in [a,b]^2 \,:\, (F^{-1}(u) - F^{-1}(v))^2 \leq \Delta(a,b)^{2-q}\, |F^{-1}(u) - F^{-1}(v)|^q,$$

so that:

$$\sigma(a,b)^2 \leq \frac{\Delta(a,b)^{2-q}}{2(1-\delta)^2} \int_0^1 \int_0^1 |F^{-1}(u) - F^{-1}(v)|^q \, du\, dv.$$

The integral in the RHS is

$$\|X - Y\|_{L^q}^q \text{ where } X, Y \sim P \text{ are independent},$$

which (by convexity) is at most $2^{q-1}\|X-\mu\|_{L^q}^q = 2^{q-1}\,\nu_q^q$. We now consider the final statement in the Lemma. Notice that:

$$\sigma^2 = (1-\delta)^2\,\sigma^2(a,b) + \frac{1}{2}\int\int_A (F^{-1}(u) - F^{-1}(v))^2 \, du\, dv$$

where $A := [0,1]^2 - [a,b]^2$ has Lebesgue measure $1-(b-a)^2 \leq 2\delta$. Hölder's inequality implies:

$$\frac{1}{2}\int\int_A (F^{-1}(u) - F^{-1}(v))^2 \, du\, dv \leq \frac{(2\delta)^{1-\frac{2}{p}}}{2} \left(\int\int_{[0,1]^2} (F^{-1}(u) - F^{-1}(v))^p \, du\, dv \right)^{\frac{2}{p}}.$$

As before, we recognize that:

$$\int\int_{[0,1]^2} (F^{-1}(u) - F^{-1}(v))^p \, du\, dv = \|X - Y\|_{L^p}^p \leq 2^{p-1}\nu_p$$

(here $X, Y \sim P$ are independent). Thus:

$$\frac{1}{2}\int\int_A (F^{-1}(u) - F^{-1}(v))^2 \, du\, dv \leq \frac{(2\delta)^{1-\frac{2}{p}}\, 2^{2-\frac{2}{p}}}{2}\,\nu_p^2 = 2^{2-\frac{4}{p}}\,\nu_p^2\,\delta^{1-\frac{2}{p}}.$$

We obtain:

$$\sigma^2 \leq (1-\delta)^2\,\sigma^2(a,b) + 4\delta^{1-\frac{2}{p}}\,\nu_p^2,$$

or

$$\sigma^2 \leq \frac{(1-\delta)^2\,\sigma^2(a,b)}{1 - 4\delta^{1-\frac{2}{p}}\,(\nu_p^2/\sigma^2)},$$

□

2.5 Estimation of the mean under moment conditions

In this section, we consider the performance of the trimmed mean under moment conditions (Assumption 2.2). In § 2.5.1 the trimmed mean under arbitrary (and possibly asymetrical) trimming. Subsequently, § 2.5.2 and § 2.5.3 show that a single choice k (in terms of the confidence parameter $1-\alpha$ and the contamination level ϵ) achieves minimax-optimal results under arbitrary moment conditions.

Remark 2.14. The results in this section all depend on a choice of trimming that depends on both ϵ and α. It is natural to ask if some choice of k could give minimax results for a wide range of α and ϵ. The answer turns out to be no, for the following reasons:

- k must depend on ϵ because a trimmed mean estimator that trims less than ϵn data points can be arbitrarily manipulated by changing ϵn points in the sample;
- the dependence on α is necessary when all one assumes is finite variance [Devroye et al., 2016, Theorem 3.2].

On the other hand, if $\epsilon = 0$ and a finite upper bound for $M_p = \nu_p/\sigma$ is known for some $p > 2$, one could use the ideas in [Devroye et al., 2016, Section 7] to obtain a single value of k that "works well" for a large range of α. Alternatively, one may use the construction of sub-Gaussian confidence intervals in §2.6.2, together with [Devroye et al., 2016, Theorem 4.2].

2.5.1 Results under arbitrary trimming choices.

In this first result, we consider what happens for an essentially arbitrary (but data-independent) choice of trimming.

Theorem 2.15. *Make assumptions 2.1 and 2.2. Then for any $k_1, k_2 > 0$ and $\alpha \in (0,1)$ satisfying:*

$$\phi := \frac{(\sqrt{k_1 + k_2 - 1} + \sqrt{\log(4/\alpha)})^2}{n} < 1$$

the following holds with probability $\geq 1-\alpha$*:*

$$\begin{aligned}\overline{X}_{n,k_1,k_2} - \mu \;\leq\; & \inf_{1\leq q\leq 2} \frac{e^{\frac{1}{q}-\frac{1}{2}}\nu_q}{(1-\phi)\, k_{\min}^{\frac{1}{q}-\frac{1}{2}}\, n^{\frac{q-1}{q}}} \sqrt{\frac{2\log(2/\alpha)}{1-\frac{k_1+k_2}{n}}} \\ & + \inf_{p>1} \nu_p \frac{\phi^{\frac{p-1}{p}}}{1-\phi} \\ & + 2 \inf_{p>1} \left(\frac{4}{\alpha}\right)^{\frac{1}{pk_{\min}}} e^{\frac{1}{p}} \nu_p \frac{\log(4/\alpha)}{3\left(1-\frac{k_1+k_2}{n}\right) k_{\min}^{\frac{1}{p}}\, n^{\frac{p-1}{p}}},\end{aligned}$$

where $k_{\min} := \min\{k_1, k_2\}$*. The same result holds when* P *is supported on* $[0,+\infty)$*, if we now set* $k_1 = 0$ *and redefine* $k := k_2$*.*

Proof of Theorem 2.15. We prove only the bound for general distributions P, as the case of P supported over $[0,+\infty)$ follows from similar ideas.

We work in the framework of Proposition 2.8 and Theorem 2.10, whereby we have i.i.d. uniform random variables U_i with $F^{-1}(U_i) = X_i$. Let:

$$B := 1 - (U_{(n-k_2+1)} - U_{(k_1)}),$$

noting that $0 < B < 1$ almost surely. Lemmas 2.11 and 2.13 give the following bounds on the trimmed population mean and variance that show up in Theorem 2.10:

$$\begin{aligned}|\mu(U_{(k_1)}, U_{(n-k_2+1)}) - \mu| \;&\leq\; \inf_{p>1} \nu_p \frac{B^{\frac{p-1}{p}}}{1-B}; \\ \sigma(U_{(k_1)}, U_{(n-k_2+1)}) \;&\leq\; \inf_{1<q\leq 2} \frac{\Delta_{n,k_1,k_2}^{1-\frac{q}{2}}\, \nu_q^{\frac{q}{2}}}{1-B}.\end{aligned}$$

Combining Theorem 2.10 with the above, and using the fact that $k_1 + k_2 \leq \phi n$, we conclude that the following inequality holds with probability $\geq 1-\alpha/2$:

$$\begin{aligned}X_{n,k_1,k_2} - \mu \;\leq\; & \inf_{1<q\leq 2} \frac{\Delta_{n,k_1,k_2}^{1-\frac{q}{2}}\, \nu_q^{\frac{q}{2}}}{1-B} \sqrt{\frac{2\log(2/\alpha)}{n\left(1-\frac{k_1+k_2}{n}\right)}} \\ & + \frac{\inf_{p>1} \nu_p\, B^{\frac{p-1}{p}}}{1-B} \\ & + \Delta_{n,k_1,k_2} \frac{\log(2/\alpha)}{3n\left(1-\frac{k_1+k_2}{n}\right)}.\end{aligned}$$

To finish the proof, we will show that:

$$\textbf{Goal 1: } \mathbb{P}\left[B > \phi\right] \leq \frac{\alpha}{4}. \tag{2.6}$$

and

$$\textbf{Goal 2: } \mathbb{P}\left[\Delta_{n,k_1,k_2} > \inf_{p>1} 2\nu_p \left(\frac{4}{\alpha}\right)^{\frac{1}{pk_{\min}}} \left(\frac{en}{k_{\min}}\right)^{\frac{1}{p}}\right] \leq \frac{\alpha}{4}. \tag{2.7}$$

To obtain (2.6), we note that:

$$B \sim U_{(k_1+k_2)}$$

and apply Lemma A.1 in the Appendix.

To prove (2.7), recall $k_{\min} = \min\{k_1, k_2\}$, and let A be the $k_{\min}$-th largest value of $|X_i - \mu|$. Note that $X_{(k_1)} - \mu \geq -A$ and $X_{(n-k_2+1)} - \mu \leq A$, so $\Delta_{n,k_1,k_2} \leq 2A$. Thus (2.7) is equivalent to:

$$\textbf{Goal 2 (sufficient): } \forall p > 1 : \mathbb{P}\left[A > \nu_p \left(\frac{4}{\alpha}\right)^{\frac{1}{pk_{\min}}} \left(\frac{en}{k_{\min}}\right)^{\frac{1}{p}}\right] < \frac{\alpha}{4}. \tag{2.8}$$

Notice that we moved the condition on p to outside the probability, but it is easy to move between one bound and the other via limiting arguments.

For any $x > 0$, $A > x$ if and only if there is a subset $S \subset [n]$ of size $\#S = k_{\min}$ such that $|X_i - \mu| > x$ for all $i \in S$. Taking a union bound, and using the independence of the X_i, we obtain (for any $p > 1$):

$$\begin{aligned}
\mathbb{P}[A_k > x] &\leq \sum_{S\subset[n],\, \#S=k_{\min}} \prod_{i\in S} \mathbb{P}[|X_i - \mu| > x] \\
&\leq \binom{n}{k_{\min}} \left(\frac{\nu_p}{x}\right)^{pk_{\min}} \leq \left[\left(\frac{en}{k_{\min}}\right)^{\frac{1}{p}} \frac{\nu_p}{x}\right]^{pk_{\min}},
\end{aligned}$$

where the last inequality above follows from the well-known bound:

$$\binom{n}{k_{\min}} \leq \left(\frac{en}{k_{\min}}\right)^{k_{\min}}.$$

The choice of

$$x_* := \nu_p \left(\frac{4}{\alpha}\right)^{\frac{1}{pk_{\min}}} \left(\frac{en}{k_{\min}}\right)^{\frac{1}{p}}$$

ensures $\mathbb{P}[A > x_*] \leq \alpha/4$. This gives us (2.8) and finishes the proof. □

2.5.2 Minimax results under finite variance.

We now argue that the trimmed mean achieves minimax results in terms of the confidence level $1 - \alpha$ and the contamination level ϵ under a range of moment assumptions on the data distribution. A key point of our result is that a *universal choice* of trimming will work in all cases.

Assumption 2.16 (Universal choice of trimming)**.** In what follows we fix $\epsilon \in [0, 1/2)$, $\alpha \in (0, 1)$, set:

$$k = k(\epsilon, \alpha) := \lfloor \epsilon n \rfloor + \lceil \log(8/\alpha) \rceil$$

and assume

$$\phi(\epsilon, \alpha) := \frac{(\sqrt{2k(\epsilon, \alpha) - 1} + \sqrt{\log(8/\alpha)})^2}{n} < 1.$$

.

The first result considers what happens under a finite-variance condition.

Theorem 2.17 (Finite-variance minimax performance)**.** *Make assumptions 2.1 (data is i.i.d. contaminated), 2.2 (moment conditions), and 2.16 (choice of k). Assume the variance of P is finite and fix $p \geq 2$, $\alpha \in (0, 1)$. Then the contaminated trimmed mean $\overline{X'}_{n,k(\epsilon,\alpha)}$ satisfies the following bound with probability at least $1 - \alpha$:*

$$|\overline{X'}_{n,k(\epsilon,\alpha)} - \mu| \leq \sigma \sqrt{\frac{2 \log(4/\alpha)}{n}} \left(1 + \delta_p(\epsilon, \alpha, n)\right) + \nu_p \frac{(4\epsilon)^{\frac{p-1}{p}}}{1 - \phi},$$

where

$$\delta_p(\epsilon, \alpha, n) := \frac{1}{(1 - \phi)^{3/2}} - 1 + \left(\frac{2\, e^{\frac{2}{p}}}{3\,(1 - \phi)} + 3^{\frac{p-1}{p}} \right) \frac{\nu_p}{\sqrt{2}\, \sigma} \left(\frac{\log(8/\alpha)}{n} \right)^{\frac{1}{2} - \frac{1}{p}}$$

and $C > 0$ is universal.

To understand this theorem, assume first $p = 2$ (finite variance only) and $\epsilon = 0$ (ie. no contamination). Then $\nu_p = \sigma$ and the theorem implies the following. There exist $c, C' > 0$ such that if $\alpha \geq 4e^{-cn}$, then:

$$\mathbb{P}\left[|\overline{X}_{n,k} - \mu| \leq C'\, \sigma \sqrt{\frac{2 \log(4/\alpha)}{n}} \right] \geq 1 - \alpha.$$

This is the kind of sub-Gaussian behavior studied in [Catoni, 2012] and [Devroye et al., 2016], which is minimax-optimal up to the value of C'.

Now assume $\nu_p < +\infty$ for some $p > 2$ and keep $\epsilon = 0$. The parameter $M_p := \nu_p / \sigma \geq 1$ is related to the kurtosis of a distribution; in fact, the kurtosis is precisely M_4^4. Our results imply that there exists an absolute constant $d > 0$ such that, for any $h \in (0, 1)$, if

$$\log(4/\alpha) \leq (d\, M_p)^{\frac{2p}{p-1}}\, n,$$

then $\delta_p(\alpha, \epsilon, n) \leq h$, and:

$$\mathbb{P}\left[|\overline{X}_{n,k} - \mu| \leq (1+h)\,\sigma\sqrt{\frac{2\log(4/\alpha)}{n}}\right] \geq 1-\alpha.$$

Thus kurtosis-style assumptions imply that the trimmed mean achieves a nearly optimal constant $(1+h)\sqrt{2}$ in front of the "sub-Gaussian" term; see [Catoni, 2012, Proposition 6.1] and [Devroye et al., 2016, Remark 1] for details.

The effect of contamination is twofold. First, it increases the value of $\delta_p(\epsilon, \alpha, n)$ by a bounded amount, as long as ϕ is bounded away from 1. Additionally, $\epsilon > 0$ causes an additional term $\nu_p\,\epsilon^{(p-1)/p}$ to appear in the bound. The latter term is a minimax-optimal error term under contamination, for any $p \geq 2$ [Minsker, 2018, Lemma 5.4][1]. We emphasize that Theorem 2.17 achieves optimal results even though the estimator does not depend at all on any properties of P.

Proof of Theorem 2.17. Set $k := k(\epsilon, \alpha)$. By Proposition 2.7,

$$\overline{X}_{n,k-\lfloor \epsilon n\rfloor, k+\lfloor \epsilon n\rfloor} \leq \overline{X'}_{n,k} \leq \overline{X}_{n,k+\lfloor \epsilon n\rfloor, k-\lfloor \epsilon n\rfloor}.$$

Theorem 2.15 (with $\alpha/2$ replacing α and $q=2$) can be applied to the upper and lower quantities appearing above; in both cases, $k_{\min} = \lceil \log(8/\alpha)\rceil$. As a consequence, the following inequality holds with probability $\geq 1-\alpha$:

$$\begin{aligned} |\overline{X'}_{n,k} - \mu| \;\leq\; & \frac{\sigma}{1-\phi}\sqrt{\frac{2\log(4/\alpha)}{n-2k}} \\ & + \nu_p\,\frac{\phi^{\frac{p-1}{p}}}{1-\phi} \\ & + 2\,e^{\frac{2}{p}}\,\nu_p\,\frac{(\log(4/\alpha))^{\frac{p-1}{p}}}{3\,\left(1-\frac{k_1+k_2}{n}\right)\,n^{\frac{p-1}{p}}}. \end{aligned}$$

To continue the proof, note that:

$$\phi \leq \frac{(\sqrt{2k-1} + \sqrt{\log(8/\alpha)})^2}{n} \leq \frac{4k - 2 + 2\log(8/\alpha)}{n},$$

and

$$k \leq \epsilon n + \lceil \log(4/\alpha)\rceil \leq \epsilon n + \log(8/\alpha) + 1.$$

We obtain:

$$\phi \leq 4\epsilon + \frac{3\log(8/\alpha)}{n},$$

[1]The result in [Minsker, 2018] assumes $2 \leq p \leq 3$ but the same argument works for any $p > 1$.

and

$$\nu_p \, \phi^{\frac{p-1}{p}} \leq \nu_p \, (4\epsilon)^{\frac{p-1}{p}} + \nu_p \left(\frac{3 \log(8/\alpha)}{n} \right)^{\frac{p-1}{p}}.$$

Using also that $n - 2k \geq (1 - \phi) \, n$, we obtain that, with probability $\geq 1 - \alpha$:

$$\begin{aligned} |\overline{X'}_{n,k} - \mu| \;\; \leq \;\; & \frac{\sigma}{(1-\phi)^{3/2}} \sqrt{\frac{2 \log(4/\alpha)}{n}} \\ & + \nu_p \, \frac{(4\epsilon)^{\frac{p-1}{p}}}{1-\phi} \\ & + \left(\frac{2 \, e^{\frac{2}{p}}}{3 \, (1-\phi)} + 10^{\frac{p-1}{p}} \right) \nu_p \left(\frac{3 \log(8/\alpha)}{n} \right)^{\frac{p-1}{p}}, \end{aligned}$$

which is the same as what we claim in the Theorem. □

2.5.3 Minimax results under possibly infinite variance.

Let us now state a more general version (with less precise constants) of the Theorem 2.17, which also works when the variance is infinite. As before, the theorem is minimax optimal up to constants, in terms of both α [Devroye et al., 2016, Theorem 3.1] and the contamination level. Notice that the choice of trimming is the same $k = k(\epsilon, \alpha)$ as in Theorem 2.17.

Theorem 2.18 (General minimax bounds)**.** *Make assumptions 2.1 (i.i.d. data with contamination), 2.2 (moment conditions), and 2.16 (choice of k). Fix $p > 1$ and $1 < q \leq 2$. Then the contaminated trimmed mean $\overline{X'}_{n,k}$ satisfies the following bound with probability $\geq 1 - \alpha$:*

$$|\overline{X'}_{n,k(\epsilon,\alpha)} - \mu| \leq \frac{C}{(1-\phi)^{3/2}} \left(\nu_q \left(\frac{\log(8/\alpha)}{n} \right)^{\frac{q-1}{q}} + C \, \nu_p \, \epsilon^{\frac{p-1}{p}} \right),$$

where $C > 0$ is universal.

Proof. The proof is very similar (although simpler) than that of Theorem 2.17; we omit the details. □

2.6 Finite-sample confidence intervals

Our goal in this section is to quantify the uncertainty in the trimmed mean estimate via confidence intervals.

The discussion right after Theorem 2.10 implies that in order to find such an interval, we must give a data-dependent estimate on the quantity $|\mu(U_{(k_1)}, U_{(n-k_1-k_2)} - \mu|$. In what follows, we identify two settings where that is possible: symmetrical data, and data satisfying kurtosis-type assumptions.

2.6.1 Data that is symmetrical around the median.

The case of P that symmetrical around the median μ corresponds to much classical work on Robust Statistics. There, it is common to assume that the uncontaminated distribution and its Huber contamination both have the same median. In that sense, our distribution P may correspond to a Huber-contaminated sample. Importantly, however, we make no moment assumptions: P may not even have a mean.

The next result gives a finite-sample confidence interval in this setting.

Theorem 2.19. *Make Assumption 2.1 with $\epsilon = 0$ (i.i.d. uncontaminated data), and also Assumption 2.3 (P is symmetrical around the median μ). Fix an integer $k > 0$, a confidence parameter $1 - \alpha \in (0, 1)$, and assume $2\sqrt{2k \log(5e/\alpha)} + 2\log(5/\alpha) + 2k < n$. Define a width parameter:*

$$\begin{aligned}\widehat{w}_{n,k}(\alpha) \;:=\; & \frac{\hat{\sigma}_{n,k}\sqrt{2\log(4/\alpha)}}{\sqrt{n-2k} - \sqrt{2\log(4/\alpha)}} \\ & + \Delta_{n,k}\left(\frac{(1/\sqrt{2} + \sqrt{5/6})\,\log(4/\alpha)}{(n-2k)\left(1 - \sqrt{\frac{2\log(4/\alpha)}{n-2k}}\right)} + \frac{\sqrt{2k\log(5e/\alpha)} + \log(5/\alpha)}{n - 2k - 2\sqrt{2k\log(5e/\alpha)} - 2\log(5/\alpha)}\right).\end{aligned}$$

Then the random interval:

$$\widehat{I}_{n,k}(\alpha) := \left[\overline{X}_{n,k} - \widehat{w}_{n,k}(\alpha), \overline{X}_{n,k} + \widehat{w}_{n,k}(\alpha)\right]$$

satisfies:

$$\mathbb{P}\left[\mu \in \widehat{I}_{n,k}(\alpha)\right] \geq 1 - \alpha.$$

Proof. The first step in the proof is to apply Theorem 2.10 (with $\alpha_1 = \alpha_2 = \alpha/4$) and obtain that, with probability $\geq 1 - 3\alpha/4$),

$$|\overline{X}_{n,k} - \mu(U_{(k)}, U_{(n-k+1)})| \leq \frac{\hat{\sigma}_{n,k}\sqrt{\frac{2\log(4/\alpha)}{n-2k}} + (1/\sqrt{2} + \sqrt{5/6})\,\Delta_{n,k}\,\frac{\log(4/\alpha)}{(n-2k)}}{1 - \sqrt{\frac{2\log(4/\alpha)}{n-2k}}}.$$

To finish, we argue that:

$$\textbf{Goal: } \mathbb{P}\left[|\mu(U_{(k)}, U_{(n-k+1)}) - \mu| > \Delta_{n,k} \frac{\sqrt{2k\log(5e/\alpha)} + \log(5/\alpha)}{n - 2k - 2\sqrt{2k\log(5e/\alpha)} - 2\log(5/\alpha)}\right] \leq \frac{\alpha}{4}.$$

As a preliminary step, define:

$$V := \max\left\{\left|U_{(k)} - \frac{k}{n}\right|, \left|1 - U_{(n-k+1)} - \left(1 - \frac{k}{n}\right)\right|\right\},$$

and assume $V \leq 1/2 - k/n$ (which implies $U_{(k)} \leq 1/2 \leq U_{(n-k+1)}$). Then Lemma 2.12 and some simple estimates give:

$$|\mu(U_{(k)}, U_{(n-k+1)}) - \mu| \leq \Delta_{n,k} \frac{V}{1 - \frac{2k}{n} - V}, \tag{2.9}$$

We now observe that, by Lemma A.1 in the appendix, for any $t \geq 0$:

$$\mathbb{P}\left[\frac{(\sqrt{k} - \sqrt{t})^2}{n} \leq U_{(k)} \leq \frac{(\sqrt{k-1} + \sqrt{t})^2}{n}\right] \geq 1 - e^{-t} - e^{-2t},$$

which implies:

$$\mathbb{P}\left[\left|U_{(k)} - \frac{k}{n}\right| \leq \frac{2\sqrt{kt} + t - 1}{n}\right] \geq 1 - e^{-t} - e^{-2t}.$$

Similarly,

$$\mathbb{P}\left[\left|1 - U_{(n-k+1)} - \frac{k}{n}\right| \leq \frac{2\sqrt{kt} + t - 1}{n}\right] \geq 1 - e^{-t} - e^{-2t}.$$

Taking

$$t := \log(5e/\alpha)$$

gives:

$$\mathbb{P}\left[V \leq \frac{\sqrt{2k\log(5e/\alpha)} + \log(5/\alpha)}{n}\right] \geq 1 - \frac{2\alpha}{5e} - \frac{2\alpha^2}{25e^2} \geq 1 - \frac{\alpha}{4}.$$

Finally, when

$$V \leq \frac{\sqrt{2k\log(5e/\alpha)} + \log(5/\alpha)}{n}$$

we have $V \leq 1/2 - k/n$ by our assumption on k and α. We conclude from (2.9) that our goal holds with the desired probability $\geq 1 - \alpha/4$. □

2.6.2 Sub-Gaussian confidence intervals.

Our next step is to give a confidence interval for the estimator $\overline{X}_{n,k}$ in a contamination-free setting. It follows from [Devroye et al., 2016, Theorems 3.2 and 4.2] that it is not possible to obtain intervals of "sub-Gaussian" length under the sole assumption of finite second moment.

Theorem 2.20. *Make Assumptions 2.1 and 2.2. Add the assumptions that $\sigma < +\infty$ and $\nu_p \leq M\,\sigma$ for some $M \geq 1$ and $p > 1$. Let $\alpha \in (0,1)$. Define $k_* := \lceil \log(8/\alpha) \rceil$ and assume*

$$\phi_* = \frac{(\sqrt{2k_*-1} + \sqrt{\log(8/\alpha)})^2}{n} \leq \frac{1}{2\,(4M^2)^{p/(p-2)}}.$$

Assume additionally that

$$\sqrt{\frac{2\log(8/\alpha)}{n-2k_*}} + 2\frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}} < 1,$$

and set:

$$A := \frac{1}{1-\sqrt{\frac{2\log(2/\alpha_1)}{n-2k_*}} - 2\frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}}.$$

Finally, define:

$$\begin{aligned}\widehat{w}_{n,M}(\alpha) \;&:=\; A\,\hat{\sigma}_{n,k}\,,\left(\sqrt{\frac{2\log(8/\alpha)}{n-2k_*}} + \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\right)\\ &\quad A\,\Delta_{n,k}\,\left(2\sqrt{2}\,\frac{\log(8/\alpha)}{n-2k_*} + \frac{\sqrt{2}M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\,\sqrt{\frac{2\log(8/\alpha)}{n-2k_*}}\right)\end{aligned}$$

and set

$$\widehat{I}_{n,M}(\alpha) := [\overline{X}_{n,k} - \widehat{w}_{n,M}(\alpha), \overline{X}_{n,k} + \widehat{w}_{n,M}(\alpha)]$$

Then there exists $C = C(\phi_*) > 0$ depending only on ϕ_* and M such that:

$$\mathbb{P}\left[\mu \in \widehat{I}_{n,M}(\alpha) \text{ and } \widehat{w}_{n,M}(\alpha) \leq C(\phi_*)\,\sigma\sqrt{\frac{\log(8/\alpha)}{n}}\right] \geq 1-\alpha.$$

Thus the Theorem says that the interval $\widehat{I}_{n,M}(\alpha)$ contains the mean with confidence $1-\alpha$, and additionally, that the length of the interval is typically within a sub-Gaussian range. One can actually show that, by making $M^{2p/(p-2)}\phi_*$ small, the constant $C(\phi_*)$ can be taken arbitrarily close to $\sqrt{2}$.

Proof. We claim that the following inequalities hold simultaneously with probability $\geq 1-\alpha$:

$$\begin{aligned}
|\overline{X}_{n,k_*} - \mu(U_{(k_*)}, U_{(n-k_*+1)})| &\leq \sigma(U_{(k_*)}, U_{(n-k_*+1)}) \sqrt{\frac{2\log(8/\alpha)}{n-2k_*}} \qquad (2.10)\\
&\quad + \Delta_{n,k_*} \frac{\log(8/\alpha)}{3(n-2k_*)};\\
\sigma(U_{(k_*)}, U_{(n-k_*)}) &\leq |\overline{X}_{n,k_*} - \mu(U_{(k_*)}, U_{(n-k_*+1)})| + \hat{\sigma}_{n,k_*} \qquad (2.11)\\
&\quad + 2\Delta_{n,k_*} \sqrt{\frac{\log(8/\alpha)}{(n-2k_*)}};\\
\hat{\sigma}_{n,k_*} &\leq \sigma(U_{(k_*)}, U_{(n-k_*+1)}) + \Delta_{n,k_*} \sqrt{\frac{\log(8/\alpha)}{2(n-2k_*)}}; \qquad (2.12)\\
\Delta_{n,k_*} &\leq 2M\,\sigma\, \frac{e^{\frac{1}{p}}}{3\,(1-2k_*/n)} \left(\frac{en}{k_*}\right)^{\frac{1}{p}}; \qquad (2.13)\\
B := 1 - (U_{(n-k_*+1)} - U_{(k_*)}) &\leq \phi_*. \qquad (2.14)
\end{aligned}$$

To see this, note that (2.10) holds with probability $\geq 1-\alpha_1$ by Theorem 2.10, part 1. Inequalities (2.11) and (2.12) hold with probability $\geq 1-\alpha_2$ by part 2 of the same theorem. Inequality (2.13) is a consequence of inequality (2.7) in the proof of Theorem 2.15. Finally, the fact that inequality (2.14) holds with probability $1-\alpha_4$ follows from the same argument as (2.6) in the proof of Theorem 2.15.

For the remainder of the proof, we will show the following claim.

If inequalities (2.10) to (2.14) all hold simultaneously, then

$$\mu \in \widehat{I}_{n,M}(\alpha) \text{ and } \widehat{w}_{n,M}(\alpha) \leq C(\phi_*)\,\sigma\sqrt{\frac{\log(8/\alpha)}{n}}.$$

So from now on, we assume the five inequalities hold. This gives us:

$$\sigma \leq \frac{(1-\phi_*)}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\,\sigma(U_{(k_*)}, U_{(n-k_*+1)}), \qquad (2.15)$$

thanks to Lemma 2.13 (the parameter δ that Lemma is $\leq \phi_*$ due to (2.14)).

Next, we consider the bias term $|\mu(U_{(k_*)}, U_{(n-k_*+1)}) - \mu|$. By Lemma 2.11, and using $\nu_p \leq M\sigma$, we obtain:

$$|\mu(U_{(k_*)}, U_{(n-k_*+1)}) - \mu| \leq \nu_p \frac{\phi_*^{\frac{p-1}{p}}}{1-\phi_*} \leq \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\,\sigma(U_{(k_*)}, U_{(n-k_*+1)}). \qquad (2.16)$$

Plugging this into (2.11) gives:

$$\sigma(U_{(k_*)}, U_{(n-k_*+1)}) \leq \frac{\hat{\sigma}_{n,k} + |\overline{X}_{n,k} - \mu| + 2\Delta_{n,k_*}\sqrt{\frac{\log(2/\alpha_2)}{(n-2k_*)}}}{1 - \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}}. \tag{2.17}$$

At the same time, plugging (2.16) into (2.10) and then applying (2.11) gives:

$$\begin{aligned}
|\overline{X}_{n,k_*} - \mu| &\leq \sigma(U_{(k_*)}, U_{(n-k_*+1)}) \left(\sqrt{\frac{2\log(2/\alpha_1)}{n-2k_*}} + \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\right) \\
&\quad + \Delta_{n,k_*}\frac{\log(2/\alpha_1)}{3(n-2k_*)} \\
&\leq \frac{\hat{\sigma}_{n,k} + |\overline{X}_{n,k} - \mu|}{1 - \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}} \left(\sqrt{\frac{2\log(2/\alpha_1)}{n-2k_*}} + \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\right) \\
&\quad + \frac{\Delta_{n,k}}{1 - \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}} \left(2\sqrt{2}\frac{\log(2/\alpha_1)}{n-2k_*} + \frac{\sqrt{2}M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}\sqrt{\frac{2\log(2/\alpha_1)}{n-2k_*}}\right).
\end{aligned}$$

We may now multiply both sides by

$$1 - \frac{M\phi_*^{\frac{p-1}{p}}}{\sqrt{1-4\phi_*^{1-2/p}M^2}}$$

and then collect the terms containing $|\overline{X}_{n,k} - \mu|$ to obtain $|\overline{X}_{n,k} - \mu| \leq \widehat{w}_{n,M}(\alpha)$.

The bound in $\widehat{w}_{n,M}(\alpha)$ follows from bounding the trimmed empirical variance via (2.12), using (2.13) to bound $\Delta_{n,k}$; appying $\sigma(U_{(k_*)}, U_{(n-k_*+1)}) \leq \sigma/(1-\phi_*)$ obtained from Lemma 2.13 and (2.14); and finally, performing some simple calculations using the estimate $\phi_* = O(\log(8/\alpha)/n)$. □

2.7 Adaptive trimming

Our final theoretical contribution is to give a general method for choosing the trimming parameter k in a data-driven fashion. Under the assumptions of Theorem 2.20, this would be possible to do via the method of sub-Gaussian confidence intervals from [Devroye et al., 2016, Section 4]. For symmetrical data, one could use similar ideas. The goal of this section a theorem that works under much more general assumptions.

Assumption 2.21. Besides Assumption 2.1, a confidence parameter $1-\alpha$, a (nonrandom) set of pairs of integers

$$\{(k_1(i), k_2(i)) \,:\, i = 1, 2, 3, \ldots, m\},$$

is fixed. We assume the following properties are satisfied for each index i:

1. $k_1(i) + k_2(i) < \eta\, n$ and
$$\frac{2\log(3m/\alpha)}{n - k_1(i) - k_2(i)} \leq \eta$$
for some fixed $\eta \in (0, 1)$;

2. $k_2(i) > 0$, and also $k_1(i) > 0$ if P is not supported on $[0, +\infty)$;

3. if $i < m$, then
$$k_1(i+1) \leq k_1(i) \text{ and } k_2(i+1) \leq k_2(i).$$

The numbers $k_1(i), k_2(i)$ correspond to different choices of trimming parameters. Condition 3 above means that, the larger i is, the less points will be trimmed.

Now define, for $i = 1, 2, \ldots, m$ and a choice of $\mu \in \mathbb{R}$, the following "bias" and "variance" parameters (recall the definitions and notation from subsection 2.4.1).

$$b(i, \mu) \;:=\; |\mu(U_{k_1(i)}, U_{(n-k_2(i)+1)}) - \mu|; \tag{2.18}$$

$$v(i) \;:=\; \frac{\hat{\sigma}_{n,k_1(i),k_2(i)} \sqrt{\frac{2\log(3\rho(i)/\alpha)}{n-k_1(i)-k_2(i)}} + c_0\, \Delta_{n,k_1(i),k_2(i)}\, \frac{\log(3\rho(i)/\alpha)}{(n-k_1-k_2)}}{1 - \sqrt{\frac{2\log(3\rho(i)/\alpha)}{n-k_1(i)-k_2(i)}}}, \tag{2.19}$$

where c_0 is the constant in Theorem 2.10.

Because larger i means less trimming, we expect that, if μ is the mean of the data distribution P, $b(i, \mu)$ should decrease with i. This is not always true, but we can nevertheless prove the next theorem. Notice that this result holds even when P does not have a well-defined mean.

Theorem 2.22. *Make Assumptions 2.1 and 2.21. Take $\mu \in \mathbb{R}$ and define $b(i, \mu)$ and $v(i)$ as in (2.18) and (2.19) (respectively). Fix $c > 1$. Then the following holds with probability $\geq 1 - \alpha$: if*

$$\widehat{I} := \min\left\{i \in [m] \,:\, \forall j, \ell \in \{i, i+1, \ldots, m\},\, |\overline{X}_{n,k_1(j),k_2(j)} - \overline{X}_{n,k_1(\ell),k_2(\ell)}| \leq c\,(v(j) + v(\ell))\right\},$$

then

$$|\overline{X}_{n,k_1(\widehat{I}),k_2(\widehat{I})} - \mu| \leq C_{c,\eta} \inf_{i \in [m]} \left(\max_{j=i,i+1,\ldots,m} b(j, \mu) + v(i) \right), \tag{2.20}$$

where $C_{c,\eta}$ depends only on c and η.

Notice that both terms in the RHS of (2.20) are sample dependent. This is also a feature of the main result of [Orenstein, 2018], which has inspired this theorem.

A somewhat simpler version of Theorem 2.22 is available for nonnegative distributions with a mean, a setting that comes up quite often in applications.

Corollary 2.23. *Add to the assumptions of Theorem 2.22 the following conditions:*

- *P is supported over $[0, +\infty)$;*
- *$\mu < +\infty$ is the mean of P;*
- *$k_1(i) = 0$ for each index i (i.e. we only trim the top of the sample).*

Then the conclusion of Theorem 2.22 can be strengthened as follows: with probability $\geq 1 - \alpha$,

$$|\overline{X}_{n,k_1(\widehat{I}),k_2(\widehat{I})} - \mu| \leq C_{c,\eta} \inf_{i \in [m]} (\mu - \mu(0, U_{(n-k_2(i)+1)}) + v(i)).$$

This result is immediate from Theorem 2.22, as in this case $b(i, \mu) = \mu - \mu(0, U_{(n-k_2(i)+1)})$ decreases with i. Let us now prove the Theorem.

Proof of Theorem 2.22. The conditions in Assumption 2.21 can be combined with Theorem 2.10 to obtain that:

$$\forall i = 1, 2, \ldots, m \, : \, \mathbb{P}\left[|\overline{X}_{n,k_1(i),k_2(i)} - \mu| \leq b(i, \mu) + v(i)\right] \geq 1 - \rho(i)\,\alpha.$$

Since the sum of the $\rho(i)$ is at most 1, we obtain that the following holds with probability $\leq 1 - \alpha$:

$$\forall i = 1, 2, \ldots, m \, : \, |\overline{X}_{n,k_1(i),k_2(i)} - \mu| \leq b(i, \mu) + v(i). \tag{2.21}$$

For the remainder of the proof, we *assume* that (2.21) holds, and obtain (2.20). To do this, we perform a simple adaptation of Lepskii's method as employed in [Orenstein, 2018] (see also [Lepskii, 1991, Mathé, 2006]). The slight difficulty here is that, unlike in the papers just mentioned, the bias term does not necessarily decrease with i, nor does the variance term increase with i.

To circumvent this, first define:

$$\tilde{v}(i) := \left(\sqrt{\frac{1}{2n^2} \sum_{i,j=k_1(i)+1}^{n-k_2(i)} (X_{(i)} - X_{(j)})^2} \right) \sqrt{\frac{2\log(3\rho(i)/\alpha)}{n}} + c_0\, \Delta_{n,k_1(i),k_2(i)}\, \frac{\log(3\rho(i)/\alpha)}{n}.$$

The quantity $\tilde{v}(i)$ can be obtained from $v(i)$ as follows: first we modify

$$\hat{\sigma}^2_{n,k_1(i),k_2(i)} = \frac{1}{2(n-k_1-k_2)^2} \sum_{i,j=k_1(i)+1}^{n-k_2(i)} (X_{(i)} - X_{(j)})^2$$

by replacing $n-k_1-k_2$ with n; we also replace other occurrences of $n-k_1-k_2$ in $v(i)$ with n; and finally, we remove the denominator in the RHS of (2.19). Part 1 of Assumption 2.21 implies that this has a bounded effect, meaning that there exists constants $a_\eta, b_\eta > 0$ depending only on η such that:

$$a_\eta \leq \frac{v(i)}{\tilde{v}(i)} \leq b_\eta.$$

Moreover, $\tilde{v}(i)$ clearly increases with i. In particular, we obtain that

$$\forall i, j = 1, \ldots, m \,:\, j \leq i \Rightarrow v(j) \leq \frac{b_\eta}{a_\eta}\, v(i). \tag{2.22}$$

Let us now perform a Lepskii-style analysis. Fixing $i \in \{1, \ldots, m\}$, we wish to argue that:

$$\textbf{Goal: } |\overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \mu| \leq C_{\eta,c} \left(\max_{j=i,i+1,\ldots,m} b(j,\mu) + v(i) \right).$$

To show this, consider first the case where $\hat{I} \leq i$. In this case we can apply the definition of $\hat{I}$ and inequality (2.22) with $j = \hat{I}$ to obtain:

$$|\overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \overline{X}_{n,k_1(i),k_2(i)}| \leq c\,(v(\hat{I}) + v(i)) \leq c\,\left(1 + \frac{b_\eta}{a_\eta}\right)\, v(i).$$

This gives:

$$|\overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \mu| \leq |\overline{X}_{n,k_1(i),k_2(i)} - \mu| + c\,\left(1 + \frac{b_\eta}{a_\eta}\right)\, v(i)$$

and (2.21) gives:

$$|\hat{I} \leq i \Rightarrow \overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \mu| \leq b(i,\mu) + \left(1 + c + \frac{c\, b_\eta}{a_\eta}\right)\, v(i). \tag{2.23}$$

The second case is when $i < \hat{I}$. Notice that (2.21) gives:

$$|\overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \mu| \leq b(\hat{I},\mu) + v(\hat{I}) \leq \max_{j=i,i+1,\ldots,m} b(j,\mu) + v(\hat{I}). \tag{2.24}$$

We now must bound the bias term.

In this case, the definition of $\hat{I}$ implies that there must exist some $j, \ell \in \{\hat{I}-1, \ldots, m\}$ with:

$$|\overline{X}_{n,k_1(j),k_2(j)} - \overline{X}_{n,k_1(\ell),k_2(\ell)}| > c\,(v(j) + v(\ell)). \tag{2.25}$$

On the other hand, applying (2.21) to indices k, ℓ gives:

$$|\overline{X}_{n,k_1(j),k_2(j)} - \overline{X}_{n,k_1(\ell),k_2(\ell)}| \leq b(j,\mu) + b(\ell,\mu) + v(j) + v(\ell).$$

Using that $j, \ell \geq \hat{I} - 1 \geq i$ (since $\hat{I} > i$), we obtain:

$$v(j) + v(\ell) \leq \frac{b(j,\mu) + b(\ell,\mu)}{c-1} \leq \frac{2}{c-1} \max_{j=i,i+1,\ldots m} b(j,\mu).$$

In order for any of the above to happen, we we must have $j \neq \ell$. Without loss of generality, assume $j > \ell$. Since $j, \ell \geq \hat{I} - 1$, we must have $\hat{I} \leq j$. Now (2.22) gives:

$$v(\hat{I}) \leq \frac{b_\eta}{a_\eta} v(j) \leq \frac{2b_\eta}{(c-1)a_\eta} \max_{j=i,i+1,\ldots m} b(j,\mu).$$

Plugging this back into (2.24) gives us:

$$\hat{I} > i \Rightarrow |\overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \mu| \leq \left(1 + \frac{2b_\eta}{(c-1)a_\eta}\right) \max_{j=i,i+1,\ldots m} b(j,\mu),$$

which, when combined with (2.23), gives:

$$|\overline{X}_{n,k_1(\hat{I}),k_2(\hat{I})} - \mu| \leq C_{c,\eta} \left(b(i,\mu) + v(i)\right)$$

for

$$C_{c,\eta} = \max\left\{\left(1 + c + \frac{c\, b_\eta}{a_\eta}\right), \left(1 + \frac{2b_\eta}{(c-1)a_\eta}\right)\right\}.$$

□

Chapter 3

Covariance results

3.1 Introduction

The problem of covariance matrix estimation is a very classical problem in Multivariate Statistics.In this paper, we are interested in estimating the covariance matrix $\Sigma = \mathbb{E}\left[XX^\top\right]$ of a random vector $X \in \mathbb{R}^d$ with zero mean from independent and identically distributed (i.i.d.) copies $X_1, \ldots, X_n$ of X. We study this problem under (relatively) heavy tails, from a nonasymptotic perspective. This less classical setting has received much recent attention [Mendelson and Paouris, 2014, Tikhomirov, 2017, Minsker, 2018].
A natural way to address the problem is using the sample covariance matrix:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i)(X_i)^\top.$$

In this line, there are important results on obtaining concentration inequalities for the operator norm of the deviation of the sample covariance operator from the true covariance operator. This line of research includes classical asymptotical work and more recent nonasymptotic bounds such as [Mendelson and Paouris, 2014, Tikhomirov, 2017, Vershynin, 2011] where the main issue is to understand the dependence on the dimension of the problem. In other papers, [Lounici, 2012, Koltchinskii and Lounici, 2014] dimesion-free results are obtained. In [Lounici, 2012] the author obtains bounds on the operator norm in terms of the "effective rank of the covariance matrix", which is defined as:

$$r(\Sigma) = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}}.$$

In [Koltchinskii and Lounici, 2014], the authors use an approach based on chaining bounds for empirical processes. They obtain concentration inequalities and expectation bounds for the operator norm for centered Gaussian random variables in a separable Banach space that do not depend on the dimension of the space. A recent preprint by Zhivotovskiy [Zhivotovskiy, 2021] obtains related dimension-free bounds under light-tail assumptions on the vector X.
Another line of analysis for the original problem is the design of an estimator of Σ that admits tight deviation bounds in the operator norm under minimal assumptions on the distribution of X. Important results in this line of research include [Minsker, 2018, Minsker and Wei, 2018, Mendelson and Zhivotovskiy, 2019, Ostrovskii and Rudi, 2019]. The current state-of-the-art regarding probability bounds in the L_4-L_2 norm equivalence case (bounded kurtosis assumption) in [Mendelson and Zhivotovskiy, 2019]. These authors find bounds that do not depend on the dimension of the space, but rather on the effective rank of Σ. Mendelson and Zhivotovskiy shows that there is an estimator for the covariance which, up to logarithmic factors, has a rate of error that is information theoretically optimal, under only the assumption of bounded kurtosis. To state their result, we first make an assumption.

Assumption 3.1 (i.i.d. data with uniformly bounded kurtosis of 1d marginals)**.** Let $X_1, \ldots, X_n$ be independent and identically distributed random column vectors in $\mathbb{R}^d$ with $\mathbb{E}\left[\|X_1\|^2\right] < +\infty$. We also assume that the mean $\mathbb{E}\left[X_1\right] = 0$ and the covariance matrix is $\Sigma \equiv \mathbb{E}\left[(X_1)(X_1)^\top\right]$. Finally, we assume $\kappa < +\infty$ is such that

$$\forall v \in \mathbb{R}^d \,:\, \mathbb{E}\left[\langle X_i, v\rangle^4\right] \leq \kappa \langle v, \Sigma v\rangle^2.$$

The main result of [Mendelson and Zhivotovskiy, 2019] is the following Theorem.

Theorem 3.2 (Theorem 1.12.(2) [Mendelson and Zhivotovskiy, 2019])**.** *Under Assumption 3.1, and for a fixed confidence parameter $1-\alpha \in (0,1)$, there is an estimator $\widehat{\Sigma}$ of Σ such that, if $n \geq c(\kappa)(r(\Sigma)\log(r(\Sigma)) + \log(1/\alpha))$, then*

$$\left\|\Sigma - \widehat{\Sigma}\right\|_{\text{op}} \leq c(\kappa)\left\|\Sigma\right\|_{\text{op}} \left(\sqrt{\frac{r(\Sigma)\log(r(\Sigma))}{n}} + \sqrt{\frac{\log(1/\alpha)}{n}}\right)$$

with probability at least $1-\alpha$. Here, $c(\kappa)$ is a constant that depends only on κ.

The present work present an estimator with better performance bounds.

Theorem 3.3 (Main result)**.** *Under Assumption 3.1, and for a fixed confidence parameter* $1-\alpha \in (0,1)$*, there is an estimator* $\widehat{\Sigma}$ *of* Σ *such that, if* $n \geq C(\kappa)(r(\Sigma) + \log(1/\alpha))$*, then*

$$\left\|\Sigma - \widehat{\Sigma}\right\|_{\mathrm{op}} \leq C(\kappa) \left\|\Sigma\right\|_{\mathrm{op}} \left(\sqrt{\frac{r(\Sigma)}{n}} + \sqrt{\frac{\log(1/\alpha)}{n}}\right)$$

with probability at least $1-\alpha$*. Here* $C(\kappa)$ *is a constant that depends only on* κ*.*

See Theorem 3.24 for a detailed formal statement. Our result is similar to [Mendelson and Zhivotovskiy, 2019] in that both estimators are not computationally efficient. The error bounds are also similar, with Theorem 3.3 removing some logarithmic factors in $r(\Sigma)$. This is relevant because, as explained in the discussion following Theorem 1.12 in [Mendelson and Zhivotovskiy, 2019], the estimator in Theorem 3.3 matches the behavior of the sample covariance would in the Gaussian setting. This is in spite of the fact that Assumption 3.1 is much weaker than Gaussianity, and allows for fairly heavy tails.

The improvement we have obtained over [Mendelson and Zhivotovskiy, 2019] requires a different proof strategy. The estimator in [Mendelson and Zhivotovskiy, 2019] is based on the median-of-means construction employed for robust vector mean estimation [Lugosi and Mendelson, 2019b]. At a key step, the analysis requires a dimension-free general matrix concentration inequality, due to Minsker [Minsker, 2017] and improved by Tropp [Tropp, 2015]. This is the step where the $\log r(\Sigma)$ factor appears. As is well known, general matrix concentration inequalities suffer from this "logarithmic drawback," and no approach requiring such inequalities could give us Theorem 3.3. On the other hand, the fact that we deal specifically with covariance-type matrices suggests that better bounds may be possible. In what follows, we give an overview of the ideas we introduce to obtain improved bounds.

3.1.1 Main proof ideas.

Our strategy for proving Theorem 3.3 is detailed in Section 3.3. Here we just discuss the main ingredients that come up later in the paper.

First of all, we use entropic (or PAC-Bayesian) inequalities in the form that has been popularized by Catoni and collaborators [Audibert and Catoni, 2011, Catoni, 2016, Catoni and Giulini, 2017]; see also [Zhivotovskiy, 2021] and [Oliveira, 2016] (more will be said about these papers later). Roughly speaking, this inequality shows that certain empirical processes automatically satisfy good concentration properties once they are smoothed with Gaussian noise. In our case, we prove a PAC-Bayesian version of Bernstein's concentration inequality (Proposition 3.6) that might be of independent interest. In the proof, we use this

inequality to control the truncated empirical process:

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, v\rangle^2 \wedge B \,:\, v \in \mathbb{S}^{d-1},$$

and also the counting functions

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\langle X_i,v\rangle^2>B} \,:\, v \in \mathbb{S}^{d-1},$$

for suitable $B > 0$. As a byproduct of this analysis, we obtain the following result.

Lemma 3.4 (Follows from Corollary 3.8 and Lemma 3.10 below)**.** *There exists an absolute constant $C > 0$ such that the following holds. make Assumption 3.1. Let $t \geq 1$ with $n > n - \lceil r(\Sigma) + t \rceil$, and set:*

$$B := C\sqrt{\kappa}\sqrt{\frac{n}{t}}\,\|\Sigma\|_{\mathrm{op}}\left(1 + \frac{r(\Sigma)}{t}\right).$$

Then with probability $\geq 1 - C\,e^{-t/C}$:

$$\forall v \in \mathbb{S}^{d-1} \,:\, \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\langle X_i,v\rangle^2>B} \leq t \text{ and}$$

$$\forall v \in \mathbb{S}^{d-1} \,:\, \left|\frac{1}{n}\sum_{i\in[n]}\langle X_i, v\rangle^2 \wedge B - \langle v, \Sigma v\rangle\right| \leq C\sqrt{\kappa}\,\|\Sigma\|_{\mathrm{op}}\sqrt{\frac{r(\Sigma)+t}{n}}.$$

Similar lemmas were proven as a first step in the analysis of the sample covariance matrices by [Adamczak et al., 2010, Mendelson and Paouris, 2014, Tikhomirov, 2017], albeit with $r(\Sigma)$ replaced by the ambient dimension. The role of these lemmas in the previous papers was to say that control of the sample covariance can achieved via a combination of these bounds for the "small" values of $\langle X_i, v\rangle^2$ (covered by the analogues of our Lemma) and some other strategy for the large values. We believe that our lemma can be used to make those analyses dimension-independent.

The second ingredient in our proofs is working with *optimally weighted samples*. Specifically, we will consider estimators of the covariance that take the form

$$\widehat{\Sigma}(\widehat{\lambda}) := \sum_{i=1}^{n}\widehat{\lambda}_i\, X_i X_i^{\top}$$

where $\widehat{\lambda} = (\widehat{\lambda}_i)_{i=1}^n$ is a vector of convex weights chosen from the sample. Our estimator is based on choosing $\widehat{\lambda}$ that solves a certain convex optimization problem. This is inspired

by the strategy used in the Computer Science literature on adversarially robust estimation [Diakonikolas and Kane, 2019], especially in reference [Hopkins et al., 2021]. While we do not obtain an efficient algorithm, we believe that it should be possible to adapt our techniques to do so.

We emphasize that the explanation we have just presented is only a very rough outline of our proof. In particular, the use of convexity and minimax-style arguments will require that we move to the convex hull of the set of matrices of the form $vv^\top$. Section 3.3 presents an overview of the whole argument.

3.1.2 Further background.

We now add a few pointers to related work that was not discussed in detail above.

The present paper belongs to a line of research that consists of estimating means and covariances of distributions with best-possible nonasymptotic performance. Although the sample mean is the optimal asymptotic estimator in one dimension, Catoni's seminal paper [Catoni, 2012] showed that it can be greatly improved in finite-sample settings with known variance. More specifically, that paper shows that Chebyshev's inequality is the tight deviation bound for the sample mean (up to constants), but there are other estimators achieving Gaussian-like behavior. So-called "sub-Gaussian mean estimators" in one dimension were further studied in [Devroye et al., 2016].

The literature soon moved to the estimation of means of vectors. Minsker [Minsker, 2015] provided a a general "geometric median" estimator for random vectors in a Banach space, with good (but suboptimal) finite-sample properties. After preliminary results by Joly et al. [Joly et al., 2017], Lugosi and Mendelson were the first to obtain a sub-Gaussian estimator for vectors in $\mathbb{R}^d$ with the Euclidean norm [Lugosi and Mendelson, 2019b]. Further results in this are include refinements of the Euclidean estimator [Lugosi and Mendelson, 2021]; computationally efficient algorithms, implementing the original Lugosi-Mendelson construction e.g. [Hopkins, 2020, Depersin and Lecué, 2022]; and nearly optimal estimators for general norms [Depersin and Lecué, 2021, Depersin and Lecué, 2021].

Estimating means of matrices (including covariance matrices) from a random sample is a particular case of mean estimation under general norms. However, the best results in that problem seem to come from approaches that are specific to matrices. The important works of Catoni and Giulini [Catoni, 2016] and [Catoni and Giulini, 2017] use PAC-Bayesian methods estimate vectors and covariance matrices; their bounds are dimension-free, but they are not

centered, and do not quite reproduce the sub-Gaussian behavior of other works. Minsker's paper [Minsker, 2018] works for general matrices, but loses logarithmic factors. The aforementioned [Mendelson and Zhivotovskiy, 2019, Minsker, 2018] also deal with matrix estimation in this sub-Gaussian sense.

The construction of our estimator is related to the weighting method from the Computer Science literature on robust estimators [Diakonikolas and Kane, 2019]. The goal in that are is to give computationally efficient estimators that can deal with arbitrary changes (often called "adversarial corruption") of a small fraction of sample points. Of the main references in the area, we cite the seminal [Diakonikolas et al., 2019], the survey [Diakonikolas and Kane, 2019], and the more recent paper by Hopkins et al. [Hopkins et al., 2021] which emphasizes the use of weights on samples. We observe that, in spite of the connection with these methods, our estimator is not computationally efficient.

3.1.3 Organization.

The remainder of the chapter is organized as follows: in the next section, we provide some preliminaries and introduce a general PAC-Bayesian method for empirical processes. We then present in section 3.3 an overview of the of the ideas used to achieve our result. A key result for counting vectors is presented in section 3.4. The analysis of the empirical process for vectors is presented in section 3.5. The analysis of our passage from vectors to matrices is in section 3.6. The final section ends by showing the final estimator. The proofs of the technical lemmas are derived in the Appendix.

3.2 Some preliminaries

3.2.1 PAC-Bayesian Bernstein inequality.

The purpose in this section is to introduce methods based on entropy inequalities to work with truncated empirical process such as the ones we encounter in our analysis. For a general result, we start with the following assumption.

Assumption 3.5. Consider a family of i.i.d. random variables $\{Z_i(\theta)_{i\in[n],\theta\in\mathbb{R}^d}\}$ defined over a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. The map
$$(\omega, \theta) \in \Omega \times \mathbb{R}^d \mapsto Z_i(\theta)(\omega) \in \mathbb{R}$$
is $(\mathcal{F} \otimes \mathcal{B}(\mathbb{R}^d))/\mathcal{B}(\mathbb{R})$-measurable.

2. Given $v \in \mathbb{R}^d$ and $\gamma > 0$, let us denote by $\Gamma_{v,\gamma}$ the Gaussian product measure over $\mathbb{R}^d$ with mean v and covariance matrix $\gamma\, I_{\mathrm{d}\times d}$. We assume

$$(\Gamma_{v,\gamma} Z_i(\theta))(\omega) = \int_{\mathbb{R}^d} Z_i(\theta)(\omega)\Gamma_{v,\gamma}(d\theta)$$

is well defined for all $\omega \in \Omega$ and depends continuously on v. We also assume that the following quantities are well defined:

$$\begin{aligned} \bar{\mu}_\gamma &:= \sup_{v\in\mathbb{S}^{d-1}} \Gamma_{v,\gamma}\mathbb{E}\left[Z_1(\theta)\right], \\ \bar{\sigma}^2_\gamma &:= \sup_{v\in\mathbb{S}^{d-1}} \Gamma_{v,\gamma}\mathbb{V}\left[Z_1(\theta)\right]. \end{aligned}$$

3. For each $\theta \in \mathbb{R}^2$, $\{Z_i(\theta)_{i\in\{1,\ldots,n\},\theta\in\mathbb{R}^d}\}$ are independent with second moment bounded, and $Z_i(\theta) - \mathbb{E}\left[Z_i(\theta)\right] \le A$ for some constant $A > 0$.

Recall the definition of the Kullback-Leiber divergence, between two probability measures: μ_0, and μ_1 on $\mathbb{R}^d$.

$$\mathrm{KL}(\mu_1|\mu_0) := \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{d\mu_1}{d\mu_0}(\theta)\right)\mu_1(d\theta) & \mu_1 \ll \mu_0 \\ +\infty & \text{otherwise} \end{cases}$$

A characterization given by the variational formula [Ledoux, 2001] implies that for all measurable and μ_1-integrable $h : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathbb{R}^d} h(\theta)\mu_1(\theta) \le \mathrm{KL}(\mu_1|\mu_0) + \log\left(\int_{\mathbb{R}^d} e^{h(\theta)}\mu_1(d\theta)\right).$$

In our case, choosing $\mu_0 = \Gamma_{0,\gamma}$ and $\mu_1 = \Gamma_{v,\gamma}$ for $v \in \mathbb{S}^{d-1}$. The KL divergence between the Gaussian measure for all $v \in \mathbb{S}^{d-1}$ is given by

$$\mathrm{KL}(\mu_1|\mu_0) = \frac{\gamma^2}{2}.$$

Therefore, using $h(\theta) = Z_\xi(\theta)$ the variational inequality gives:

$$\sup_{v\in\mathbb{S}^{d-1}} \left(\Gamma_{v,\gamma} h(\theta) - \frac{1}{2\gamma^2}\right) \le \log \Gamma_{o,\gamma} e^{Z_\xi(\theta)} \le +\infty. \tag{3.1}$$

The Assumptions 3.5 that we made on the $Z_i(\theta)$ imply that for each $\theta \in \mathbb{R}^d$ and $\alpha \in (0,1)$:

$$\mathbb{P}\left[\sum_{i=1}^n Z_i(\theta) - n\mathbb{E}\left[Z_i(\theta)\right] \ge \mathbb{V}\left[Z_i(\theta)\right]\sqrt{2n\log(1/\alpha)} + \frac{2A\log(1/\alpha)}{2}\right] \le \alpha.$$

Our next result is a Bernstein-type inequality for the supremum of a Gaussian smoothing process: supremum of $\Gamma_{v,\gamma}\sum_{i=1}^n Z_i(\theta)$ over $v \in \mathbb{S}^{d-1}$.

Proposition 3.6 (Bernstein-type concentration inequality for Gaussian smoothed process)**.** *Suppose Assumption 3.5 holds, an let be $\gamma > 0$. Then for all $\alpha \in (0,1)$,*

$$\sup_{v\in\mathbb{S}^{d-1}} \sum_{i=1}^{n} \Gamma_{v,\gamma}\left(Z_i(\theta)\right) \leq n\bar{\mu}_\gamma + \bar{\sigma}_\gamma\sqrt{n(\gamma^{-2}+2\,\log(1/\alpha))} + \frac{A\left(\gamma^{-2}+2\,\log(1/\alpha)\right)}{6}$$

with probability $\geq 1-\alpha$.

Proof. First, observe that $\sup_{v\in\mathbb{S}^{d-1}}\Gamma_{v,\gamma}(\sum_{i=1}^{n} Z_i(\theta)(\omega))$ is a measurable function of $\omega\in\Omega$ because of the continuity assumption. Now, let $\xi\in(0,3/A)$, $\theta\in\mathbb{R}^d$ and let us define

$$Z_\xi(\theta) := \sum_{i=1}^{n} \xi\left[Z_i(\theta) - \mathbb{E}\left[Z_i(\theta)\right] - \frac{n\,\xi\,\mathbb{V}\left[Z_i(\theta)\right]}{2\left(1-\xi A/3\right)}\right].$$

The variational inequality (3.1) implies

$$\frac{\sup_{v\in\mathbb{S}^{d-1}}\Gamma_{v,\gamma}Z_\xi(\theta)}{\xi} \leq \frac{\frac{\gamma^{-2}}{2}+\log\Gamma_{0,\gamma}e^{Z_\xi(\theta)}}{\xi}$$

Next, we claim that for $\xi\in(0,3/M)$:

$$\mathbb{P}\left[\log\Gamma_{0,\gamma}e^{Z_\xi(\theta)} \leq \log(1/\alpha)\right] \leq 1-\alpha.$$

Indeed, by the Markov's Inequality and Fubini, it follows

$$\mathbb{P}\left[\log\Gamma_{0,\gamma}e^{Z_\xi(\theta)} \leq \log(1/\alpha)\right] \geq 1-\alpha\,\Gamma_{0,\gamma}\mathbb{E}\left[e^{Z_\xi(\theta)}\right].$$

Furthermore, a computation with moment generating functions as in the proof of Bernstein's [Boucheron et al., 2013, 2.8] inequality gives

$$\forall\theta\in\mathbb{R}^d : \mathbb{E}\left[e^{Z_\xi(\theta)}\right] = \prod_{i=1}^{n}\left(\mathbb{E}\left[\exp\left\{\,\xi(Z_i(\theta)-\mathbb{E}\left[Z_i(\theta)\right]) - \frac{\xi^2\mathbb{V}\left[Z_i(\theta)\right]}{2-\frac{2\xi M}{3}}\right\}\right]\right) \leq 1.$$

We deduce from the above that:

$$\mathbb{P}\left[\frac{\sup_{v\in\mathbb{S}^{d-1}}\Gamma_{v,\gamma}Z_\xi(\theta)}{\xi} \leq \log(1/\alpha)\right] \geq 1-\alpha.$$

The definitions of $\bar{\mu}_\gamma$ and $\bar{\sigma}_\gamma$ in Assumption 3.5 imply:

$$\frac{\Gamma_{v,\gamma}\,Z_\xi(\theta)}{\xi} \geq \sum_{i=1}^{n}\Gamma_{v,\gamma}\,Z_i(\theta) - n\bar{\mu}_\gamma - \frac{n\,\xi\,\bar{\sigma}_\gamma^2}{2\left(1-\xi A/3\right)},$$

so we obtain:

$$\mathbb{P}\left[\sup_{v\in\mathbb{S}^{d-1}}\sum_{i=1}^{n}\Gamma_{v,\gamma}\left(Z_i(\theta)\right) \leq n\bar{\mu}_\gamma + \frac{n\xi\bar{\sigma}_\gamma^2}{2-\frac{2A\xi}{3}} + \frac{\frac{\gamma^{-2}}{2}+\log\left(1/\alpha\right)}{\xi}\right] \geq 1-\alpha.$$

This holds for any $\xi \in (0, 3/A)$. Choosing

$$\xi^* := \frac{\sqrt{\gamma^{-2} + 2\log(1/\alpha)}}{\sqrt{n}\bar{\sigma}_\gamma \left(1 + \frac{A\sqrt{\gamma^{-2}+2\log(1/\alpha)}}{3\sqrt{n}\bar{\sigma}_\gamma}\right)}$$

gives:

$$\frac{n\xi^*\ \bar{\sigma}_\gamma^2}{2 - \frac{2M\xi^*}{3}} + \frac{\frac{\gamma^{-2}}{2} + \log(1/\alpha)}{\xi^*} = \bar{\sigma}_\gamma\sqrt{\gamma^{-2} + 2\log(1/\alpha)} + \frac{A(\gamma^{-2} + 2\log(1/\alpha))}{6}.$$

Hence, the result holds. □

3.3 Proof elements and overview

In this section, we present a general overview of how our proofs are developed, and introduce some of our main tools.

As explained above, we follow previous work such as [Hopkins et al., 2021] and consider weighted covariance estimators. Define a set of weight vectors,

$$\Delta_{n,k} := \left\{\lambda \in \mathbb{R}_+^n \ : \ \sum_{i=1}^n \lambda_i = 1,\ \max_{i\le n} \lambda_i \le \frac{1}{n-k}\right\},$$

and write

$$\widehat{\Sigma}(\lambda) := \sum_{i=1}^n \lambda_i X_i X_i^\top.$$

Ultimately, our goal is to choose $\widehat{\lambda} \in \Delta_{n,k}$ (for some suitable k) so that $\|\widehat{\Sigma}(\widehat{\lambda}) - \Sigma\|_{\text{op}}$ is small. To shed some intuition on the weight set, notice that $\Delta_{n,k}$ is the convex hull of vectors of the form

$$\frac{\mathbf{1}_S}{n-k},\ \text{where } S \subset [n] \text{ has cardinality } n-k$$

and $\mathbf{1}_S$ has coordinates $\mathbf{1}_{S,i} = \mathbf{1}_{i\in S}$ (this result is Lemma A.2 in the appendix). So in a way, $\Delta_{n,k}$ "convexifies" the idea of choosing a subset of the vectors X_i in order to estimate the covariance (which would allow one to avoiding outliers).

3.3.1 Controlling the norm: a first step.

For each $\lambda \in \Delta_{n,k}$, the norm $\|\widehat{\Sigma}(\lambda) - \Sigma\|_{\text{op}}$ is given by:

$$\sup_{v\in\mathbb{S}^{d-1}} \left|\sum_{i=1}^n \lambda_i \langle X_i, v\rangle^2 - \langle v, \Sigma v\rangle\right|.$$

This means we must obtain both upper and lower bounds on the empirical process inside the absolute value. However, it is known that lower tails of sample covariances behave much more nicely than the upper tails [Oliveira, 2016]. As a first step, then, we focus on the upper tail. To bound the *largest* eigenvalue of $\widehat{\Sigma}(\lambda) - \Sigma$ for the *best possible* choice of λ, we consider the minimax problem:

$$\inf_{\lambda \in \Delta_{n,k}} \sup_{v \in \mathbb{S}^{d-1}} \sum_{i=1}^{n} \lambda_i \langle X_i, v \rangle^2 - \langle v, \Sigma v \rangle. \tag{3.2}$$

Let us pretend for a moment that some minimax theorem applies, and we can exchange the order of inf and sup in the previous display. Then we want to bound:

$$\sup_{v \in \mathbb{S}^{d-1}} \inf_{\lambda \in \Delta_{n,k}} \sum_{i=1}^{n} \lambda_i \langle X_i, v \rangle^2 - \langle v, \Sigma v \rangle. \tag{3.3}$$

Now the infimum inside is easy to compute. By Lemma A.2 in the appendix,

$$\inf_{\lambda \in \Delta_{n,k}} \sum_{i=1}^{n} \lambda_i \langle X_i, v \rangle^2 - \langle v, \Sigma v \rangle = \frac{1}{n-k} \sum_{i \in S(v)} \langle X_i, v \rangle^2 - \langle v, \Sigma v \rangle,$$

where $S(v)$ is the set of indices $i \in [n]$ achieving the $n-k$ smallest values of $\langle X_i, v \rangle^2$.

For our next step, assume B satisfies the following "counting condition."

$$\textbf{Counting condition: } \forall v \in \mathbb{S}^{d-1} \, : \, \#\{i \in [n] \, : \, \langle X_i, v \rangle^2 > B\} \leq k. \tag{3.4}$$

That implies that $\langle X_i, v \rangle^2 \leq B$ for all $i \in S(v)$. From this one may conclude that:

$$\frac{1}{n-k} \sum_{i \in S(v)} \langle X_i, v \rangle^2 - \langle v, \Sigma v \rangle \leq \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \wedge B - \mathbb{E}\left[\langle X_1, v \rangle^2 \wedge B \right].$$

In other words, we have obtained:

$$\sup_{v \in \mathbb{S}^{d-1}} \inf_{\lambda \in \Delta_{n,k}} \sum_{i=1}^{n} \lambda_i \langle X_i, v \rangle^2 - \langle v, \Sigma v \rangle \leq \sup_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \wedge B - \mathbb{E}\left[\langle X_1, v \rangle^2 \wedge B \right] \right|. \tag{3.5}$$

It turns out that the "truncated empirical process" in the RHS of (3.5) and the counting condition (3.4) can both be shown to hold in high probability (for suitable B) via PAC-Bayesian methods. This will require first applying Gaussian smoothing and then comparing the smoothed and unsmoothed processes. Additionally, we will also argue that the lower tail of $\widehat{\Sigma}(\lambda) - \Sigma$ is controlled by the truncated process for any possible choice of $\lambda \in \Delta_{n,k}$.

3.3.2 A minimax argument via matrices.

The above argument would be enough for our purposes if the passage from (3.2) to (3.3) was valid. That, however, is not the case: one would need concavity instead of convexity in v to apply a minimax argument. To work around this issue, we move to a matrix setting.

Given $v \in \mathbb{S}^{d-1}$, let $M(v) = vv^T$. The minimax problem in (3.2) can be rewritten as follows.

$$\inf_{\lambda \in \Delta_{n,k}} \sup_{v \in \mathbb{S}^{d-1}} \sum_{i=1}^{n} \lambda_i \langle X_i, M(v) X_i \rangle - \operatorname{tr}(\Sigma M(v)),$$

Notice that the function of λ and $M(v)$ appearing above is affine in λ and linear in $M(v)$. The only thing preventing a proper application of a minimax theorem is the fact that the set of matrices $\{M(v) : v \in \mathbb{S}^{d-1}\}$ is not convex. This can be fixed by passing to its convex hull, which is easily seen to be the set of "density matrices:"

$$\mathcal{D}(\mathbb{R}^d) := \{M \in \mathbb{R}^{d \times d} : M = M^T,\, M \geq 0 \text{ and } \operatorname{tr}(M) = 1\}.$$

This means that the minimax problem that matters to us is:

$$\inf_{\lambda \in \Delta_{n,k}} \sup_{M \in \mathcal{D}(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i \langle X_i, M X_i \rangle - \operatorname{tr}(\Sigma M). \tag{3.6}$$

To this problem we may safely apply Sion's theorem. Moreover, there is a randomized construction mapping M to unit vectors g_M that are "close" to M in some sense. With this, one can prove analogues of the counting condition (3.4) and the truncated empirical process bound (3.5) for problem (3.6).

3.3.3 The final estimator.

In the end, the above reasoning will allow us to prove the following result. For a matrix A, let $\sigma_{\max}(A)$ denote its largest eigenvalue.

$$\widehat{\lambda}_k \in \arg\min_{\lambda \in \Delta_{n,k}} \left(\max_{\lambda' \in \Delta_{n,k}} \sigma_{\max}(\widehat{\Sigma}(\lambda) - \widehat{\Sigma}(\lambda')) \right).$$

This estimator will be shown to satisfy an error bound of the type:

$$\|\widehat{\Sigma}(\widehat{\lambda}_k) - \Sigma\|_{\mathrm{op}} \leq \varepsilon(k)$$

with high probability. However, it will be clear from our bounds that, in order to obtain our main theorem we will need to take $k \approx r(\Sigma) + \log(1/\alpha)$. This is problematic because $r(\Sigma) = \operatorname{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$ is unknown.

Following [Mendelson and Zhivotovskiy, 2019], we use a three step procedure. In the first step, we estimate the trace of Σ via the median-of-means method. In the second step, we estimate the operator norm $\|\Sigma\|_{\mathrm{op}}$ via $\|\widehat{\Sigma}(\widehat{\lambda}_{cn})\|_{\mathrm{op}}$ for some small $c > 0$. Finally, we now have a good estimate of $r(\Sigma)$ and can use that to compute the final estimator, with the desired error rate.

3.4 Counting arguments for vectors

This section presents a probabilistic argument that will allow us to show that the counting condition in (3.4) holds with high probability. This is the content of the following lemma.

Lemma 3.7. *Under Assumption 3.1. Assume $t > 0$, and*

$$B = 4\sqrt{2\,\kappa}\sqrt{\frac{n}{t}}\,\|\Sigma\|_{\mathrm{op}}\left(1+\gamma^2\,r(\Sigma)\right),$$

Then

$$\mathbb{P}\left[\forall v \in \mathbb{S}^{d-1} : \sum_{i=1}^{n} \mathbf{1}_{\langle X_i, v\rangle^2 \geq B} \leq \frac{t}{2} + \sqrt{2\,t^2 + t\,\gamma^{-2}} + \frac{2\,t+\gamma^{-2}}{3}\right] \geq 1 - e^{-t}.$$

we conclude thanks to (3.4).

Once this lemma is in place, we can easily obtain the following corollary (take $\gamma^2 = 1/2t$).

Corollary 3.8 (Proof omitted)**.** *Make Assumption 3.1. Assume also $B > 0$ such that:*

$$B = 4\sqrt{2\,\kappa}\sqrt{\frac{n}{t}}\,\|\Sigma\|_{\mathrm{op}}\left(1+\frac{r(\Sigma)}{2\,t}\right),$$

Then

$$\mathbb{P}\left[\forall v \in \mathbb{S}^{d-1} : \sum_{i=1}^{n} \mathbf{1}_{\langle X_i, v\rangle^2 \geq B} \leq \frac{23}{6}\,t\right] \geq 1 - e^{-t}.$$

Proof of Lemma 3.7. First, observe that for all $v \in \mathbb{S}^{d-1}$

$$\sum_{i=1}^{n} \mathbf{1}_{\langle X_i, v\rangle^2 \geq B} \leq 2\sum_{i=1}^{n} \Gamma_{v,\gamma}\mathbf{1}_{\langle X_i, \theta\rangle^2 \geq B}. \tag{3.7}$$

Then, applying the choice of $Z_i(\theta) := \mathbf{1}_{\langle X_i,\theta\rangle^2 \geq B}$ in Proposition 3.6, it follows that $\forall v \in \mathbb{S}^{d-1}$

$$\sum_{i=1}^{n} \Gamma_{v,\gamma}\mathbf{1}_{\langle X_i, \theta\rangle^2 \geq B} \leq \sum_{i=1}^{n} \Gamma_{v,\gamma}\mathbb{E}\left[\mathbf{1}_{\langle X_i, \theta\rangle^2 \geq B}\right] + \left(\sum_{i=1}^{n} \Gamma_{v,\gamma}\mathbb{V}\left[\mathbf{1}_{\langle X_i, \theta\rangle^2 \geq B}\right](2\,t+\gamma^{-2})\right)^{\frac{1}{2}} + \frac{2\,t+\gamma^{-2}}{6},$$

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

Next, let us now compute both integrals: $\Gamma_{v,\gamma}\mathbb{E}\left[\mathbf{1}_{\langle X_i,\theta\rangle^2\geq B}\right]$ and $\Gamma_{v,\gamma}\mathbb{V}\left[\mathbf{1}_{\langle X_i,\theta\rangle^2\geq B}\right]$ for $v \in \mathbb{S}^{d-1}$. By Markov inequality,

$$\mathbb{E}\left[\mathbf{1}_{\langle X_i,v\rangle^2\geq B}\right] = \mathbb{P}\left[\langle X_i, v\rangle^2 \geq B\right] \leq \frac{\mathbb{E}\left[\langle X_i, v\rangle\right]^4}{B^2},$$

then by Fubini,

$$\Gamma_{v,\gamma}\mathbb{E}\left[\mathbf{1}_{\langle X_i,\theta\rangle^2\geq B}\right] \leq \frac{1}{B^2}\mathbb{E}\left[\Gamma_{v,\gamma}\langle X_i, \theta\rangle^4\right].$$

Observe that we have the same bound for $\Gamma_{v,\gamma}\mathbb{V}\left[\mathbf{1}_{\langle X_i,\theta\rangle\geq B}\right]$. Indeed

$$\mathbb{V}\left[\mathbf{1}_{\langle X_i,v\rangle^2\geq B}\right] = \mathbb{P}\left[\langle X_i, v\rangle^2 \geq B\right]\left(1 - \mathbb{P}\left[\langle X_i, v\rangle^2 \geq B\right]\right) \leq \frac{\mathbb{E}\left[\langle X_i, v\rangle\right]^4}{B^2}.$$

Finally, we compute $\mathbb{E}\left[\Gamma_{v,\gamma}\langle X_i, v\rangle^4\right]$. It follows from calculations with the normal distribution,

$$\begin{aligned}
\Gamma_{v,\gamma}\mathbb{E}\left[\left(\mathbf{1}_{\langle X_i,\theta\rangle^2\geq B}\right)^2\right] &\leq \frac{1}{B^2}\Gamma_{v,\gamma}\,\mathbb{E}\left[\langle X_i, \theta\rangle^4\right] \\
&\leq \frac{\kappa}{B^2}\Gamma_{v,\gamma}\langle \theta, \Sigma\theta\rangle^2 \\
&\leq \frac{\kappa}{B^2}\Gamma_{0,1}\|\Sigma^{1/2}(v + \gamma\theta)\|^4 \\
&\leq \frac{8\,\kappa}{B^2}\Gamma_{0,1}\left(\|\Sigma^{1/2}v\|^4 + \gamma^4\|\Sigma^{1/2}\theta\|^4\right) \\
&\leq \frac{8\,\kappa}{B^2}\left(\|\Sigma\|_{\mathrm{op}}^2 + \gamma^4\,\mathrm{tr}\,(\Sigma)^2\right), \qquad (3.8)
\end{aligned}$$

and by choosing

$$B = 4\sqrt{2\,\kappa}\sqrt{\frac{n}{t}}\,\|\Sigma\|_{\mathrm{op}}\left(1 + \gamma^2\,r(\Sigma)\right),$$

we conclude

$$\forall v \in \mathbb{S}^{d-1} : \sum_{i=1}^{n}\Gamma_{\gamma,c}\mathbf{1}_{\langle X_i,\theta\rangle^2\geq B} \leq \frac{t}{4} + \frac{\sqrt{t}}{2}\sqrt{2\,t + \gamma^{-2}} + \frac{2\,t + \gamma^{-2}}{6},$$

with probability at least $1 - e^{-t}$. We conclude thanks to (3.4). □

3.5 Truncated empirical processes for vectors

The main goal of this section is to analyze the truncated empirical process in the RHS of equation (3.5), which was discussed in subsection 3.3.1.

We start by defining both the truncated empirical processes and its Gaussian-smoothed version.

$$\begin{aligned}\varepsilon_\theta(B) &:= \sup_{\|v\|=1} \frac{1}{n}\left|\sum_{i=1}^n \left(\langle X_i, v\rangle^2 \wedge B - \mathbb{E}\left[\langle X_i, v\rangle^2 \wedge B\right]\right)\right| \text{ and} \\ \widetilde{\varepsilon}_\theta(B) &:= \sup_{\|v\|=1} \frac{1}{n}\left|\sum_{i=1}^n \Gamma_{v,\gamma}\left(\langle X_i, \theta\rangle^2 \wedge B - \mathbb{E}\left[\langle X_i, \theta\rangle^2 \wedge B\right]\right)\right|.\end{aligned}$$

The main results of this section are twofold. First, we prove a deterministic statement that bounds the *lower tail* of $\widehat{\Sigma}(\lambda) - \Sigma$ uniformly over $\lambda \in \Delta_{n,k}$ via the truncated empirical process, $\varepsilon_\theta(B)$.

Proposition 3.9 (Proof in sub. 3.5.1)**.** *Under Assumption 3.1. Set $B > 0$ is such that:*

$$\forall v \in \mathbb{S}^{d-1} \,:\, B_k(v) := \text{ } (n-k) \text{ smallest value of } \langle v, X_i\rangle^2 \leq B.$$

Then for all $\lambda \in \Delta_{n,k}$ and $v \in \mathbb{S}^{d-1}$,

$$\inf_{\lambda \in \Delta_{n,k}} \langle v, \widehat{\Sigma}(\lambda) v\rangle \;\geq\; \langle v, \Sigma\, v\rangle - \frac{\kappa\, \langle v, \Sigma\, v\rangle^2}{B} - \frac{B\, k}{n} - \varepsilon_\theta(B). \tag{3.9}$$

Second, we give a probabilistic bound on the truncated empirical process.

Lemma 3.10. *Make Assumption 3.1. Suppose $B > 0$ and $\gamma > 0$. Then with probability at least $1 - \alpha - e^{-k}/6$*

$$\begin{aligned}\varepsilon_\theta(B) \;\leq\; & \gamma^2 \operatorname{tr}(\Sigma) \sqrt{\frac{2\,\kappa \log(2/\alpha)}{n}} + \frac{2\,B\, \log(2/\alpha)}{3n} + \frac{B\,k}{n}\, c \\ +\; & \sqrt{\frac{8\,\kappa}{n}} \left(\|\Sigma\|_{\mathrm{op}} + \gamma^2 \operatorname{tr}(\Sigma)\right) \sqrt{2\log(2/\alpha) + \gamma^{-2}} + \frac{(2\log(2/\alpha) + \gamma^{-2})\, B}{6\, n},\end{aligned}$$

where

$$c = \sum_{j=1}^{+\infty} \frac{32\sqrt{2}}{j^{3/2}} \exp\left(\frac{-j}{2}\right) + 8.$$

We note that the proof of this Lemma requires several steps. In subsection 3.5.2 we use a PAC-Bayesian argument to control $\widetilde{\varepsilon}_\theta(B)$. In subsection 3.5.3 we develop tools to bound the difference $|\varepsilon_\theta(B) - \widetilde{\varepsilon}_\theta(B)|$. Finally, we put the previous bounds together and obtain Lemma 3.10.

3.5.1 Control of the lower tail.

We prove here Proposition 3.9 following.

Proof. As observed at the beginning of section 3.3, $\lambda \in \Delta_{k,n}$ is a convex combination of indicator vectors of the form: $\mathbf{1}_S/(n-k)$, with $S \in \binom{[n]}{n-k}$.

Therefore, for a fixed $v \in \mathbb{S}^{d-1}$, observe that

$$\begin{aligned}
\inf_{\lambda \in \Delta_{n,k}} \langle v, \widehat{\Sigma}(\lambda) v \rangle &= \text{mean of the smallest } n-k \text{ values of } \langle v, X_i \rangle^2 \\
&\geq \frac{1}{n}(\text{ sum of the smallest } n-k \text{ values of } \langle X_i, v \rangle^2 \wedge B \ ) \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 - \frac{B\,k}{n}, \qquad (3.10)
\end{aligned}$$

for all $B \geq B_k(v)$.

Note that

$$\frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \wedge B \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\langle X_i, v \rangle^2 \wedge B \right] - \varepsilon_\theta(B).$$

To finish, we find a lower bound for the first term in the RHS.

$$\begin{aligned}
\mathbb{E}\left[\langle X_i, v \rangle^2 \wedge B \right] &\geq \mathbb{E}\left[\langle X_i, v \rangle^2 \right] \mathbf{1}_{\langle X_i, v \rangle^2 \leq B} \\
&= \mathbb{E}\left[\langle X_i, v \rangle^2 \right] - \mathbb{E}\left[\left[\langle X_i, v \rangle^2 \mathbf{1}_{\langle X_i, v \rangle^2 \geq B} \right] \right] \\
&\geq \mathbb{E}\left[\langle X_i, v \rangle^2 \right] - \frac{\mathbb{E}\left[\langle X_i, v \rangle^4 \right]}{B} \\
&\geq \langle v, \Sigma\, v \rangle - \frac{\kappa \langle v, \Sigma\, v \rangle^2}{B}.
\end{aligned}$$

We finish the proof combining these bounds on (3.10). □

3.5.2 The smoothed empirical process.

We now apply the PAC-Bayesian method to the smoother empirical process with truncated terms defined in the beginning of this section.

Lemma 3.11. *Under Assumption 3.1. Consider $B, \gamma > 0$. Then*

$$\widetilde{\varepsilon}_\theta(B) \leq \sqrt{\frac{8\,\kappa}{n}} \left(\|\Sigma\|_{\text{op}} + \gamma^2 \operatorname{tr}(\Sigma) \right) \sqrt{2\log(1/\alpha) + \gamma^{-2}} + \frac{(2\log(1/\alpha) + \gamma^{-2})\, B}{6\, n},$$

with probability at least $1-\alpha$.

Proof. Applying Proposition 3.6 with

$$\begin{aligned} Z_i(\theta) &= (\langle X_i, \theta\rangle^2 \wedge B)_{i\in[n],\theta\in\mathbb{R}^d}, \text{ and} \\ \bar{\sigma}_\xi &= \sup_{\theta\in\mathbb{R}^d} \sum_{i=1}^n \Gamma_{v,\gamma} \mathbb{V}\left[\langle X_i, \theta\rangle^2 \wedge B\right] \end{aligned}$$

it follows with probability at least $1-\alpha$ that for all $v \in \mathbb{S}^{d-1}$,

$$\sum_{i=1}^n \Gamma_{v,\gamma}\left(\langle X_i, \theta\rangle^2 \wedge B - \mathbb{E}\left[\langle X_i, \theta\rangle^2 \wedge B\right]\right)) \le S\sqrt{2\log(1/\alpha)+\gamma^{-2}} + \frac{(2\log(1/\alpha)+\gamma^{-2})B}{6}. \tag{3.11}$$

Let us now observe that

$$\begin{aligned} \Gamma_{v,\gamma}\mathbb{V}\left[\langle X_i, \theta\rangle^2 \wedge B\right] &= \Gamma_{v,\gamma}\mathbb{E}\left[\langle X_i, \theta\rangle^4\right] - \Gamma_{v,\gamma}\left(\mathbb{E}\left[\langle X_i, \theta\rangle^2\right]\right)^2 \\ &\le \Gamma_{0,1}\mathbb{E}\left[\langle X_i, v + c\,\theta\rangle^4\right]. \\ &\le 8\,\kappa\left(\|\Sigma\|_{\mathrm{op}}^2 + \gamma^4 \operatorname{tr}(\Sigma)^2\right) \\ &\le 8\,\kappa\left(\|\Sigma\|_{\mathrm{op}} + \gamma^2 \operatorname{tr}(\Sigma)\right)^2. \end{aligned}$$

Using this bound in (3.11), the claim follows. □

3.5.3 Comparison of empirical processes.

As our next step, we find a bound for the difference between the smoothed and unsmoothed truncated empirical processes $\widetilde{\varepsilon}_\theta(B)$ and $\varepsilon_\theta(B)$ defined above. We will need the following lemma.

Lemma 3.12. *Let $v \in \mathbb{R}^d$, $\gamma, B > 0$. Assume $\{X_i : i \in S\}$ is a set of vectors indexed by the set S. Assume further that $2\,|\langle X_i, v\rangle| \le B$ and $2\,\gamma^2\|x\| \le B$ for all $i \in S$. Then*

$$\left|\Gamma_{v,\gamma}(\langle X_i, \theta\rangle^2 \wedge B) - \left(\langle X_i, v\rangle^2 + \gamma^2\|X_i\|^2\right)\right| \le \exp\left(\frac{-B}{8\gamma^2\|X_i\|^2}\right)\left(\frac{8\gamma^3\|X_i\|^3}{B^{\frac{1}{2}}} + \frac{16\gamma^5\|X_i\|^5}{B^{\frac{3}{2}}}\right).$$

Proof. We observe that for each $\{X_i : i \in S\}$ and $\theta \in \mathbb{R}^d$

$$\Gamma_{v,\gamma}(\langle X_i, \theta\rangle^2 \wedge B) = \mathbb{E}\left[N^2 \wedge B\right], \text{ with } N \sim \mathcal{N}(\langle X_i, v\rangle, \gamma^2\|X_i\|^2).$$

Note that $\mathbb{E}[N^2] = \langle X_i, v\rangle^2 + \gamma^2\|X_i\|^2$, then

$$\begin{aligned} \left|\mathbb{E}\left[N^2 \wedge B\right] - \left(\langle X_i, v\rangle^2 + \gamma^2\|X_i\|^2\right)\right| &= \mathbb{E}\left[N^2 \wedge B\right]_+ \\ &= \frac{\int_{\sqrt{B}}^{+\infty}(t^2 - B)\left(e^{-\frac{(t-\langle X_i, v\rangle)^2}{2\,\gamma^2\|X_i\|^2}} + e^{-\frac{(t+\langle X_i, v\rangle)^2}{2\,\gamma^2\|X_i\|^2}}\right)dt}{\gamma\|X_i\|\sqrt{2\,\pi}}. \end{aligned}$$

By change of variables, for all $t \geq \sqrt{B}$:

$$\exp\left(-\frac{(t \pm \langle X_i, v\rangle)^2}{2\,\gamma^2\|X_i\|^2}\right) \leq \exp\left(-\frac{(\sqrt{B} \pm \langle X_i, v\rangle)^2}{2\,\gamma^2\|X_i\|^2}\right)\exp\left(\frac{\sqrt{B}\langle X_i, v\rangle}{2\gamma^2\|X_i\|^2}\right),$$

and $t^2 - B = 2u\sqrt{B} + u^2$. Under our assumptions,

$$\mathbb{E}\left[N^2 \wedge B\right]_+ \leq \frac{\exp\left(-\frac{B}{8\gamma^2\|X_i\|^2}\right)}{\gamma\|X_i\|}\int_0^{+\infty}\left(2\sqrt{B}u + u^2\right)\exp\left(\frac{\sqrt{B}u}{2\gamma^2\|X_i\|^2}\right)du.$$

Now let us use the formula:

$$\forall \eta > 0, \forall a \in \mathbb{N}: \quad \int_0^{+\infty} u^a e^{-\eta u}du = \frac{a!}{\eta^{a+1}\gamma^2\|X_i\|^2},$$

with $a = 1, 2$ and $\eta = \sqrt{B}/(2)$, we obtain the result. □

The final result of this subsection establishes that the difference between the two processes can be controlled via a counting condition related to (3.4). This will require the introduction of a new event.

$$\text{Norm}(k) := \bigcap_{j\geq 1}\left\{\#\left\{i \in [n] \,:\, \|X_i\| \geq \sqrt{e\,\kappa}\left(\frac{n}{k}\right)^{\frac{1}{4}}\sqrt{\text{tr}\,(\Sigma)}\right\} \leq jk\right\}. \tag{3.12}$$

In the Appendix A.4 we show that $\mathbb{P}\left[\text{Norm}(k)\right] \geq 1 - e^{-k}/6$.

Lemma 3.13. *Make Assumption 3.1. Given $B, \gamma > 0$ and a vector $v \in \mathbb{S}^{d-1}$, define the set:*

$$\mathsf{Good}_B(v) := \left\{i \in [n] \,:\, B \geq 2\,\gamma^2\,\|X_i\|^2 \text{ and } B \geq 2|\langle X_i, v\rangle|^2\right\},$$

and assume that $\#\mathsf{Good}_B(v) \geq n - k$ for all $v \in \mathbb{S}^{d-1}$. Also assume that the event Norm*(k) holds. Then*

$$|\varepsilon_\theta(B) - \widetilde{\varepsilon}_\theta(B)| \leq \left|\frac{1}{n}\sum_{i=1}^n\left((\gamma^2\|X_i\|^2) \wedge B - \mathbb{E}\left[(\gamma^2\|X_i\|^2) \wedge B\right]\right)\right| + \frac{B\,k}{n}\,c$$

where $c > 0$ is the same as in Lemma 3.10.

Proof. Let us define $f(\theta, x) := \langle X_i, \theta\rangle^2 \wedge B + (\gamma^2\|X_i\|^2) \wedge B$ for all $\theta \in \mathbb{R}^d$. Then, observe that

$|\varepsilon_\theta(B) - \widetilde{\varepsilon}_\theta(B)|$ is upper bounded by

$$\begin{aligned}
& \sup_{\|v\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \Gamma_{v,\gamma} \langle X_i, \theta \rangle^2 \wedge B - \langle X_i, v \rangle^2 \wedge B - \left(\Gamma_{v,\gamma} \mathbb{E}\left[\langle X_i, \theta \rangle^2 \wedge B \right] - \mathbb{E}\left[\langle X_i, v \rangle^2 \wedge B \right] \right) \right| \\
\leq & \left| \frac{1}{n} \sum_{i=1}^{n} \left((\gamma^2 \|X_i\|^2) \wedge B - \mathbb{E}\left[(\gamma^2 \|X_i\|^2) \wedge B \right] \right) \right| \\
+ & \sup_{\|v\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \Gamma_{v,\gamma} f(\theta, X_i) - f(v, X_i) - \mathbb{E}\left[\Gamma_{v,\gamma} f(\theta, X_i) - f(v, X_i) \right] \right|.
\end{aligned}$$

Let us first find a bound for the last term in the inequality above. Fix a vector $v \in \mathbb{S}^{d-1}$ and consider the indices $i \in \mathsf{Good}_{\mathsf{B}}(\mathsf{v})$. Lemma 3.12 implies:

$$\begin{aligned}
\left| \Gamma_{v,\gamma}(\langle X_i, \theta \rangle^2 \wedge B) - \left(\langle X_i, v \rangle^2 + \gamma^2 \|X_i\|^2 \right) \right| & \leq \exp\left(\frac{-B}{8\gamma^2 \|X_i\|^2} \right) \left(\frac{8\gamma^3 \|X_i\|^3}{B^{\frac{1}{2}}} + \frac{16\gamma^5 \|X_i\|^5}{B^{\frac{3}{2}}} \right) \\
& \leq \frac{16\gamma^3 \|X_i\|^2}{\sqrt{B}} \exp\left(-\frac{B}{8\gamma^2 \|X_i\|^2} \right) \\
& \leq \frac{32\sqrt{2}\, B}{2\, j_i^{3/2}} \exp\left(\frac{-j_i}{4} \right), \qquad (3.13)
\end{aligned}$$

where in the last inequality $j_i \in [n]$ is such that

$$\frac{B}{2(j_i+1)^{1/2}} \leq \gamma \|X_i\| \leq \frac{B}{2 j_i^{1/2}}.$$

Consider now the indices in $\mathsf{Bad}_B(v) := i \in [n] \setminus \mathsf{Good}_B(v)$. We use the bound $0 \leq f(\theta, x) \leq B$ for all $(\theta, x) \in \mathbb{R}^d \times \mathbb{R}^d$, then

$$\Gamma_{v,\gamma} f(\theta, X_i) - f(v, X_i) \leq B.$$

Therefore, by the bound (3.13) and the assumption (3.12), it follows

$$\begin{aligned}
& \sup_{\|v\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \Gamma_{v,\gamma} f(\theta, X_i) - f(v, X_i) - \mathbb{E}\left[\Gamma_{v,\gamma} f(\theta, X_i) - f(v, X_i) \right] \right| \\
\leq & \frac{1}{n} \sum_{i \in \mathsf{Good}(v)} 32\sqrt{2}\, \ell_j \frac{B}{2\, j_i^{3/2}} e^{-j_i/2} + \frac{\#(\mathsf{Bad}_B(v))\, B}{n} \\
\leq & \frac{1}{n} \sum_{j=1}^{+\infty} 32\sqrt{2}\, \ell_j \frac{B}{2\, j^{3/2}} e^{-j/2} + \frac{8\, B\, k}{n}. \qquad (3.14)
\end{aligned}$$

Here ℓ_j denote the number of indices $i \in [n]$ with $j_i = j$.
Given that $\mathsf{Norm}(k)$ holds,

$$\forall j \in \mathbb{N} \setminus \{0, 1\} \sum_{m=1}^{j-1} \ell_m \leq \# \left\{ i \in [n] \,:\, \|X_i\| \geq \sqrt{e\,\kappa} \left(\frac{n}{jk} \right)^{\frac{1}{4}} \sqrt{\mathrm{tr}\,(\Sigma)} \right\} \leq jk.$$

Then, the RHS first term in (3.14) is upper bounded by

$$\sup\left\{\sum_{j=1}^{\infty} 32\sqrt{2}\,\ell_j \frac{B}{2\,j^{3/2}} e^{-j/2} : \{\ell_m\}_{m\in\mathbb{N}}, \forall j>1, \sum_{m=1}^{j-1}\ell_m \le jk\right\}.$$

Since ℓ_j in the sum are multiplied by terms that decrease with j, the supremum is thus achieved at $\ell_1 = 2k$ and $\ell_2 = \ell_3 = \cdots = k$. As a consequence, the claim follows. □

3.5.4 Bounding the truncated empirical process. To finish the section, we combine the tools of the previous subsections to obtain Lemma 3.10.

Proof of Lemma 3.10. Assume that Norm(k) holds. Note that $\left(\gamma^2\|X_i\|^2\right)\wedge B \le B$, and the variance of each term is at most $\mathbb{E}\left[\left(\left(\gamma^2\|X_i\|^2\right)\wedge B\right)^2\right] \le \kappa\,\gamma^4\,\mathrm{tr}\left(\Sigma\right)^2$. We may use the Bernstein's concentration inequality in Lemma 3.13 above to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left(\gamma^2\|X_i\|^2\right)\wedge B - \mathbb{E}\left[\left(\gamma^2\|X_i\|^2\right)\wedge B\right]\right) \le \gamma^2\,\mathrm{tr}\left(\Sigma\right)\sqrt{\frac{2\,\kappa\log(2/\alpha)}{n}} + \frac{2\,B\log(2/\alpha)}{3n},$$

with probability at least $1-\alpha/2$. Then, Lemma 3.11 implies the result.

□

3.6 From vectors to matrices

The tools we have developed in the two previous sections control expressions involving terms like $\langle X_i, v\rangle^2$ where $v \in \mathbb{S}^{d-1}$ has unit norm. As explained in subsection 3.3.2, this will not suffice for our minimax-based analysis. We will thus need to extend our control of expressions involving $\langle X_i, v\rangle^2$ to objects of the form $\mathrm{tr}\left(X_i X_i^T M\right) = \langle X_i, MX_i\rangle$ with $M \in \mathcal{D}(\mathbb{R}^d)$ (recall that $\mathcal{D}(\mathbb{R}^d)$ is the set of density matrices introduced in subsection 3.3.2). The following construction will be useful to move between matrices and vectors.

Definition 3.14. Consider $M \in \mathcal{D}(\mathbb{R}^d)$ and its spectral decomposition:

$$M = \sum_{s=1}^{d}\lambda_s \xi_s \xi_s^\top,$$

where ξ_s are orthonormal vectors in $\mathbb{R}^d$, $\lambda_s \ge 0$, and $\sum_{s=1}^d \lambda_s = 1$. We define

$$\mathsf{g}_M := \sum_{s=1}^{d}\sqrt{\lambda_s}\epsilon_s\xi_s$$

where the $\epsilon_s \sim \{-1, +1\}$ are i.i.d. uniform (Rademacher) random variables. $\mathbf{E}_{\mathtt{g}_M}, \mathbf{P}_{\mathtt{g}_M}$ denote expectation and probability with respect only to $\mathtt{g}_M$.

It is easy to verify that:

$$\|\mathtt{g}_M\| = 1 \text{ almost surely and } \mathbf{E}_{\mathtt{g}_M}\left\{\mathtt{g}_M \mathtt{g}_M^\top\right\} = M.$$

From now on, we explore the $\mathtt{g}_M$ construction to obtain counting and concentration results for expressions involving matrices. The main result of this section is the control of the upper tail:

Proposition 3.15. *Make Assumption 3.1. Consider $r \in \mathbb{N} \setminus \{0, 1\}$ and $B > 0$ such that:*

$$B = 24\sqrt{2\kappa}\sqrt{\frac{n}{\lceil \frac{k}{92} \rceil}}\,\|\Sigma\|_{\mathrm{op}}\,.$$

Then,

$$\inf_{\lambda \in \Delta_{n,k}} \sup_{M \in \mathcal{D}(\mathbb{R}^d)} \left\{\mathrm{tr}\left(\widehat{\Sigma}(\lambda) M\right) - \mathrm{tr}\left(\Sigma M\right)\right\} \leq \frac{C\,B\,k}{n} + \varepsilon_\theta(B)$$

with probability at least $1 - C_0\, e^{-c_0\, k}$. Here the constant $C > 0$ is the same as in Lemma 3.18 above, and c_0 and c_1 are absolute constants in $(0, 1)$.

The proof of 3.15 splits into two main parts. One component of the proof is to solve a minmax problem (Proposition 3.17). The second component of the proof is to control the following expression.

$$\frac{1}{n}\sum_{i=1}^{n} \langle X_i, MX_i\rangle \wedge B - \mathbf{E}_{\mathtt{g}_M}\left[\langle X_i, \mathtt{g}_M\rangle^2 \wedge B\right].$$

For solving both parts, it is crucial the passage from M to $\mathtt{g}_M$.

We start by using the general PAC-Bayesian statement from section 3.2.1 for counting the number of terms $\langle X_i, MX_i\rangle$ that can be large.

Lemma 3.16. *Under Assumption 3.1. Suppose $v \in \mathbb{S}^{d-1}$ and consider $B > 0$ such that:*

$$B = 24\sqrt{2\,\kappa}\sqrt{\frac{n}{t}}\,\|\Sigma\|_{\mathrm{op}}\,,$$

Let be $M \in \mathcal{D}(\mathbb{R}^d)$, then

$$\sup_{M \in \mathcal{D}(\mathbb{R}^d)} \sum_{i=1}^{n} \mathbf{1}_{\langle X_i, MX_i\rangle \geq B} \leq \frac{92}{3}\,t$$

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

Proof. Observe that $\|\mathrm{g}_M\| = 1$ a.s. and $\mathbf{E}_{\mathrm{g}_M}\left\{\mathrm{g}_M \mathrm{g}_M^\top\right\} = M$ imply

$$\mathbf{E}_{\mathrm{g}_M}\langle \mathrm{g}_M, X_i\rangle^2 = \mathbf{E}_{\mathrm{g}_M}\langle X_i, \mathrm{g}_M \mathrm{g}_M^\top X_i\rangle = \langle X_i, MX_i\rangle \text{ and}$$

$$\langle X_i, MX_i\rangle \wedge B \geq 2\,\mathbf{E}_{\mathrm{g}_M}\{4\,\langle X_i, \mathrm{g}_M\rangle^2 \wedge B\}. \tag{3.15}$$

Therefore, since $D(\mathbb{R}^d) =$ convex hull $\left\{vv^\top : v \in \mathbb{R}^d, \|v\| = 1\right\}$,

$$\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \sum_{i=1}^n \mathbf{1}_{\langle X_i, MX_i\rangle \geq B} \;\leq\; 8 \sup_{M\in\mathcal{D}(\mathbb{R}^d)} \mathbf{E}_{\mathrm{g}_M} \sum_{i=1}^n \mathbf{1}_{\langle X_i, \mathrm{g}_M\rangle^2 \geq B/4}.$$

Therefore, by Corollary 3.8, the result follows. □

The next result shows that the minimax problem can be upper bounded via truncation.

Proposition 3.17. *Under Assumption 3.1. Set $B > 0$ as in Lemma 3.16, and $k \in \mathbb{N} \setminus \{0, 1\}$. Then,*

$$\begin{aligned} &\inf_{\lambda\in\Delta_{n,k}} \sup_{M\in\mathcal{D}(\mathbb{R}^d)} \left\{\mathrm{tr}\left(\widehat{\Sigma}(\lambda)M\right) - \mathrm{tr}\left(\Sigma M\right)\right\} \\ \leq\;& \sup_{M\in\mathcal{D}(\mathbb{R}^d)} \frac{1}{n}\left\{\sum_{i=1}^n \langle X_i, MX_i\rangle \wedge B - \mathbf{E}_{\mathrm{g}_M}[\langle X_i, \mathrm{g}_M\rangle^2 \wedge B]\right\} + \varepsilon_\theta(B), \end{aligned} \tag{3.16}$$

with probability at least $1 - e^{-\lceil c_0\, k\rceil}$.

Proof. $\Delta_{n,k}$ and $\mathcal{D}(\mathbb{R}^d)$ are convex and compact sets, and $\mathrm{tr}\left(\widehat{\Sigma}(\lambda)M - \Sigma\, M\right)$ is affine in both λ and M. By Sion's minimax theorem:

$$\inf_{\lambda\in\Delta_{n,k}} \sup_{M\in\mathcal{D}(\mathbb{R}^d)} \left\{\mathrm{tr}\left(\widehat{\Sigma}(\lambda)M\right) - \mathrm{tr}\left(\Sigma M\right)\right\} = \sup_{M\in\mathcal{D}(\mathbb{R}^d)} \inf_{\lambda\in\Delta_{n,k}} \left\{\mathrm{tr}\left(\widehat{\Sigma}(\lambda)M\right) - \mathrm{tr}\left(\Sigma M\right)\right\}.$$

Now, by definition the RHS equals

$$\begin{aligned} &\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \inf_{\lambda\in\Delta_{n,k}} \left\{\sum_{i=1}^n \lambda_i \langle X_i, MX_i\rangle - \mathbb{E}\left[\langle X_i, MX_i\rangle\right]\right\} \\ =\;& \sup_{M\in\mathcal{D}(\mathbb{R}^d)} \frac{1}{n-k}\left\{\text{sum of the smallest } n-k \text{ values of } \langle X_i, MX_i\rangle - \mathbb{E}\left[\langle X_i, MX_i\rangle\right]\right\} \end{aligned} \tag{3.17}$$

Setting $k = \lceil (92/3)\, r\rceil$, by Lemma 3.16 $\#\{i \in [n] : |\langle X_1, MX_i\rangle| \geq B\} \leq k$ with probability at least $1 - e^{-r}$. Since $\mathbb{E}\left[\langle X_i, MX_i\rangle\right] \geq \mathbb{E}\left[\langle X_i, MX_i\rangle \wedge B\right]$, the sum in (3.17) is less or equal to

the sum of $\langle X_i, MX_i\rangle \wedge B - \mathbb{E}\left[\langle X_i, MX_i\rangle \wedge B\right]$. This implies that the RHS is upper bounded by

$$\begin{aligned}
&\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \frac{1}{n}\left\{\sum_{i=1}^n \langle X_i, MX_i\rangle \wedge B - \mathbb{E}\left[\langle X_i, MX_i\rangle \wedge B\right]\right\}\\
\leq\ &\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \frac{1}{n}\left\{\sum_{i=1}^n \langle X_i, MX_i\rangle \wedge B - \mathbf{E}_{\mathsf{g}_M}[\langle X_i, \mathsf{g}_M\rangle^2 \wedge B]\right\}\\
+\ &\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \frac{1}{n}\left\{\sum_{i=1}^n \mathbf{E}_{\mathsf{g}_M}[\langle X_i, \mathsf{g}_M\rangle^2 \wedge B] - \mathbb{E}\left[\langle X_i, MX_i\rangle \wedge B\right]\right\}.
\end{aligned}$$

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

We find a bound for the second term in the RHS. Since the map $\psi(t) = t \wedge B$ ($t \in \mathbb{R}$) is concave,

$$\begin{aligned}
\langle X_i, MX_i\rangle \wedge B &= \psi(\langle X_i, MX_i\rangle)\\
&= \psi(\mathbf{E}_{\mathsf{g}_M}[\langle X_i, g_M\rangle^2])\\
\text{(Jensen's ineq.)} \quad &\geq \mathbf{E}_{\mathsf{g}_M}[\psi(\langle X_i, g_M\rangle^2)]\\
&= \mathbf{E}_{\mathsf{g}_M}\left[\langle X_i, g_M\rangle^2) \wedge B\right].
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \frac{1}{n}\left\{\mathbf{E}_{\mathsf{g}_M}[\langle X_i, \mathsf{g}_M\rangle^2 \wedge B] - \mathbb{E}\left[\langle X_i, MX_i\rangle \wedge B\right]\right\}\\
\leq\ &\sup_{M\in\mathcal{D}(\mathbb{R}^d)} \mathbf{E}_{\mathsf{g}_M}\left[\frac{1}{n}\sum_{i=1}^n \left(\langle X_i, \mathsf{g}_M\rangle^2 \wedge B - \mathbb{E}\left[\langle X_i, \mathsf{g}_M\rangle^2 \wedge B\right]\right)\right]\\
\leq\ &\sup_{v\in\mathbb{R}^d,\, \|v\|=1} \left\{\frac{1}{n}\sum_{i=1}^n \left(\langle X_i, v\rangle^2 \wedge B - \mathbb{E}\left[\langle X_i, v\rangle^2 \wedge B\right]\right)\right\}\\
\leq\ &\varepsilon_\theta(B).
\end{aligned}$$

Therefore, the result follows. □

Lemma 3.11 gives a bound for the last term ($\varepsilon_\theta(B)$ in the RHS of 3.16. The following Lemma analyses the first term. For that we first need to define the counting event:

$$\mathsf{Count}_k(M) := \bigcap_{j=1}^{\infty}\left\{\#\left(i \in [n]\ :\ |\langle X_i, MX_i\rangle| \geq 12\sqrt{\frac{2\,\kappa\, n}{\lceil \frac{k\,j}{92}\rceil}}\,\|\Sigma\|_{\mathrm{op}}\right) \leq jk\right\}. \tag{3.18}$$

Lemma A.5 in the appendix, based on Lemma 3.16, shows that or all $r \in \mathbb{N}\setminus\{0,1\}$:

$$\mathbb{P}\left[\bigcap_{M\in\mathcal{D}(\mathbb{R}^d)} \mathsf{Count}_k(M)\right] \geq 1 - \frac{e^{-\lceil \frac{k}{92}\rceil}}{1 - e^{-\lceil \frac{k}{92}\rceil}}.$$

Lemma 3.18. *Make Assumption 3.1. Assume $B > 0$, and $k \in \mathbb{N} \setminus \{0, 1\}$. Then, in the event* $\mathsf{Count}_k(M)$

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, MX_i\rangle \wedge B - \mathbf{E}_{\mathsf{g}_M}\left[\langle X_i, \mathsf{g}_M\rangle^2 \wedge B\right] \leq \frac{C\,B\,k}{n}.$$

where C is a universal positive constant.

Proof. Fix $M \in \mathcal{D}(\mathbb{R}^d)$. Define the sets of indices:

$$\mathsf{Good}_k(M) := \{i \in [n] \,:\, \langle X_i, MX_i\rangle \leq B\} \text{ and } \mathsf{Bad}_k(M) := [n]\backslash\mathsf{Good}_k(M).$$

When $\mathsf{Count}_k(M)$ holds, $\#\mathsf{Bad}_k(M) \leq k$ and therefore:

$$\frac{1}{n}\sum_{i\in\mathsf{Bad}_k(M)}\langle X_i, MX_i\rangle \wedge B - \mathbf{E}_{\mathsf{g}_M}\left[\langle X_i, \mathsf{g}_M\rangle^2 \wedge B\right] \leq \frac{B\,k}{n}. \tag{3.19}$$

We now consider the "good indices" $i \in \mathsf{Good}_k(M)$ for which $\langle X_i, MX_i\rangle \wedge B = \langle X_i, MX_i\rangle$. For each such index:

$$\langle X_i, MX_i\rangle - \mathbf{E}_{\mathsf{g}_M}\langle X_i, \mathsf{g}_M\rangle^2 \wedge B = \mathbf{E}_{\mathsf{g}_M}\left[\langle X_i, \mathsf{g}_M\rangle^2 - \langle X_i, \mathsf{g}_M\rangle^2 \wedge B\right] = \mathbf{E}_{\mathsf{g}_M}\left[(\langle X_i, \mathsf{g}_M\rangle^2 - B)_+\right].$$

Furthermore,

$$\mathbf{E}_{\mathsf{g}_M}\left(\langle X_i, \mathsf{g}_M\rangle^2 - B\right)_+ = \int_0^\infty \mathbf{P}_{\mathsf{g}_M}\left(\langle X_i, \mathsf{g}_M\rangle^2 - B > t\right)dt. \tag{3.20}$$

By Definition 3.14,

$$\langle X_i, \mathsf{g}_M\rangle = \sum_{s=1}^{d}\epsilon_s\, a_s$$

where the ϵ_s are independent Rademacher r.v.'s and the weights $a_s := \sqrt{\lambda_s}\langle \xi_s, X_i\rangle$ satisfy:

$$\sum_{s=1}^{d} a_s^2 = \sum_{s=1}^{d}\lambda_s\langle \xi_s, X_i\rangle^2 = \langle X_i, MX_i\rangle.$$

Therefore,

$$\forall x \geq 0 \,:\, \mathbf{P}_{\mathsf{g}_M}\{|\langle X_i, \mathsf{g}_M\rangle| \geq x\} \leq 2\exp\left(-\frac{x^2}{2\langle X_i, MX_i\rangle}\right).$$

Plugging this into (3.20) we obtain:

$$\mathbf{E}_{\mathsf{g}_M}\left(\langle X_i, \mathsf{g}_M\rangle^2 - B\right)_+ \leq \int_0^{+\infty} 2\exp\left(-\frac{t+B}{2\langle X_i, MX_i\rangle}\right) = 4\,\langle X_i, MX_i\rangle\, e^{-\frac{B}{2\langle X_i, MX_i\rangle}}. \tag{3.21}$$

Then, for each $i \in \mathsf{Good}_k(M)$, one can find an integer $j_i \geq 1$ such that:

$$B(j_i^{-1/2}) \leq \langle X_i, MX_i\rangle \leq B(j_{i+1}^{-1/2}).$$

This bound in (3.21) implies:

$$\begin{aligned}\sum_{i\in\mathsf{Good}_k(M)} \mathbf{E}_{\mathbf{g}_M}\left(\langle X_i,\mathbf{g}_M\rangle^2 - B\right)_+ &\leq \sum_{i\in\mathsf{Good}_k(M)} 4\,\langle X_i, MX_i\rangle\, e^{-\frac{B}{2\langle X_i, MX_i\rangle}} \\ &\leq \sum_{i\in\mathsf{Good}_k(M)} 4\,\ell_j\, B\,\frac{e^{-\frac{j_i^{1/2}}{2}}}{j_i^{1/2}} \\ &\leq \sum_{j=1}^{+\infty} 4\,\ell_j\, B\,\frac{e^{-\frac{j^{1/2}}{2}}}{j^{1/2}}. \end{aligned} \tag{3.22}$$

where ℓ_j is the number of indices $i \in \mathsf{Good}_k(M)$ with $j_i = j$. Given that $\mathsf{Count}_k(M)$ holds,

$$\forall j \in \mathbb{N}\setminus\{0,1\} \sum_{m=1}^{j-1} \ell_m \leq \#\left(i\in[n] \,:\, |\langle X_i, MX_i\rangle| \geq 24\sqrt{\frac{\kappa\, n}{\lceil \frac{kj}{92}\rceil}}\,\|\Sigma\|_{\mathrm{op}}\right) \leq rj.$$

Then, the RHS in (3.22) is upper bounded by

$$\sup\left\{\sum_{j=1}^{\infty} 4\,\ell_j\,\frac{B}{j^{1/2}} e^{-\frac{j^{1/2}}{2}} \,:\, \{\ell_m\}_{m\in\mathbb{N}}, \forall j>1, \sum_{m=1}^{j-1}\ell_m \leq jk\right\}.$$

Since ℓ_j in the sum are multiplied by terms that decrease with j, the supremum is thus achieved at $\ell_1 = 2r$ and $\ell_2 = \ell_3 = \cdots = r$. As a consequence, under $\mathsf{Count}_k(M)$,

$$\sum_{i\in\mathsf{Good}_k(M)} \mathbf{E}_{\mathbf{g}_M}\left(\langle X_i,\mathbf{g}_M\rangle^2 - B\right)_+ \leq 8\,B\,k \sum_{j\geq 1}\frac{e^{-\frac{j^{1/2}}{2}}}{j^{1/2}}.$$

Combining this with (3.19) finishes the proof. □

The combination of the last two propositions immediately gives the main result (Proposition 3.15).

3.7 The final estimator

3.7.1 The estimator. In this section, we construct the covariance estimator. The previous two sections allow us to control the lower and upper tails respectively. Therefore, we can prove the following result.

Theorem 3.19. *Under Assumption 3.1. Consider $k \in \mathbb{N}\setminus\{0,1\}$. Then the following holds with probability at least $1 - C_1\, e^{-c_1\, k} - \alpha$, :*

1. *for all* $\lambda \in \Delta_{n,k}$*:*

$$\inf_{v \in \mathbb{S}^{d-1}} \langle v, \widehat{\Sigma}(\lambda)\, v\rangle - \langle v, \Sigma v\rangle \geq \varepsilon_1(k);$$

2. *there is* $\lambda_\star \in \Delta_{n,k}$ *such that:*

$$\|\widehat{\Sigma}(\lambda_\star) - \Sigma\|_{\mathrm{op}} \leq \max\{\varepsilon_1(k), \varepsilon_2(k)\}.$$

Above:

$$\varepsilon_1(k) = c(\kappa)\, \|\Sigma\|_{\mathrm{op}} \left(\sqrt{\frac{k}{n}} + \sqrt{\frac{\log(2/\alpha)}{n}} + \sqrt{\frac{\log(2/\alpha) + r(\Sigma)}{n}} + \frac{\log(2/\alpha) + r(\Sigma)}{\sqrt{nk}} + \frac{\log(2/\alpha)}{\sqrt{n\,k}} \right)$$

and $\varepsilon_2(k)$ *is equal to a positive universal constant times* $\varepsilon_1(k)$*.*

At the end of this section, we prove this statement. We will also obtain a result that allows us to select a good weight vector $\widehat{\lambda}$.

Proposition 3.20. *Assume that items 1 and 2 in Theorem 3.19hold. Select:*

$$\widehat{\lambda}_k = \arg \min_{\lambda \in \Delta_{n,k}} \sup_{\lambda' \in \Delta_{n,k}} \sigma_{\max}(\widehat{\Sigma}(\lambda) - \widehat{\Sigma}(\lambda')).$$

Then:

$$\|\widehat{\Sigma}(\lambda) - \Sigma\|_{\mathrm{op}} \leq 3\, \varepsilon_1(k) \vee \varepsilon_2(k).$$

Proof. Theorem 3.19 above implies that if we choose $\lambda_\star$ as in its item 2,

$$\|\widehat{\Sigma}(\lambda_\star) - \Sigma\|_{\mathrm{op}} \leq \varepsilon_1(k) \vee \varepsilon_2(k).$$

Consider now $\widehat{\lambda}_k$ and let us define:

$$R_k(\lambda) := \sup_{\lambda' \in \Delta_{n,k}} \sigma_{\max}(\widehat{\Sigma}(\lambda) - \widehat{\Sigma}(\lambda')).$$

Since $\widehat{\lambda}_k$ minimizes this risk, it follows

$$R_k(\widehat{\lambda}_k) \leq R(\lambda_\star).$$

At the same time,

$$\begin{aligned} R_k(\widehat{\lambda}_k) &\geq \sigma_{\max}(\widehat{\Sigma}(\widehat{\lambda}_k) - \widehat{\Sigma}(\lambda_\star)) \\ &\geq \sigma_{\max}(\widehat{\Sigma}(\widehat{\lambda}_k) - \Sigma) - \|\widehat{\Sigma}(\lambda_\star) - \Sigma\|_{\mathrm{op}} \\ &\geq \sigma_{\max}(\widehat{\Sigma}(\widehat{\lambda}_k) - \Sigma) - \varepsilon_1(k) \vee \varepsilon_2(k), \end{aligned}$$

and

$$\begin{aligned} R_k(\lambda_\star) &= \sup_{\|v\|=1} \sup_{\lambda' \in \Delta_{n,k}} \langle v, (\widehat{\Sigma}(\lambda_\star) - \widehat{\Sigma}(\lambda'))\, v \rangle \\ &\leq \sup_{\|v\|=1} \langle v, (\widehat{\Sigma}(\lambda_\star) - \Sigma)\, v \rangle + \varepsilon_1(k) \\ &\leq \|\widehat{\Sigma}(\lambda_\star) - \Sigma\|_{\text{op}} + \varepsilon_1(k) \\ &\leq \varepsilon_1(k) \vee \varepsilon_2(k) + \varepsilon_1(k). \end{aligned}$$

We conclude that

$$\sigma_{\max}(\widehat{\Sigma}(\widehat{\lambda}_k) - \Sigma) \leq 3\varepsilon_1(k) \vee \varepsilon_2(k).$$

By part 1 of Theorem 3.20, the smallest eigenvalue of $\widehat{\Sigma}(\widehat{\lambda}_k) - \Sigma$ is greater than or equal to $-\varepsilon_1(k)$. We conclude:

$$\|\widehat{\Sigma}(\widehat{\lambda}_k) - \Sigma\|_{\text{op}} \leq 3\varepsilon_1(k) \vee \varepsilon_3(k).$$

□

Observe that this result holds for any $k < n$. For the choice of k to give the optimal bounds in the main result, two conditions are necessary. The parameter k should be large enough such that

$$\frac{\log(1/\alpha) + r(\Sigma)}{\sqrt{nk}} = O\left(\sqrt{\frac{\log(1/\alpha)}{n}}\right),$$

and simultaneously we need a k small enough such that $\sqrt{k/n}$ also has that order. Therefore, we need k to be of the order $r(\Sigma) + \log(2/\alpha)$. Here the difficulty is that $r(\Sigma)$ is unknown. Therefore, first we need to estimate $r(\Sigma)$ as the procedure presented in section 3.3.3.
We end this section with the proof of Theorem 3.19. Each item of the result will be proved in the following two lemmas respectively.

Lemma 3.21. *Make Assumption 3.1. Assume $B > 0$ such that:*

$$B = 6\sqrt{2\,\kappa}\sqrt{\frac{n}{k}}\,\|\Sigma\|_{\text{op}}\,,$$

then for all $\lambda \in \Delta_{n,k}$ and $v \in \mathbb{S}^{d-1}$ with probability at least $1 - \alpha - e^{-k}/6$:

$$\begin{aligned} \inf_{\lambda \in \Delta_{n,k}} \langle v, \widehat{\Sigma}(\lambda) v \rangle &\geq \langle v, \Sigma\, v \rangle - c(\kappa)\,\|\Sigma\|_{\text{op}} \left(\sqrt{\frac{k}{n}} + \sqrt{\frac{2\log(2/\alpha)}{n}}\right) \\ &- c(\kappa)\,\|\Sigma\|_{\text{op}} \left(\sqrt{\frac{k}{n}} + \sqrt{\frac{\log(2/\alpha) + r(\Sigma)}{n}} + \frac{\log(2/\alpha) + r(\Sigma)}{\sqrt{nk}} + \frac{\log(2/\alpha)}{\sqrt{nk}}\right). \end{aligned}$$

Above, $c(\kappa)$ is a positive constant that depends on κ only.

Proof. In Norm(k), by combining Lemma 3.9 with the Lemma 3.10, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\langle X_i, v\rangle^2 \wedge B &\geq \langle v, \Sigma\, v\rangle - \frac{\kappa\,\langle v, \Sigma\, v\rangle^2}{B} - \gamma^2\,\mathrm{tr}\,(\Sigma)\,\sqrt{\frac{2\,\kappa\log(2/\alpha)}{n}} - \frac{2\,B\log(2/\alpha)}{3n} \\ &- \frac{B\,k}{n}\,c - 2\sqrt{2\,\kappa}\left(\|\Sigma\|_{\mathrm{op}} + \gamma^2\,\mathrm{tr}\,(\Sigma)\right)\sqrt{\frac{2\log(2/\alpha)+\gamma^{-2}}{n}} \\ &- \frac{(2\log(2/\alpha)+\gamma^{-2})\,B}{6\,n}.\end{aligned}$$

with probability at least $1-\alpha-e^{-k}/6$. As a consequence of the choice of B and $\gamma^{-2}=2r(\Sigma)$, the result follows. □

Lemma 3.22. *Make Assumption 3.1. Define:*

$$B = 24\sqrt{2\kappa}\sqrt{\frac{n}{\lceil\frac{3k}{92}\rceil}}\,\|\Sigma\|_{\mathrm{op}}\,.$$

Then with probability at least $1-C_1\,e^{-c_1\,k}-\alpha$,

$$\begin{aligned}&\inf_{\lambda\in\Delta_{n,k}}\sup_{M\in\mathcal{D}(\mathbb{R}^d)}\left\{\mathrm{tr}\left(\widehat{\Sigma}(\lambda)M\right)-\mathrm{tr}\,(\Sigma M)\right\} \leq c(\kappa)\sqrt{\frac{k}{n}}\,\|\Sigma\|_{\mathrm{op}} \\ +\;& c(\kappa)\,\|\Sigma\|_{\mathrm{op}}\left(\frac{\log(2/\alpha)+r(\Sigma)}{\sqrt{n\,k}}+\frac{\log(2/\alpha)}{\sqrt{n\,k}}+\sqrt{\frac{\log(2/\alpha)}{n}}+\sqrt{\frac{\log(2/\alpha)+r(\Sigma)}{n}}\right).\end{aligned}$$

Proof. Assume that Norm(k) and Count$_r(M)$ hold. Combing both Proposition 3.15 and Lemma 3.10, if we set $B>0$ like in Lemma 3.15, it follows

$$\begin{aligned}&\inf_{\lambda\in\Delta_{n,k}}\sup_{M\in\mathcal{D}(\mathbb{R}^d)}\left\{\mathrm{tr}\left(\widehat{\Sigma}(\lambda)M\right)-\mathrm{tr}\,(\Sigma M)\right\} \\ \leq\;& \frac{C\,B\,k}{n}+\sqrt{\frac{8\,\kappa}{n}}\,\|\Sigma\|_{\mathrm{op}}\left(1+\gamma^2 r(\Sigma)\right)\sqrt{2\log(2/\alpha)+\gamma^{-2}}+\frac{(2\log(2/\alpha)+\gamma^{-2})\,B}{6\,n} \\ +\;& \gamma^2\,\mathrm{tr}\,(\Sigma)\sqrt{\frac{2\,\kappa\log(2/\alpha)}{n}}+\frac{2\,B\log(2/\alpha)}{3n}+\frac{B\,k}{n}\,c.\end{aligned}$$

with probability at least $1-C_1\,e^{-c_1\,k}-\alpha$, with constants $c_1\in(0,1)$ and $C>0$. Above, c and C are positive universal constants as in Lemma 3.13 and Lemma 3.18, respectively. Hence, setting $\gamma^{-2}=2\,r(\Sigma)$, the result follows. □

3.7.2 The final estimator.

The main goal of this section is obtain a good estimator for $r(\Sigma)$ and compute the final estimator. We begin the section by estimating the trace. Denote by $\widehat{\mathrm{tr}}(\Sigma)$ the trimmed mean estimator of $\mathrm{tr}\,(\Sigma)$. Since

$$\mathrm{tr}\,(\Sigma) = \mathbb{E}\left[\|X_1\|^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[\langle X_1, e_i\rangle^2\right],$$

we can use Theorem 2.17 in §2 for the random variable

$$\sum_{i=1}^{d} \mathbb{E}\left[\langle X_1, e_i\rangle^2\right].$$

Hence, given $\alpha \geq 4\, e^{-c\, n}$ and $k = \lfloor \lceil \log(8/\alpha) \rceil$, it follows with probability at least $1-\alpha$

$$\begin{aligned} |\,\mathrm{tr}\,(\Sigma) - \widehat{\mathrm{tr}}(\Sigma)| &\leq (1+h)\sqrt{2}\left(\mathbb{V}\left[\sum_{i=1}^{d} \mathbb{E}\left[\langle X_1, e_i\rangle^2\right]\right]\right)^{1/2} \sqrt{\frac{2\log(4/\alpha)}{n}} \\ &\leq (1+h)\sqrt{2\kappa}\sqrt{\frac{2\log(4/\alpha)}{n}}\,\mathrm{tr}\,(\Sigma), \end{aligned}$$

with $h \in (0,1)$.

As a consequence, if $n \geq c'(\kappa)(\log(4/\alpha)$, the following holds with probability at least $1-\alpha$.

$$\frac{1}{2}\,\mathrm{tr}\,(\Sigma) \geq \widehat{\mathrm{tr}}(\Sigma) \geq 2\,\mathrm{tr}\,(\Sigma). \tag{3.23}$$

Next, we estimate the norm $\|\Sigma\|_{\mathrm{op}}$ via $\|\widehat{\Sigma}(\widehat{\lambda}_{\epsilon n})\|_{\mathrm{op}}$, $\widehat{\Sigma}(\widehat{\lambda}_{\epsilon n})$ is as the estimator $\widehat{\Sigma}(\widehat{\lambda})$, but considering k as a small fraction $\epsilon > 0$ of n.

Lemma 3.23. *Make Assumption 3.1. Set $n > C(\kappa)(2\,\log(2/\alpha) + 2\,r(\Sigma))$. Then:*

$$\|\widehat{\Sigma}(\widehat{\lambda}_{\epsilon n}) - \Sigma\|_{\mathrm{op}} \leq \frac{\|\Sigma\|_{\mathrm{op}}}{2}$$

with probability at least $1 - C_1\, e^{-c_1\, k} - \alpha$,

Proof. Let us consider k as fraction ϵ of n. Then, by Lemma 3.20, with probability $1 - C_1\, e^{-c_1\, k} - \alpha$, it follows

$$\|\widehat{\Sigma}(\widehat{\lambda}_{\epsilon n}) - \Sigma\|_{\mathrm{op}} \leq \varepsilon(\epsilon n),$$

where

$$\begin{aligned}\varepsilon_1(\epsilon n) &= c(\kappa)\,\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\epsilon}+\sqrt{\frac{\log(2/\alpha)}{n}}+\sqrt{\frac{\log(2/\alpha)+r(\Sigma)}{n}}+\frac{\log(2/\alpha)+2\,r(\Sigma)}{n\sqrt{\epsilon}}\right)\\ &+ c(\kappa)\,\|\Sigma\|_{\mathrm{op}}\left(\frac{\log(2/\alpha)}{n\sqrt{\epsilon}}\right).\end{aligned}$$

Consider $\epsilon^{1/2} = 1/(10\,c(\kappa))$ and the universal constant $C(\kappa) > 100c(\kappa)^2$. Therefore, we conclude the proof. □

Now, we can estimate the effective rank. Define the effective rank estimator as

$$\widehat{r}(\Sigma) := \frac{\widehat{\mathrm{tr}}(\Sigma)}{\|\widehat{\Sigma}(\widehat{\lambda}_{\epsilon n})\|_{\mathrm{op}}}.$$

Under the assumptions of Lemma 3.23. Combining the norm bound from Lemma 3.23 above with (3.23), it follows directly

$$\frac{\widehat{r}(\Sigma)}{4} \le r(\Sigma) \le 4\,\widehat{r}(\Sigma)$$

with probability at least $1 - C_1\,e^{-c_1\,k} - 2\,\alpha$.

Finally, we end the section by showing the final estimator. Once we have a good estimate for $r(\Sigma)$, we can compute that the covariance matrix estimator has an error as following.

Proposition 3.24. *Under Assumption 3.1. Given a fixed confidence parameter* $1-\alpha \in (0,1)$*, set*

$$K_0 := C\left(\log(10/\alpha) + \frac{r(\Sigma)}{4}\right).$$

There exists an estimator $\widehat{\Sigma}$ *of* Σ *such that, if* $n > C(\kappa)\,K_0$*, then*

$$\|\widehat{\Sigma}(\widehat{\lambda}) - \Sigma\| \le C'(\kappa)\,\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\frac{\log(10/\alpha)+r(\Sigma)}{n}}+\sqrt{\frac{\log(10/\alpha)}{n}}\right)$$

with probability at least $1-\alpha$.

Proof. Consider $K_0 := C(\log(10/\alpha) + r(\Sigma)/4)$. Theorem 3.19 implies that for $k \ge K_0$, the following holds with probability at least

$$1 - \sum_{k\ge K_0}\left(C_1\,e^{-c_1\,k} - 4\,e^{r(\Sigma)/4}e^{-c_3\,k}\right) \ge 1 - C_0\,e^{-c_0 K_0}e^{r(\Sigma)/4}.$$

$$\|\widehat{\Sigma}(\widehat{\lambda}) - \Sigma\|_{\mathrm{op}} \le \varepsilon(k)$$

where

$$\varepsilon(k) = C(\kappa)\,\|\Sigma\|_{\mathrm{op}} \left(\sqrt{\frac{k}{n}} + \sqrt{\frac{\log(2/\alpha)}{n}} + \sqrt{\frac{\log(2/\alpha) + r(\Sigma)/4}{n}} + \frac{\log(2/\alpha) + r(\Sigma)/4}{\sqrt{nk}} + \frac{\log(2/\alpha)}{\sqrt{n\,k}} \right).$$

At the same time, if we set $k_0 = C(\log(10/\alpha) + \widehat{r}(\Sigma))$, then $k_0 \ge K_0$ with probability $1 - C_1 e^{c_1 k} e^{r(\Sigma)/4}$. Therefore,

$$\|\widehat{\Sigma}(\widehat{\lambda}) - \Sigma\|_{\mathrm{op}} \le \varepsilon(k_0)$$

with probability $1 - C_0\, e^{-c_0 k} e^{r(\Sigma)/4} - C_1 e^{-c_1\, k} e^{r(\Sigma)/4}$.

Finally, since

$$\frac{\widehat{r}(\Sigma)}{r(\Sigma)} \in [1/4, 4]$$

with probability at least $1 - C_2 e^{-c_2\, k}$, the result follows with probability

$$1 - C_0\, e^{-c_0 k} e^{r(\Sigma)/4} - C_1 e^{-c_1\, k} e^{r(\Sigma)/4} - C_2 e^{-c_2\, k} e^{r(\Sigma)/4} := 1 - \alpha.$$

□

Chapter 4

Conclusions

One of our main results in §2 is that the trimmed mean achieves minimax-optimal performance, up to constant factors, when the trimming parameter $k \approx \log(1/\alpha)$ under different moment conditions.

Research direction 1. Higher dimensions: There has also been great interest in extending the trimmed means results to higher dimensions. In [Lugosi and Mendelson, 2021], its statistical optimal performance is proved. Incidentally, that paper was inspired by an early version of this work. In this context, natural questions are the following.

1. Can we obtain better bounds?

2. Can we compute the trimmed means estimator in high dimensions?

Regarding the last question, Hopkins and Li in [Hopkins and Li, 2019] suggest that this may be difficult or impossible. Settling this problem is perhaps the most interesting question regarding computationally efficient estimation of high-dimensional means.

Chapter §3 provides an estimator of the covariance matrix Σ of random d-dimensional vector from an i.i.d. sample of size n. Our sole hypothesis is that this vector satisfies a bounded kurtosis (or $L^4 - L^2$ equivalence) assumption over its one-dimensional marginals. Given this, we show that Σ can be estimated from the sample with the same high-probability error rates that the sample covariance matrix achieves in the case of Gaussian observations. Our work leaves open important avenues for future research, which we outline following.

Research direction 2. Efficient algorithm: Our final estimator achieves the best possible

statistical performance. There has been progress progress on the computational side for the adversarial contamination and heavy-tailed data. However, many computational questions remain open. For instance, can we construct an optimal robust covariance matrix estimator under minimal assumptions that is computationally efficient?

Research direction 3. Linear Regression: Consider the problem of linear regression under adversarial contamination and heavy-tailed distribution. Assume also that sample size n is smaller than the dimension d assuming a sparse parameter. Using the mathematical techniques of the covariance matrix estimation in chapter §3, can we develop an iterative algorithm that is able to achieve the optimal estimation rate in reasonable time complexity? Currently, I am working with Roberto Imbuzeiro Oliveira (IMPA) and Philip Thompson (Purdue University) in that problem.

Appendix A

A.1 Some auxiliary technical results for Chapter §2

Lemma A.1 (Upper tail concentration of order statistics)**.** *Let $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ be the order statistics of an i.i.d.* Uniform$[0,1]$ *random sample. Then for all $k \in [n]$ and $t > 0$:*

$$\mathbb{P}\left[U_{(k)} > \frac{(\sqrt{k-1}+\sqrt{t})^2}{n}\right] = \mathbb{P}\left[1 - U_{(n-k+1)} > \frac{(\sqrt{k-1}+\sqrt{t})^2}{n}\right] \leq e^{-t}, \tag{A.1}$$

$$\mathbb{P}\left[U_{(k)} < \frac{(\sqrt{k}-\sqrt{t})^2}{n}\right] = \mathbb{P}\left[1 - U_{(n-k+1)} < \frac{(\sqrt{k}-\sqrt{t})^2}{n}\right] \leq e^{-2t}. \tag{A.2}$$

Proof. The equality of the two probabilities in each line follows from the symmetry of the uniform distribution under the transformation "$u \mapsto 1-u$."

We will use two bounds for the binomial distribution proven in [Okamoto, 1958, Theorems 3 and 4] (see also [Boucheron et al., 2013, Exercise 2.13]): for all $c > 0$,

$$\mathbb{P}\left[\sqrt{\frac{\text{Binomial}(n,\lambda)}{n}} < \sqrt{\lambda} - c\right] \leq e^{-c^2 n},$$

$$\mathbb{P}\left[\sqrt{\frac{\text{Binomial}(n,\lambda)}{n}} > \sqrt{\lambda} + c\right] \leq e^{-2c^2 n}.$$

Now, for any $\lambda \in (0,1)$, $U_{(k)} > \lambda$ if and only if the number of $i \in [n]$ with $U_i \leq \lambda$ is less than k. This gives:

$$\forall \lambda \in (0,1) \,:\, \mathbb{P}\left[U_{(k)} > \lambda\right] = \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{I}\{U_i \leq \lambda\} \leq k-1\right] = \mathbb{P}\left[\text{Binomial}(n,\lambda) \leq k-1\right].$$

For $\lambda \geq (k-1)/n$,

$$\mathbb{P}\left[\text{Binomial}(n,\lambda) \leq k-1\right] = \mathbb{P}\left[\sqrt{\frac{\text{Binomial}(n,\lambda)}{n}} < \sqrt{\lambda} - (\sqrt{\lambda} - \sqrt{\frac{k-1}{n}})\right] \leq e^{-(\sqrt{\lambda n}-\sqrt{k-1})^2}.$$

Taking:

$$\lambda = \frac{(\sqrt{k-1}+\sqrt{t})^2}{n}$$

as in the statement of the Lemma gives us (A.1). For (A.2), we note that $U_{(n-k+1)} > 1-\lambda$ if there are at least k points $U_i \in [1-\lambda, 1]$. Using [Okamoto, 1958, Theorem 3]:

$$\mathbb{P}\left[\sum_{i=1}^{n} \mathbb{I}\{U_i > 1-\lambda\} \geq k\right] = \mathbb{P}\left[\sqrt{\frac{\text{Binomial}(n,\lambda)}{n}} > \sqrt{\lambda} + (\sqrt{\frac{k}{n}} - \sqrt{\lambda})\right] \leq e^{-2(\sqrt{k}-\sqrt{\lambda n})^2}.$$

The choice of $\lambda = (\sqrt{k}-\sqrt{t})^2/n$ gives us (A.2). □

A.2 Technical lemmas for Chapter §3

Lemma A.2. *The set of weight vectors $\Delta_{n,k}$ defined at the beginning of section 3.3 is the convex hull of*

$$\left\{\frac{\mathbf{1}_S}{n-k} \,:\, S \subset [n],\, |S| = n-k\right\},$$

where $S \subset [n]$, and $\mathbf{1}_S \in \{0,1\}^n$ is the characteristic vector of S.

For prove this Lemma, we will need the following result:

Lemma A.3. *Define the index set*

$$H := \left\{i \in [n] : \lambda_i = \frac{1}{n-k}\right\}, \text{ and } \mathcal{A}_{n,s} := \left\{\frac{\mathbf{1}_S}{n-k} \,:\, S \subset [n],\, |S| = n-k\right\},$$

where $S \subset [n]$ and $\mathbf{1}_S \in \{0,1\}^n$ are as in Lemma A.2 above. Take $\lambda \in \Delta_{n,k} \setminus \mathcal{A}_{n,k}$. Then

1. *$\#H < n-k$,*
2. *there exist j_1 and $j_2 \in [n] \setminus H$ such that $j_1 \neq j_2$.*

Proof. We start the prove for the first item. Since,

$$\sum_{i \in H} \frac{1}{n-k} = \sum_{i \in H} \lambda_i \leq \sum_{i \in [n]} \lambda_i = 1.$$

Then, $\#H \leq n-k$. For proving the strict inequality, we assume $\#H = n-k$, then $\sum_{i\in H} \lambda_i = 1$ implies $\lambda_j = 0$ for $j \notin [n] \setminus H$. Therefore,

$$\lambda = \frac{\mathbf{1}_H}{n-k} \in \mathcal{A}_{n,k}.$$

This contradicts our hypothesis about λ.

Consider now $\lambda_i = 1/(n-k)$ for $i \in H$, with $\#H \leq n-k-1$, $\sum_{i\in H} \lambda_1 < 1$, then $0 < \lambda_{j_1} < 1/(n-k)$ for some $j_1 \notin H$. Then

$$\sum_{i\in H\cup\{j_1\}} \lambda_i = \frac{\#H}{n-k} + \lambda_{j_1} < \frac{n-k-1}{n-k} + \frac{1}{n-k} = 1.$$

It follows that $\exists j_2 \in [n] \setminus (H \cup \{j_1\})$. □

Proof of Lemma A.2 . Observe that $\Delta_{n,k}$ defined in Section 3.5 is a convex compact subset of $\mathbb{R}^n$ and it contains $\mathcal{A}_{n,k}$. Then, *convex hull* $(\mathcal{A}_{n,k}) \subset \Delta_{n,k}$.

Next, let us show the other inclusion. For this, we claim that

$$\lambda \in \Delta_{n,k} \text{ is a endpoint if and only if } \lambda \in \mathcal{A}_{n,k}.$$

We use a proof by contraposition for the first direction, so assume that $\lambda \notin \mathcal{A}_{n,k}$. Take $\lambda \in \Delta_{n,k} \setminus \mathcal{A}_{n,k}$. By Lemma A.3, we have λ such that it has $\#H$ entries taking the value $1/(n-k)$ and two entries: λ_{j_1} and $\lambda_{j_2} \in (0, 1/(n-k))$. Then, take $\epsilon > 0$ such that

$$\begin{aligned} 0 < \lambda_{j_1} - \epsilon &< \lambda_{j_1} + \epsilon < 1/(n-k), \text{ and} \\ 0 < \lambda_{j_2} - \epsilon &< \lambda_{j_2} + \epsilon < 1/(n-k). \end{aligned}$$

It follows $\lambda_1 = \lambda + \epsilon e_{j_1} - \epsilon e_{j_2}$ and $\lambda_2 = \lambda - \epsilon e_{j_1} + \epsilon e_{j_2}$ belong to $\Delta_{n,k}$. Furthermore, $\lambda_1 \neq \lambda_2$ and $\lambda = (\lambda_1 + \lambda_2)/2$.

As a consequence, λ is not an endpoint.

Conversely, let us fix $\lambda \in \mathcal{A}_{n,k}$. We claim that $\forall\, \lambda_1$ and $\lambda_2 \in \Delta_{n,k}$ and $\forall\, \theta \in (0,1)$

$$(1-\theta)\lambda_1 + \theta\lambda_2 = \lambda \text{ implies } \lambda_1 = \lambda_2 = \lambda.$$

In fact, $\lambda = \mathbf{1}_S/(n-k) \in \mathcal{A}_{n,k}$, then $\forall\, i \in S$,

$$\lambda_i = \frac{1}{n-k} = (1-\theta)\lambda_{1,i} + \theta\lambda_{2,i}.$$

Since λ_1 and $\lambda_2 \in \Delta_{n,k}$, it follows $\lambda_{1,i} \leq 1/(n-k)$ and $\lambda_{2,i} \leq 1/(n-k)$. Then

$$\begin{aligned} \frac{1}{n-k} &= (1-\theta)\lambda_{1,i} + \theta\lambda_{2,i} \\ &= (1-\theta)\frac{1}{n-k} + \theta\frac{1}{n-k} \\ &= \frac{1}{n-k}. \end{aligned}$$

Therefore, $\theta \neq 0, 1$ implies $\lambda_{1,i} = 1/(n-k)$ and $\lambda_{2,i} = 1/(n-k)$. It follows that $\forall\, i \in S, \lambda_{1,i} = \lambda_{2,i} = \lambda_i = 1/(n-k)$ and $\forall\, j \in [n] \setminus S, \lambda_j = 0 = (1-\theta)\lambda_{1,j} + \theta\lambda_{2,j}$. Hence, $\forall\, j \in [n] \setminus S, \lambda_j = 0 = (1-\theta)\lambda_{1,j} + \theta\lambda_{2,j}$ implies $\lambda_{1,j} = \lambda_{2,j} = 0$. □

The next Lemma show a bound to the probability that the event Norm(k) occurs.

Lemma A.4.

$$\mathbb{P}\left[\mathrm{Norm}(k)\right] \leq 1 - \frac{e^{-k}}{6}.$$

Proof. We first claim that for any $k \in \mathbb{N} \setminus \{0, 1\}$,

$$\mathbb{P}\left[\#\left\{i \in [n] \,:\, \|X_i\| \geq \sqrt{e\,\kappa}\left(\frac{n}{k}\right)^{\frac{1}{4}}\sqrt{\mathrm{tr}\left(\Sigma\right)}\right\} > k\right] \leq e^{-k-2}. \tag{A.3}$$

Indeed, by the hypothesis of fourth-moment and the Minkowski's inequality,

$$\left(\mathbb{E}\left[\|X_1\|^4\right]\right)^{\frac{1}{2}} \leq \kappa\,\mathrm{tr}\left(\Sigma\right).$$

Therefore,

$$\forall i \in [n], \lambda > 0 : \mathbb{P}\left[\|X_1\| \geq \sqrt{\kappa}\lambda\sqrt{\mathrm{tr}\left(\Sigma\right)}\right] \leq \frac{1}{\lambda^4}$$

Now, consider for any $\lambda > 0$ the probability that there exists a $S \subset [n]$ of cardinality $k+1$ such that $\|X_i\| \geq \lambda\sqrt{\kappa}\lambda\,\mathrm{tr}\left(\Sigma\right)$ for all $i \in S$

$$P(\lambda) := \mathbb{P}\left[\#\{i \in [n] : \|X_i\| \geq \sqrt{\kappa}\lambda\sqrt{\mathrm{tr}\left(\Sigma\right)}\} > k\right].$$

Using a union bound, the probability above, $P(\lambda)$, is upper bounded by

$$\sum_{|S|=k+1}\prod_{i\in S}\mathbb{P}\left[\|X_i\| \geq \sqrt{\kappa}\sqrt{\mathrm{tr}\left(\Sigma\right)}\right] \leq \binom{n}{k+1}\frac{1}{\lambda^{4(k+1)}} \leq \left(\frac{e\,n}{(k+1)\lambda^4}\right)^{k+1}.$$

Taking

$$\lambda^* := \sqrt{e}\left(\frac{n}{k}\right)^{1/4},$$

it follows

$$P(\lambda^*) \leq \left(\frac{ke^{-1}}{k+1}\right)^{k+1} \leq e^{-(k+2)}.$$

Then, the claim follows. Finally, the union bound for $k, 2k, \ldots$ gives

$$\mathbb{P}\left[\bigcup_{j=1}^{\infty}\left\{\#\left\{i \in [n] \,:\, \|X_i\| \geq \sqrt{e\,\kappa}\left(\frac{n}{jk}\right)^{\frac{1}{4}}\sqrt{\mathrm{tr}\left(\Sigma\right)}\right\} \leq jk\right\}\right] \leq \sum_{j\geq 1} e^{-jk-2} = \frac{e^{-k}}{e^2 - e^{2-k}}.$$

The RHS is bounded by $e^{-k}/6$. Thus, the lemma follows. □

Lemma A.5. *Under Assumption 3.1, for all $k \in \mathbb{N} \setminus \{0, 1\}$:*

$$\mathbb{P}\left[\bigcap_{M \in \mathcal{D}(\mathbb{R}^d)} \mathsf{Count}_k(M)\right] \geq 1 - \frac{e^{-\lceil \frac{kj}{92} \rceil}}{1 - e^{-\lceil \frac{kj}{92} \rceil}}$$

Proof. Our strategy will be to use the g_M construction in Definition 3.14 to pass to a counting event involving unit vectors. To start, we make the following claim.

Claim. *For any $M \in \mathcal{D}(\mathbb{R}^d)$ and $B > 0$,*

$$\sum_{i=1}^{n} \mathbf{1}_{\langle X_i, MX_i \rangle \geq B} \leq 24\, \mathbf{E}_{\mathsf{g}_M}\left\{\sum_{i=1}^{n} \mathbf{1}_{\langle X_i, v \rangle^2 \geq \frac{B}{2}}\right\}.$$

To see this, fix an index $i \in [n]$. Recall the spectral decomposition of $M = \sum_s \lambda_s \xi_s \xi_s^T$ as in Definition 3.14. Note that

$$\mathbf{E}_{\mathsf{g}_M}(\langle X_i, \mathsf{g}_M \rangle)^4 = \sum_{s=1}^{d} \lambda_s^2 \langle X_i, \xi_s \rangle^4 + 6 \sum_{1 \leq s_1, s_2 \leq d,\, s_1 \neq s_2} \lambda_{s_1} \lambda_{s_2} \langle X_i, \xi_{s_1} \rangle^2 \langle X_i, \xi_{s_2} \rangle^2.$$

In particular,

$$\mathbf{E}_{\mathsf{g}_M}(\langle X_i, \mathsf{g}_M \rangle)^4 \leq 6 \left(\sum_{s=1}^{d} \lambda_s \langle X_i, \xi_s \rangle^2\right)^2 = 6 \langle X_i, MX_i \rangle^2.$$

By the Paley–Zygmund inequality,

$$\mathbf{P}_{\mathsf{g}_M}\left(\langle X_i, g_M \rangle^2 \geq \frac{\langle X_i, MX_i \rangle}{2}\right) \geq \frac{1}{24}.$$

In particular, if $\langle X_i, MX_i \rangle \geq B$, then:

$$\mathbf{P}_{\mathsf{g}_M}\left(\langle X_i, \mathsf{g}_M \rangle^2 \geq \frac{B}{2}\right) \geq \mathbf{P}_{\mathsf{g}_M}\left(\langle X_i, \mathsf{g}_M \rangle \geq \sqrt{\frac{\langle X_i, MX_i \rangle}{2}}\right) \geq \frac{1}{24}.$$

Therefore, for each $i \in [n]$:

$$\mathbf{1}_{\langle X_i, MX_i \rangle \geq B} \leq 24 \mathbf{E}_{\mathsf{g}_M}[\mathbf{1}_{\langle X_i, v \rangle^2 \geq \frac{B}{2}}],$$

and the claim follows from summing over $i \in [n]$.

By the claim, we obtain that, $\forall\, M \in \mathcal{D}(\mathbb{R}^d)$ and $B > 0$:

$$\sum_{i=1}^{n} \mathbf{1}_{\langle X_i, MX_i \rangle \geq B} \leq 24 \left\{\sup_{||v||=1} \sum_{i=1}^{n} \mathbf{1}_{\langle X_i, v \rangle^2 \geq \frac{B}{2}}\right\}. \tag{A.4}$$

By choosing:

$$B = B(k) = 12\sqrt{2\,\kappa}\sqrt{\frac{n}{\lceil \frac{k}{92} \rceil}}\,\|\Sigma\|_{\mathrm{op}}\,.$$

it follows from Lemma 3.7 that

$$\sup_{\|v\|=1} \sum_{i=1}^{n} \mathbf{1}_{\langle X_i, v\rangle^2 \geq \frac{B}{2}} \leq \frac{k}{24}$$

holds with probability at least $1 - e^{-\lceil \frac{r}{92} \rceil}$. Then

$$\sup_{M \in \mathcal{D}(\mathbb{R}^d)} \sum_{i=1}^{n} \mathbf{1}_{\langle X_i, MX_i\rangle \geq B(k)} \leq k$$

with probability at least $1 - e^{-\lceil \frac{k}{92} \rceil}$. If we now consider $k, 2k, 3k \ldots$, and take a union bound, we conclude:

$$\begin{aligned} \mathbb{P}\left[\bigcup_{j=1}^{\infty}\left(\bigcup_{M \in \mathcal{D}(\mathbb{R}^d)} \{\#\,(i \in [n] \,:\, |\langle X_i, MX_i\rangle| \geq B(jk)) > jk\}\right)\right] &\leq \sum_{j \geq 1} e^{-\lceil \frac{k}{92} \rceil} \\ &\leq \frac{e^{-\lceil \frac{k}{92} \rceil}}{1 - e^{-\lceil \frac{k}{92} \rceil}}. \end{aligned}$$

This last event is precisely the complement of $\mathrm{Count}_r(M)$. □

Bibliography

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

[Audibert and Catoni, 2011] Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766 – 2794.

[Bickel, 1965] Bickel, P. J. (1965). On Some Robust Estimates of Location. *The Annals of Mathematical Statistics*, 36(3):847 – 858.

[Blair, 1985] Blair, C. (1985). Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). *SIAM Review*, 27(2):264–265.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

[Bubeck et al., 2013] Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

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

[Catoni, 2016] Catoni, O. (2016). Pac-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*.

[Catoni and Giulini, 2017] Catoni, O. and Giulini, I. (2017). Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*.

[Cherapanamjeri et al., 2019] Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019).

Fast mean estimation with sub-gaussian rates.

[Depersin and Lecué, 2021] Depersin, J. and Lecué, G. (2021). Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms.

[Depersin and Lecué, 2022] Depersin, J. and Lecué, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511 – 536.

[Devroye et al., 2016] Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

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

[Diakonikolas and Kane, 2019] Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics.

[Diakonikolas et al., 2020] Diakonikolas, I., Kane, D. M., and Pensia, A. (2020). Outlier robust mean estimation with subgaussian rates via stability.

[Hall, 1981] Hall, P. (1981). Large sample property of Jaeckel's adaptive trimmed mean. *Annals of the Institute of Statistical Mathematics*, 33(A):449–462.

[Hogg, 1974] Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, 69(348):909–923.

[Hopkins, 2020] Hopkins, S. B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193 – 1213.

[Hopkins and Li, 2019] Hopkins, S. B. and Li, J. (2019). How hard is robust mean estimation?

[Hopkins et al., 2021] Hopkins, S. B., Li, J., and Zhang, F. (2021). Robust and heavy-tailed mean estimation made simple, via regret minimization.

[Huber, 1964] Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

[Huber, 1972] Huber, P. J. (1972). The 1972 Wald Lecture Robust Statistics: A Review. *The Annals of Mathematical Statistics*, 43(4):1041 – 1067.

[Huber and Ronchetti, 2009] Huber, P. J. and Ronchetti, E. (2009). *Robust statistics (2nd edition)*. Wiley New York.

[Jaeckel, 1971] Jaeckel, L. A. (1971). Some Flexible Estimates of Location. *The Annals of Mathematical Statistics*, 42(5):1540 – 1552.

[Jana Jurecková, 1994] Jana Jurecková, Roger Koenker, A. H. W. (1994). Adaptive choice of trimming proportions. *Annals of the Institute of Statistical Mathematics*, 46(4):737–755.

[Joly et al., 2017] Joly, E., Lugosi, G., and Oliveira, R. I. (2017). On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11(1):440 – 451.

[Koltchinskii and Lounici, 2014] Koltchinskii, V. and Lounici, K. (2014). Concentration inequalities and moment bounds for sample covariance operators.

[Ledoux, 2001] Ledoux, M. (2001). The concentration of measure phenomenon. *American Mathematical Society*.

[Lee and Valiant, 2020] Lee, J. C. H. and Valiant, P. (2020). Optimal sub-gaussian mean estimation in $\mathbb{R}$.

[Lee, 2004] Lee, J.-Y. (2004). Adaptive choice of trimming proportions for location estimation of the mean. *Communications in Statistics - Simulation and Computation*, 33(3):673–684.

[Lepskii, 1991] Lepskii, O. V. (1991). On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466.

[Lerasle and Oliveira, 2011] Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. *arXiv: Statistics Theory*.

[Lounici, 2012] Lounici, K. (2012). High-dimensional covariance matrix estimation with missing observations.

[Lugosi and Mendelson, 2019a] Lugosi, G. and Mendelson, S. (2019a). Mean Estimation and Regression Under Heavy-Tailed Distributions: A survey. *Foundations of Computational Mathematics*, 19(19):1145 – 1190.

[Lugosi and Mendelson, 2019b] Lugosi, G. and Mendelson, S. (2019b). Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794.

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

[Léger and Romano, 1990] Léger, C. and Romano, J. P. (1990). Bootstrap adaptive estimation: The trimmed-mean example. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 18(4):297–314.

[Maller, 1988] Maller, R. A. (1988). Asymptotic Normality of Trimmed Means in Higher Dimensions. *The Annals of Probability*, 16(4):1608 – 1622.

[Mathé, 2006] Mathé, P. (2006). The lepskii principle revisited. *Inverse Problems*, 22(3):L11–L15.

[Mendelson and Paouris, 2014] Mendelson, S. and Paouris, G. (2014). On the singular values of random matrices. *Journal of the European Mathematical Society*, 16(4):823–834.

[Mendelson and Zhivotovskiy, 2019] Mendelson, S. and Zhivotovskiy, N. (2019). Robust covariance estimation under $l_4 - l_2$ norm equivalence.

[Minsker, 2015] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308 – 2335.

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

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

[Minsker and Wei, 2018] Minsker, S. and Wei, X. (2018). Robust modifications of u-statistics and applications to covariance estimation problems.

[Okamoto, 1958] Okamoto, M. (1958). Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10(1):73 – 101.

[Oliveira, 2016] Oliveira, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194.

[Orenstein, 2018] Orenstein, P. (2018). Robust importance sampling with adaptive winsorization. *arXiv preprint arXiv:1810.11130*. To appear in *Bernoulli*.

[Ostrovskii and Rudi, 2019] Ostrovskii, D. and Rudi, A. (2019). Affine invariant covariance estimation for heavy-tailed distributions.

[Rocke et al., 1982] Rocke, D. M., Downs, G. W., and Rocke, A. J. (1982). Are robust estimators really necessary? *Technometrics*, 24(2):95–101.

[Shi Jian, Zheng Zhongguo , 1996] Shi Jian, Zheng Zhongguo (1996). Choice of optimal trimming proportion by the random weighting method. *Acta Mathematica Sinica*, 12:326–336.

[Stigler, 1973] Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3):472–477.

[Stigler, 1974] Stigler, S. M. (1974). Linear Functions of Order Statistics with Smooth Weight Functions. *The Annals of Statistics*, 2(4):676 – 693.

[Stigler, 1977] Stigler, S. M. (1977). Do Robust Estimators Work with Real Data? *The Annals of Statistics*, 5(6):1055 – 1098.

[Stigler, 2010] Stigler, S. M. (2010). The changing history of robustness. *The American Statistician*, 64(4):277–281.

[Tikhomirov, 2017] Tikhomirov, K. (2017). Sample Covariance Matrices of Heavy-Tailed Distributions. *International Mathematics Research Notices*, 2018(20):6254–6289.

[Tropp, 2015] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

[Tukey, 1962] Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.

[Tukey and Mclaughlin, 1963] Tukey, J. W. and Mclaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample : Trimming/winsorization 1. *Sankhyā, The Indian Journal of Statistics*.

[Vershynin, 2011] Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.

[Wilfrid J. Dixon, Karen K. Yuen, 1974] Wilfrid J. Dixon, Karen K. Yuen (1974). Trimming and winsorization: A review. *Statistische Hefte*, 15.

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