Robust subgaussian estimation with VC-dimension

Jules Depersin*

University of Amsterdam.
Roetersstraat 11, 1018 WB Amsterdam.
E-mail: j.d.p.depersin@uva.nl

Abstract. Median-of-means (MOM) based procedures provide non-asymptotic and strong deviation bounds even when data are heavy tailed and/or corrupted. This work proposes a new general and systematic way to bound the excess risk for MOM estimators. The core technique is the use of the VC-dimension (instead of Rademacher complexity) to measure the statistical complexity. In particular, this allows one to give the first robust estimators for sparse estimation which achieves the so-called subgaussian rate, only assuming a finite second moment for the uncorrupted data.

By comparison, previous works using Rademacher complexities required a number of finite moments that grows logarithmically with the dimension. With this technique, we derive new robust subgaussian bounds for mean estimation in any norm.

Résumé. Les procédures basées sur la médiane des moyennes (MOM) fournissent des bornes non-asymptotiques et fortes, même lorsque les données ont des queues de distribution lourdes et/ou sont corrompues. Ce travail propose une nouvelle méthode générale et systématique pour limiter le risque des estimateurs MOM. La technique de base est l’utilisation de la dimension VC (au lieu de la complexité de Rademacher) pour mesurer la complexité statistique. Cela permet en particulier de trouver des estimateurs robustes pour l’estimation sparse qui atteignent le taux dit sous-gaussien en supposant seulement un second moment fini pour les données non corrompues.

En comparaison, les travaux précédents utilisant les complexités de Rademacher nécessitaient un nombre de moments finis de l’ordre du logarithme de la dimension, donc dépendant de la dimension. Grâce à cette technique, nous proposons de nouvelles bornes sous-gaussiennes robustes pour l’estimation de la moyenne dans n’importe quelle norme.

MSC2020 subject classifications: 62F35
Keywords: Robustness, heavy tailed data, sparse estimation.

Contents

1 Introduction .......................................................... 2
2 Setting and main tools ............................................... 4
  2.1 The contamination model ...................................... 4
  2.2 VC-dimension ..................................................... 5
  2.3 Median-of-means .................................................. 6
3 Mean estimation ....................................................... 6
4 Sparse setting and other estimation tasks ......................... 8
  4.1 Sparse mean estimation ........................................ 8
  4.2 Regression ....................................................... 9
  4.3 Low rank matrix estimation .................................... 10
  4.4 Covariance estimation ........................................... 11
5 Conclusion: concurrent work and discussion ..................... 11
6 Main Proofs ........................................................... 12
  6.1 Facts about the VC-dimension ................................ 12
  6.2 General methodology ........................................... 13

*J.D. thanks INSEE, ENSAE and CREST for their support.
1. Introduction

Robustness has been a classical topic in statistics since the work of Hampel [23, 24], Huber ([27, 28]) and Tukey [53]. In recent years, two main lines of research have been followed on the subject:

- **Robustness against outliers**: The classical setup in statistics is that the observations were generated independently from a given probabilistic model. This first kind of robustness aims at relaxing this strong assumption, as real datasets are typically exposed to some source of contamination. Robustness to outliers has first been described by Huber in [28]: in this context, the dataset is “close” to the ideal setup, but has been corrupted. The model that we deal with in this work, sometimes referred to as “adversarial contamination”, has been described and popularised by Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart in [17] and is now widely used in the robust statistics literature. The samples are generated from the following process: First, $N$ samples are drawn independently from some unknown distribution. An adversary is then allowed to look at the samples and arbitrarily corrupt an $\epsilon$-fraction of them. In this setup, not only can outliers be correlated to each other and to inliers, but inliers can also be correlated to one another (because the adversary can choose which original samples to keep and in doing so correlates the samples that they keeps, for instance only keeping the largest samples when they are real-valued). In recent years, designing outlier-robust estimators for high-dimensional settings has become a challenge in a number of applications (see [58] for a survey).

While robustness to contamination follows naturally from our analysis, it is not the main focus of this paper. Indeed, most of the problems we deal with in this paper have already been tackled from the point of view of robustness to outliers: sparse estimation in [4], linear regression in [19], covariance estimation in [11], etc. However, those papers do not address the problem of robustness to heavy tailed data: the inliers are assumed to follow normal distribution in all these works.

- **Robustness against heavy tailed data**: In this work, we are interested in estimators whose risks are controlled with very high probability without making either boundedness or Gaussian assumptions on the data (or any other strong concentration assumptions). Indeed, we want to avoid those assumptions that severely limit the applicability of the results.

Let us first give a simple example: For univariate mean estimation, if we are given $N$ independent realizations of a random variable with mean $\mu$ and variance $\sigma^2$, the empirical mean $\hat{\mu}$ satisfies $|\hat{\mu} - \mu| \leq \sigma/\sqrt{N\delta}$, with probability $1 - \delta$. It means that, for a given radius $R$, the empirical mean is outside the ball of center $\mu$ and of radius $R$, denoted $B(\mu, R)$, with probability bounded by $\frac{(\sigma/R)^2}{N}$. In general this bound is sharp for the empirical mean, as shown by Catoni in [8].

We informally call robust to heavy tail an estimator whose probability to be outside a ball centered at the estimand and of a fixed radius $R$ decreases exponentially with the number of observations $N$, even when the data is heavy tailed.

We notice that the empirical mean is thus not robust to heavy tailed data in the mean estimation problem, since this probability only decreases at a rate proportional to $1/N$. In order to get estimators robust to heavy tail, it is in general not sufficient to bound the expected loss of an estimator, since exponentially high probability guarantees cannot easily be derived from such bounds.

We now give another informal definition which we will show is related but distinct. We informally say that an estimator achieves a subgaussian rate when its probability to be outside a ball centered at the estimand and of a fixed radius $R$ achieves the same upper bound, up to multiplicative constants, as the empirical mean (or empirical risk minimizer in the case of regression) of a $N$-sample of i.i.d. gaussian variables, even when the data is heavy tailed.

An estimator is said to achieve a subgaussian rate when it is as good as if the data were Gaussian, even when it is not.
Let us show how these two definitions can be different by considering high-dimensional mean estimation. Let $X_1, \ldots, X_N \in \mathbb{R}^d$ denote $N$ i.i.d. random variables with mean $\mu$ and covariance matrix $\text{Id}$. The estimators $\hat{\mu}$ described by Minsker [46] or by Vidyasagar [26] both achieve, with probability greater than $1 - \delta$,

$$
\|\hat{\mu} - \mu\|^2 \lesssim \frac{d \log(1/\delta)}{N},
$$

(1)

where $\| \cdot \|$ is the canonical Euclidean norm on $\mathbb{R}^d$ and where $X \preceq Y$ means $X \leq CY$ for some universal constant $C$. For a fixed radius $R$, the probability that the estimators fail to be in $B(\mu, R)$ is bounded by $\exp(-Nc\sqrt{R}/d)$ for some universal constant $c$, thus the estimators are robust to heavy tails. However, they do not achieve a subgaussian rate. If $Z_1, Z_2, \ldots, Z_N$ are independent identically distributed Gaussian variables $N(\mu, \text{Id})$, it follows from Gaussian concentration inequality (see [33, Theorem 7.1] or [34, pages 56-57]) that with probability at least $1 - \delta$,

$$
\|\bar{Z}_N - \mu\|_2 = \sup_{\|v\|_2 \leq 1} \langle \bar{Z}_N - \mu, v \rangle \leq \mathbb{E} \sup_{\|v\|_2 \leq 1} \langle \bar{Z}_N - \mu, v \rangle + \sigma \sqrt{2 \log(1/\delta)},
$$

where $\sigma = \sup_{\|v\|_2 \leq 1} \sqrt{\mathbb{E} \langle \bar{Z}_N - \mu, v \rangle^2}$. We check that $\mathbb{E} \sup_{\|v\|_2 \leq 1} \langle \bar{Z}_N - \mu, v \rangle \leq \sqrt{d/N}$ and $\sigma = \sqrt{1/N}$, which leads to the rate in (2), which is called the subgaussian rate

$$
\|\bar{Z}_N - \mu\|_2^2 \lesssim \left( \frac{d}{N} + \frac{\log(1/\delta)}{N} \right).
$$

(2)

The main difference with (1) is that, in the subgaussian rate, two terms appear: a first that we will call the “complexity term” in which the dimension of the ambient space $d$ appears which measures the “complexity” of the problem, and a second which will call “deviation term” where the failure-dependent factor $\log(1/\delta)$ appears. What is remarkable about this rate is that the complexity and the deviation are “decoupled” in the sense that the dimension of the problem does not appear in the deviation term and $\delta$ does not appear in the complexity term. In (1), by contrast, the quantities are multiplied together instead of being added. This rate is shown to be deviation-minimax up to constant factors by Lugosi and Mendelson [39]. Recently, a seminal paper by the same authors [41] described the first estimator to reach the subgaussian rate only assuming finite second moment, soon followed by other works in that sense ([31], [15], [12], [25]).

Let us now consider $\mathcal{U}_s = \{ \sum_i \lambda_i e_i \mid \lambda_i \in \mathbb{R} \& \sum_i \lambda_i = s \leq s \}$, the set of $s$-sparse vectors, where $(e_i)_i$ is the canonical basis of $\mathbb{R}^d$. Let us assume that $\mu$ belongs to $\mathcal{U}_s$, an assumption that has come to be widely used in a number of settings following the work of Donoho [20]; see [62] for a survey. We see, taking again the Gaussian variables $Z_i$, and considering the projection $\bar{Z}_N$ of $Z_N$ on the set of sparse vectors (taking the $s$ largest coordinates in absolute value) that, with probability at least $1 - \delta$, (still from [33, Theorem 7.1])

$$
\|\bar{Z}_N - \mu\|_2 = \sup_{\|v\|_2 \leq 1, v \in \mathcal{U}_s} \langle \bar{Z}_N - \mu, v \rangle \\
\leq \mathbb{E} \sup_{\|v\|_2 \leq 1, v \in \mathcal{U}_s} \langle \bar{Z}_N - \mu, v \rangle + \sigma \sqrt{2 \log(1/\delta)} \\
\leq C \left( \frac{s \log(d/s)}{N} + \frac{\log(1/\delta)}{N} \right)^{1/2}.
$$

(3)

The subgaussian rate for the sparse mean estimation problem (3) is different from (2) while the “deviation term” remains unchanged. The “complexity term” (the one that does not depend on $\delta$) goes from $d$ to $s \log(d/s)$. This paper is the first to answer whether this rate can be reached only assuming a second order moment on the random variables.

In this work, we show that the analysis presented by Lugosi and Mendelson in [41], by Lecué and Lerasle in [31], [35] or [30], all based on the median-of-means principle and the use of Rademacher complexities, can be modified in order to achieve sub-gaussian rates for sparse or structured problems assuming only bounded moments of order two. The method developed in [35] or in [30] requires data to have at least $\log(d)$ finite moments (where $d$ is the dimension of the space) in order to exploit the sparsity of the problem and offers no guarantees without that requirement, to date is the best known method. We show that we can drop this condition by judiciously introducing the VC-dimension in the different proofs, and exploit the sparsity of the problem with only moments of order two. Classical approaches using local Rademacher complexities cannot achieve this type of subgaussian bounds under only a second moment assumption. Indeed, as shown
in the counter-example from Section 3.2.3 of [14], local Rademacher complexities may scale like $d^{1/8}$ whereas the correct Gaussian bound should be of the order of $\sqrt{\log d}$. Our VC-dimension based approach allows one to overcome this issue and to go beyond this $\log d$ subgaussian moments assumption that has appeared in all works on subgaussian estimation involving structured or sparse problems in high dimensions (for instance in the work of Lerasle [35]). We also show that this general technique can be easily applied to a wide range of problems and achieve state-of-the-art bounds for different estimation tasks such as:

- Regression, already studied by Lecué and Lerasle in [30] where our estimator’s rate matches that from [30], and sparse regression where our estimator’s rate is the first to match that from [30] with only moments of order two.
- Mean estimation with non-Euclidean norms, studied by Lugosi and Mendelson in [39], where our analysis gives a different rate that is better for some norms.
- Robust low-rank matrix estimation.

The main contribution of this work is the following: while the classical approach used so far provides, for most problems, a decoupling between a complexity term and a deviation term, we show that the complexity term it provides -in which local Rademacher complexity appears- does not capture the correct statistical complexity of high-dimensional problems in general. In particular, under low-dimensional structural assumptions it seems that the Rademacher complexity is not the right way to measure the complexity of the problem of robust estimation. We show another analysis that provides a new complexity term, which is better in some cases, but not in all. This work leaves open the question of the right statistical complexity of robust estimation problems, which is discussed in the conclusion.

This paper is not the first to introduce VC-dimension in robust estimation problems: this has been done by Chen, Gao, and Ren in [9] and by Gao in [22] for instance. In those two papers, estimation and regression with possible sparsity and outliers are also achieved with optimal rates, using VC-dimension techniques. The main differences lie in the model assumptions. For example, [9] estimates the center of symmetric distributions without moment assumption. In comparison, our estimators are for mean and covariance, and thus moment assumption is needed, but we do not need distributions to be symmetric.

The paper is organised as follows. In the next section, we present our setting and the main tools needed for the analysis. In Section 3, we present our analysis for the problem of mean estimation with any norm, and in Section 4.1, we present it in sparse or structured setups. The last section concludes.

2. Setting and main tools

In this section, we present the adversarial contamination model and we introduce the main tools needed for our analysis, namely the VC-dimension and the median-of-means paradigm.

2.1. The contamination model

The setting that we will use throughout the paper and that we call contamination model is the following:

**Setting 1.** Let $(Y_1, \ldots, Y_N)$ denote $N$ independent and identically distributed random vectors in $\mathbb{R}^d$. We want to estimate $\mathbb{E}(Y_1) = \mu$, assuming that $Y_1$ has a finite second moment. Let $\Sigma = \mathbb{E}((Y_1 - \mu)(Y_1 - \mu)^T)$ denote the unknown covariance matrix of $Y_1$.

The vectors $Y_1, \ldots, Y_N$ are not observed. Instead, this dataset may have been corrupted, and this corruption may be adversarial: there exists a (possibly random) set $O$ such that, for any $i \in O^c$, $X_i = Y_i$. We denote $\epsilon := |O|/N$ the outlier rate.

The observed dataset is $\{X_i : i = 1, \ldots, N\}$, and we want to estimate $\mu$.

Notice that there are no assumptions on the data $\{X_i, i \in O\}$. In particular these may be dependent on $\{Y_i : i = 1, \ldots, N\}$, and the $\{X_i : i \in O\}$ may have an arbitrary dependence structure. The set $O$ itself may be dependent on $\{Y_i : i = 1, \ldots, N\}$. This model is sometimes called the model of $\epsilon$-corruption or the strong contamination model, and is described in details by Diakonikolas and Kane [18].

We will now introduce the tools needed for our analysis.
2.2. VC-dimension

We start this part by recalling some basic facts about the VC-dimension that appear for instance in [56], Section 8.3.

Definition 1. Let \( \mathcal{F} \) be a set of Boolean functions on an euclidean space \( E \). We say that a finite set \( S \subset X \) is shattered by \( \mathcal{F} \) if, for every subset \( B \subset S \), there exists \( f \in \mathcal{F} \) such that \( S \cap f^{-1}(\{1\}) = B \). We call the VC-dimension of \( \mathcal{F} \) (and denote it as \( \text{VC}(\mathcal{F}) \)) the largest integer \( n \) that there exists a set \( S \) of cardinal \( n \) that is shattered by \( \mathcal{F} \).

We call VC-dimension of a set \( C \subset E \) and denote \( \text{VC}(C) \) the VC-dimension of the set of half-spaces generated by the vectors of \( C \):

\[
\text{VC}(C) = \text{VC}(\{x \in E \to 1_{\langle x,v \rangle \geq 0}, v \in C\}),
\]

where \( \langle x,y \rangle \) denote the scalar product of the set \( E \). This is an helpful abuse of notation. The following facts will be useful for the rest of this paper.

1. If \( F \) is a subspace of \( E \) of dimension \( k \), then \( \text{VC}(F) = k + 1 \). More generally, if \( F \) is a set of real-valued functions in a \( k \)-dimensional linear space, then \( \text{Pos}(F) = \{x \to 1_{f(x) \geq 0}, f \in F\} \) has VC-dimension \( k + 1 \) (see for instance [21], Theorem 7.2).

2. For a function \( g : \mathcal{X} \to \mathcal{X} \) and for a set \( \mathcal{F} \) of Boolean functions, if we denote \( \mathcal{F} \circ g = \{f \circ g \mid f \in \mathcal{F}\} \), then we have \( \text{VC}(\mathcal{F} \circ g) \leq \text{VC}(\mathcal{F}) \).

3. Sauer’s Lemma [51]: Let \( \mathcal{F} \) denote a set of functions with VC-dimension \( \nu \) and let \( S \) be a set of \( n \geq \nu \) points. Let \( \mathcal{F} * S = S \cap f^{-1}(\{1\}), f \in \mathcal{F} \)\), then

\[
\text{Card}(\mathcal{F} * S) \leq \left(\frac{en}{\nu}\right)^{\nu}.
\]

4. Let us denote, for any set \( C \subset E \), \( C - C = \{x - y \mid (x,y) \in C \times C\} \). For any \( r > 0 \), \( \text{VC}(\{x \in E \to 1_{\langle x,v \rangle \geq r}, v \in C\}) \leq \text{VC}(C - C) \lesssim \text{VC}(C) \), see Section 6 for a proof of this fact.

The following lemma is a straightforward extension of Theorem 6 in [1]. We present a proof in Section 6.

Lemma 1. Let \( \mathcal{F}_1, ..., \mathcal{F}_n \) denote \( n \) sets of boolean functions, each having VC-dimension \( \leq \nu \). Then,

\[
\text{VC}(\mathcal{F}_1 \cup \mathcal{F}_2 \cup \ldots \cup \mathcal{F}_n) \leq 4\nu + 2\log_2(n).
\]

When writing the set \( \mathcal{U}_s \) as a union of \( \binom{d}{s} \) \( s \)-dimensional subspaces, we find the following corollary that is the main tool we use to deal with sparse estimation.

Corollary 1. Fix \( v_1, ..., v_d \in \mathbb{R}^n \) and denote \( \mathcal{U}_s = \{\sum_i \lambda_i v_i \mid \lambda_i \in \mathbb{R} \& \sum_i 1_{\lambda_i \neq 0} \leq s\} \) the set of \( s \)-sparse vectors, then

\[
\text{VC}(\mathcal{U}_s) \leq 4s \log_2(ed/s).
\]

As a side remark, we note that [1] also shows that this bound is tight up to multiplicative constants whenever the set of vectors \( (v_1, ..., v_d) \) is linearly independent: there exists an absolute constant \( c \) such that \( \text{VC}(\mathcal{U}_s) \geq cs \log_2(ed/s) \).

Let us now recall a theorem that will be very useful in regression and covariance estimation. Let \( P = \{P_1, ..., P_m\} \) denote a set of multivariate polynomials. A sign assignment is an element \( w \) of \( \{+,-\}^m \). The sign assignment \( w \) is consistent with \( P \) if there exists \( x \in \mathbb{R}^n \) such that \( P_i(x) \geq 0 \iff w_i = + \).

Theorem 1 (Warren, [59]). Let \( P = \{P_1, ..., P_m\} \) denote a set of polynomials of degree at most \( \nu \) in \( n \) real variables with \( m > n \). The number of sign assignments consistent for \( P \) is then at most \( (4e\nu m/n)^n \).

We denote \( \mathbb{R}_0^m[X] \) the set of all polynomials of degree at most \( \nu \) in \( n \) real variables.

Corollary 2. Assume that the set of functions \( \mathcal{F} \) can be written \( \mathcal{F} = \{P \in \mathbb{R}_0^m[X] \to 1_{P(x) \geq 0}, x \in \mathbb{R}^n\} \), then \( \text{VC}(\mathcal{F}) \leq 2n \log_2(4e\nu) \).

The following example will be useful in some applications concerning matrix estimation (we note that this is not a new idea. A similar result can be found, for instance, in [61], Theorem 2).

Proposition 1. Let \( r \geq 0 \) and call \( \mathcal{M}_d^r(\mathbb{R}) \) the set of rank \( k \), symmetric, \( d \)-dimensional matrices. When applied to matrices, the notation \( \langle A, B \rangle \) denotes the Frobenius inner product: \( \langle A, B \rangle = \text{Tr}(AB^T) \).

Let \( \mathcal{F} = \{M \in \mathcal{M}_d(\mathbb{R}) \to 1_{\langle x,M \rangle \geq r}, x \in \mathcal{M}_d^r(\mathbb{R})\} \). Then \( \text{VC}(\mathcal{F}) = \text{VC}(\mathcal{M}_d^r(\mathbb{R})) \leq 2(d + 1)k \log_2(12e) \).
Proof. Any $X \in M^k_d(\mathbb{R})$ can be written $X = \sum_{i=1}^k \lambda_i x_i x_i^T$, with $(\lambda_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$. Besides, for any $M$, the function $(\lambda_i, x_i)_{i \leq k} \rightarrow (X, M) - r$ is a polynomial of degree 3 in $k(d + 1)$ variables. Hence, the result follows from Corollary 2.

2.3. Median-of-means

This work uses the median-of-means (MOM) approach which was introduced in [2, 29, 48] and has received a lot of attention recently in the statistical and machine learning communities [7, 16, 36, 46, 47]. This approach allows one to build estimators that are robust to outliers and heavy tailed data in various settings [2, 5, 29]. It can be defined as follows: we first randomly split the data into $K$ blocks $B_1, \ldots, B_K$ of equal size $m$ (if $K$ does not divide $N$, we just remove some data). We then compute the empirical mean within each block: for $k = 1, \ldots, K$,

$$\bar{X}_k = \frac{1}{m} \sum_{i \in B_k} X_i.$$ 

In the one-dimensional case, the final estimator is the median of the latter $K$ empirical means. This estimator has subgaussian deviations as shown by Devroye, Lerasle, Lugosi, and Oliveira in [16]. The extension of this result to higher dimensions is not trivial as several possible generalizations of the one dimensional median exist, see for instance [46] by Minsker.

Let us assume Setting 1. For any $k \in \{1, \ldots, K\}$, let $X_k := (X_i)_{i \in B_k}$ and $Y_k := (Y_i)_{i \in B_k}$. We start with a basic observation.

Remark 1. When $K \geq |O|$, there are at least $K - |O|$ blocks $B_k$ on which $X_k = Y_k$.

For instance, if $K \geq 4|O|$, then, there exist at least three quarters of the blocks $B_k$ where $X_k = Y_k$. We can now state the main lemma.

Lemma 2. Let $\mathcal{F}$ be a set of Boolean functions satisfying the following assumptions.

- For all $f \in \mathcal{F}$, $\mathbb{P} (f(Y_1) = 0) \geq 15/16$.
- $K \geq C(VC(\mathcal{F}) \vee |O|)$ where $C$ is a universal constant.

Then, with probability $\geq 1 - \exp(-K/128)$, for all $f \in \mathcal{F}$, there are at least $3K/4$ blocks $B_k$ on which $f(X_k) = 0$.

In words, if each property $f$ is true for one non-corrupted block with constant probability, (here 15/16 but it could be any fixed constant $\alpha \geq 1/2$) and if $K$ is large enough, then, with very high probability, all properties are “true for most of the blocks”. The Boolean functions that we will consider to construct estimators will measure whether the mean of the block is far from the true mean. For instance, for mean estimation, we take the set

$$\mathcal{F} = \{(x_i)_{i \leq m} \rightarrow 1(\pm \sum_{i \in O} x_i, v) \geq r_{K}, v \in V\}.$$ 

This result is an alternative the one Lugosi and Mendelson obtain in [39, Theorem 2] where the complexity is measured with a VC-dimension instead of the Rademacher complexity. We show below that this difference leads to substantial improvements in some examples such as sparse multivariate mean estimation compared with the bounds in [39]. The strength of this result is that it is uniform in $\mathcal{F}$ and gives an exponentially low failure probability, but its proof is quite simple. The proof of this result is given in Section 6.2.

Clearly, the fraction $3/4$ of the block is arbitrary in Lemma 2. In fact, up to some modifications of the constants, the same result holds for any fixed fraction $\alpha < 1$.

3. Mean estimation

We start with the mean estimation problem in $\mathbb{R}^d$ that illustrates our technique. We assume Setting 1, the goal is to estimate the mean $\mathbb{E}[Y]$ of a random vector $Y$ in $\mathbb{R}^d$ given a possibly corrupted dataset of i.i.d. copies of $Y$.

Let $\|\cdot\|$ denote a norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ denote its dual norm. Let $B$ denote the unit ball for the norm $\|\cdot\|$ and $B^*$ that for the norm $\|\cdot\|_*$. Let $B^*_0$ denote the set of extremal points of $B^*$: we recall that an extremal point of $B^*$ is a point in
Remark 4. Our aim is to show that VC type bounds are particularly efficient in structured cases, when Rademacher complexity fails.

In particular, for any $\delta \in [e^{-N/128}, 1/2]$, there exists an estimator $\mu_\delta$ such that, with probability greater than $1 - \delta$,

$$
\|\mu - \mu_\delta\| \lesssim \left\| \Sigma^{1/2} \right\| \left( \sqrt{\frac{\text{VC}(B_0^*)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon} \right).
$$

The 'outlier' term $\sqrt{\epsilon}$ is optimal (see the remarks after [10, Theorem 1.3]). The deviation term $\left\| \Sigma^{1/2} \right\| \sqrt{\log(1/\delta)/N}$ is the same as in the Gaussian concentration inequality, $\left\| \Sigma^{1/2} \right\|$ being the weak variance term. It is optimal as shown in [39], and can also be achieved by other estimators such as the multivariate trimmed mean estimator as proven in [42]. The difference with [39] is the complexity term, which here is $\left\| \Sigma^{1/2} \right\| \sqrt{\text{VC}(B_0^*)/N}$. Neither [39] nor this work build estimators achieving the true subgaussian rate in every case, where this complexity is $\mathbb{E}(\|G\|)/\sqrt{N}$, $G$ being a centered Gaussian vector with the same covariance as $Y$. For now, it is not known whether MOM estimators can or cannot achieve this rate in general, for all possible norms. However, as we will show, our rate matches the true subgaussian rate in some special cases (so does that from [39] for some other special cases).

Remark 2. The inequality $\text{VC}(B_0^*) \leq d + 1$ establishes a general bound on the complexity term.

The complexity term in [39], also found in [35, Chapter 4, Lemma 47] is $\mathbb{E}(\|\tilde{Y}\|)/N$ where $\tilde{Y} = \sum \epsilon_i(Y_i - \mu)$, $\epsilon_i$ being i.i.d. Rademacher variables. Here it is $\left\| \Sigma^{1/2} \right\| \sqrt{\text{VC}(B_0^*)/N}$. Which of them is best depends on the situation. For instance, when one wishes to estimate with respect to $\|\cdot\|_2$, the Euclidean norm on $\mathbb{R}^d$, $\mathbb{E}(\|\tilde{Y}\|)/N \simeq \sqrt{\text{Tr}(\Sigma)/N}$, while $\left\| \Sigma^{1/2} \right\| \sqrt{\text{VC}(B_0^*)/N} = \sqrt{\lambda_1 d/N}$, $\lambda_1$ being the largest eigenvalue of $\Sigma$, the former is better. In this example, the bound using the VC-dimension loses the dependence in the covariance structure. On the other hand, suppose that we want to estimate $\mu$ with respect to the sup norm $\|a\|_\infty = \max\{a_1, \ldots, a_n\}$ and assume that $\Sigma = \text{Id}$ for simplicity. Then $\|\text{Id}\| = 1$ and $\text{VC}(B_0^*) \lesssim \log(d)$ so

$$
\left\| \Sigma^{1/2} \right\| \sqrt{\text{VC}(B_0^*)/N} \simeq \sqrt{\log(d)/N}.
$$

On the other hand, if we only have two moments on the coordinates of $Y$, then the Rademacher bound $\mathbb{E}(\|\tilde{Y}\|_2)/N$ is of order $\sqrt{d/N}$ in general (to see that, take for $Y_1$ a random vector whose coordinates are independent, equal to $\sqrt{d/N}$ with probability $1/(dN)$ and 0 otherwise).

Remark 3. The analysis of Section 6.2 and in particular Lemma 5, shows that the estimator $\hat{\mu}_K$ achieves the bound $\left\| \Sigma^{1/2} \right\| \sqrt{K/N}$ when $K \geq C \vee |\mathcal{O}|$, where the complexity $C$ is the minimum between the VC-dimension $\text{VC}(B_0^*)$ and the Rademacher complexity $\mathbb{E}(\|\tilde{Y}\|^2/(N \left\| \Sigma^{1/2} \right\|))$. Therefore both our bounds and the bound of [39] hold simultaneously and we can always keep the “best complexity term” between VC and Rademacher complexity. As the main novelty here is the introduction of the VC-dimension, we do not recall this fact in each application. The interested reader can have in mind that, in most examples, the same result holds and the estimators have risk bounds lesser than both complexities. Our aim is to show that VC type bounds are particularly efficient in structured cases, when Rademacher complexity fails to achieve optimal bounds.

Remark 4. For the $\|\cdot\|_\infty$ norm, other estimators can achieve the correct rate $\sqrt{\log(d)/N}$ without resorting to VC-dimensions, for instance by using the one-dimensional estimator of Catoni [8] coordinate-wise. However this case is a good illustration of the limitations of the Rademacher classical analysis. As we will see in Section 4.3, in some more involved setups, the VC-dimension analysis is to this time the only one to achieve a sharp rate.
4. Sparse setting and other estimation tasks

This section shows that the methodology of Theorem 2 also applies to a great variety of estimation tasks. Let us start with the example of sparse mean estimation for the Euclidean norm.

4.1. Sparse mean estimation

For any $v_1, ..., v_d \in \mathbb{R}^n$, let $\mathcal{U}_s(v_1, ..., v_d) = \{ \sum_i \lambda_i v_i \mid \lambda_i \in \mathbb{R} \& \sum_i \lambda_i \neq 0 \leq s \}$ denote the set of $s$-sparse vectors over the dictionary $\{v_1, ..., v_d\}$. For this part, we fix the vectors $v_1, ..., v_d$ and we denote $\mathcal{U}_s = \mathcal{U}_s(v_1, ..., v_d)$. We consider Setting 1 and assume furthermore that $\mu$ belongs to $\mathcal{U}_s$. We denote $B_2$ the unit ball for the canonical Euclidean norm in $\mathbb{R}^n$, and propose the estimator

$$\hat{\mu}_K = \arg\min_{a \in \mathcal{U}_s} \max_{u \in B_2} \text{Med} \left\langle \bar{X}_k - a, u \right\rangle.$$

**Theorem 3.** There exists an absolute constant $C$ such that, if $K \geq C(s \log(d/s) \vee |\Omega|)$, then, with probability greater than $1 - \exp(-K/128)$,

$$\|\hat{\mu}_K - \mu\|_2 \leq 8 \sqrt{\lambda_1(\Sigma) K N}.$$

Here, $\lambda_1(\Sigma)$ is the largest eigenvalue of $\Sigma$.

The conclusion of Theorem 3 can be written as follows. For any $\delta \in \left[ e^{-N/128}, 1/2 \right]$, there exists an estimator $\mu_\delta$ such that, with probability greater than $1 - \delta$,

$$\|\mu - \mu_\delta\|_2 \lesssim \sqrt{\lambda_1(\Sigma)} \left( \sqrt{\frac{s \log(d/s)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon} \right).$$

The complexity term (with “complexity” $s \log(d/s)$) is once again decoupled from the deviation term, which is not the case in works such as the one of Hsu and Sabato [26] where those two terms are multiplied together. The complexity term exhibited here is not optimal in general because it does not depend on the structure of $\Sigma$: for instance if the eigenvalues of $\Sigma$ decrease extremely rapidly, one could have $\sqrt{\text{Tr}(\Sigma)} < \sqrt{\lambda_1(\Sigma)} \sqrt{s \log(d/s)}$, and the complexity term of the true subgaussian rate, obtained with the Gaussian concentration inequality, would be even smaller than $\sqrt{\text{Tr}(\Sigma) N}$, since:

$$\mathbb{E} \sup_{\|v\|_2 \leq 1, v \in \mathcal{U}_s} \langle \bar{Z}_N - \mu, v \rangle \leq \mathbb{E} \sup_{\|v\|_2 \leq 1} \langle \bar{Z}_N - \mu, v \rangle \leq \sqrt{\frac{\text{Tr}(\Sigma)}{N}},$$

taking Gaussian variable $(Z_i)_i$ with covariance $\Sigma$. However, our complexity term is interesting for two principal reasons:

- This is the first sparsity dependent bound that holds without higher moment conditions than the $L_2$ ones. In contrast, [35] or [30] from Lecué and Lerasle need to assume the existence of $\log(d)$ subgaussian moments in order to make the sparsity appear, and offer no guarantees without that requirement.

- It is theoretically optimal when $\Sigma = \lambda \text{Id}$, since in that case, the complexity term from the Gaussian concentration inequality is proportional to $\sqrt{\lambda \frac{s \log(d/s)}{N}}$.

**Remark 5.** This theorem can be obtained without resorting to VC-dimensions, simply by using the analysis of Lugosi and Mendelson in [41] on the $\binom{d}{s}$ subspaces of $\mathcal{U}_s$ and a union bound. In other words, we can obtain similar rates with the standard median-of-means approach and simple manipulations. However, in other cases where the set considered is not a simple union of subspaces, for instance in Section 4.3, such a trick is no longer possible.
4.2. Regression

In this section, we consider the standard linear regression setting where data are pairs \((Y_i, V_i)\), \(i \in \mathbb{R}^d \times \mathbb{R}\) and we look for the best linear combination of the coordinates of \(Y_i\) to predict \(V_i\) using the squared loss \(l(\beta) = \mathbb{E}(V_i - \langle \beta, Y_i \rangle)^2\). In other words, we look for \(\beta^*\) defined as follows. Given \(S \subset \mathbb{R}^d\) (in this paper, we will only study \(S = \mathbb{R}^d\) or \(S = \mathcal{U}_k\)),

\[
\beta^* = \arg\min_{\beta \in S} l(\beta) = \arg\min_{\beta \in S} \mathbb{E}(V_i - \langle \beta, Y_i \rangle)^2.
\]

As in the previous section, the observed dataset \((X_i, Z_i)\), \(i \in \mathbb{R}^d \times \mathbb{R}\) is a corrupted version of the i.i.d. dataset \((Y_i, V_i)\), \(i \in \{1, \ldots, N\}\) in a possibly adversarial way. The assumptions made on good data \((Y_i, V_i)\) are gathered in the following setting: (see also [35] or [3]).

**Setting 2.** There exists a (possibly random) set \(O\) such that, for any \(i \in O^c\), \((X_i, Z_i) = (Y_i, V_i)\), where \((Y_i, V_i)\) are independent identically distributed observations in \(\mathbb{R}^d \times \mathbb{R}\). Let \(\xi_i = V_i - \langle \beta^*, Y_i \rangle\). We make four assumptions:

1. \(Y_1\) has a finite second moment and we write its \(L^2\)-moments matrix \(\Sigma = \mathbb{E}(Y_1Y_1^T)\). Let also \(B_\Sigma = \{x \in S - S \mid \langle x, \Sigma x \rangle \leq 1\}\) be the ellipsoid associated with this \(L_2\) structure.
2. Let \(\sigma^2 := \sup_{u \in B_\Sigma} \mathbb{E}(\xi_i^2 \langle u, Y_i \rangle^2)\) and assume that \(\sigma^2 < \infty\).
3. There exists an universal constant \(\gamma\) such that, for all \(u \in S - S\), \(\mathbb{E}(\|u, Y_i\|) \geq \gamma \sqrt{\mathbb{E}(\langle Y_1, Y_1 \rangle)}\).
4. \(\mathbb{E}(\xi_1) = 0\)

Condition 2 is implied by Assumptions 3.5 and 3.7 in [3], the same assumption is made in [35].

Condition 3 is called the “small ball hypothesis”, it is described in detail in [44] or in [32] for instance. It is implied by Condition 3.5 in [3], it is stated similarly in [35].

Condition 4 is always true when \(S = \mathbb{R}^d\). In other cases, it is still true in a number of applications, for instance when the noise \(\xi\) and \(Y\) are independent.

The two last conditions may seem exotic, we refer to [3, Section 3] for detailed discussions and examples where these are satisfied. For the moment, we may emphasize that they involve only first and second moment conditions on \(\xi_1\) and \(\langle u, Y_1 \rangle\).

Let \(Q_{1/4}^K\) denote the first quartile over \(k \leq K\): for any sequence \(x_1, \ldots, x_K \in \mathbb{R}\), if we denote \(x_1^*, \ldots, x_K^*\) the corresponding increasingly ordered sequence, \(Q_{1/4}^K x_K = x_K^*\). Let \(\hat{B}_\Sigma = \{u \in S - S \mid Q_{1/4}^K \sum_{i \in B_k} |\langle u, X_i \rangle|^2 \leq 1\}\). Our estimator is the following:

\[
\hat{\beta} = \arg\max_{a \in S} \text{Med} \sum_{u \in B_\Sigma} \sum_{i \in B_k} (Z_i - \langle a, X_i \rangle) \langle u, X_i \rangle.
\]

This new estimator satisfies the following result.

**Theorem 4.** There exists an absolute constant \(C\) such that the following holds. Let \(S = \mathbb{R}^d\) or \(\mathcal{U}_K\) and let \(N \gamma^2 / 64 \geq K \geq C(VC(S - S) \vee |O|)\). Then, with probability \(\geq 1 - \exp(-K/128)\),

\[
|\langle \hat{\beta} - \beta^*, \Sigma(\hat{\beta} - \beta^*) \rangle| \leq 128 \sigma^2 \frac{K}{N \gamma^2}.
\]

For all \(\beta\),

\[
l(\beta) = l(\beta^*) + 2\mathbb{E}(\xi_i \langle \beta - \beta^*, Y_i \rangle) + (\beta - \beta^*) \Sigma(\beta - \beta^*) \leq l(\beta^*) + (\beta - \beta^*) \Sigma(\beta - \beta^*),
\]

where we recall that \(l(\beta) = \mathbb{E}(V_i - \langle \beta, Y_i \rangle)^2\). So, if \(r = \sqrt{128} \sigma \sqrt{\frac{K}{N}}\), then

\[
l(\hat{\beta}) - l(\beta^*) \leq r^2 / \gamma^4.
\]

The conclusion of Theorem 4 can be written as follows: for any \(\delta \in [e^{-N/128}, 1/2]\), there exists an estimator \(\mu_\delta\) such that, with probability greater than \(1 - \delta\),

\[
|\langle \hat{\beta} - \beta^*, \Sigma(\hat{\beta} - \beta^*) \rangle| \leq \sigma \left(\sqrt{\frac{VC(S)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon}\right).
\]
Once again we notice the decoupling between complexity and deviation. This result is interesting for several reasons, mainly because it is the first that gives a bound holding with exponential probability, that holds without assuming more than 2 moments on the design $Y_i$, even in the sparse setting. By comparison works such as [30] of Lecué and Lerasle or [40] of Lugosi and Mendelson assume that at least $\log(d)$ subgaussian moments exist to achieve such a rate and offer no guarantees without that requirement and are the best to date.

4.3. Low rank matrix estimation

We now turn to the problem of matrix estimation, presented for instance in [52, 63, 64].

Let $\mathcal{M}_d^k(\mathbb{R})$ be the set of square matrices of size $d$, $\mathcal{M}_d^k(\mathbb{R})$ be the set of rank $k$, symmetric, $d$-dimensional matrices, let $\langle \cdot \rangle_F$ denote the usual scalar product on $\mathcal{M}_d(\mathbb{R})$ and $\|\cdot\|_F$ denote the associated norm, sometime called Frobenius norm.

We have observations in $\mathcal{M}_d(\mathbb{R})$ and we try to estimate their mean, assuming a low-ranked structure. The setting is the following: we have, as in setting 1, $N$ (corrupted) observations $(X_i)_i \in \mathcal{M}_d(\mathbb{R})$ of original $(Y_i)$ satisfying $\mathbb{E}(Y_i) = B$ and we try to estimate the (non necessarily low rank) mean $B$. We will assume for simplicity that $\mathbb{E}((Y_i^j - B^j)(Y_i^k - B^k)) = \sigma^2 1((i,j) = (k,l))$. We try to estimate $B$ with respect to the following norm:

$$\|A\|_r = \sup_{U \in \mathcal{M}_d^r(\mathbb{R}), \|U\|_F = 1} \langle U, A \rangle_F,$$

where we recall that $\mathcal{M}_d^r(\mathbb{R})$ is the set of symmetric, $d$-dimensional matrices of rank $k$. We will try to show that this structure can not be estimated through the analysis based on Rademacher complexity: we give this example to illustrate the benefit of our approach.

$$\hat{B}_k = \arg\min_{M \in \mathcal{M}_d(\mathbb{R})} \sup_{U \in \mathcal{M}_d^r(\mathbb{R}), \|U\|_F = 1} \operatorname{Med}_k \left( \frac{1}{m} \sum_{i=1}^m \langle U, X_i - M \rangle_F \right),$$

**Theorem 5.** There exists an absolute constant $C$ such that, if $K \geq C(kd \vee |O|)$, then, with probability greater than $1 - \exp(-K/128)$,

$$\|\hat{B}_k - B\|_r \leq 8 \sqrt{\frac{\sigma K}{N}}.$$

The conclusion of Theorem 5 can be written as follows. For any $\delta \in [e^{-N}/128, 1/2]$, there exists an estimator $\hat{B}_\delta$ such that, with probability greater than $1 - \delta$,

$$\|\hat{B}_\delta - B\|_r \leq \sigma \left( \sqrt{\frac{kd}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon} \right).$$

While in [63] authors need $X_i$ to be i.i.d. gaussian variables, we only need finite second-order moments on those variables.

At this point, we want to show that those results could not be obtained using standard analysis with Rademacher complexity ([35, 39]). Indeed this analysis would give a bound of order $\mathbb{E}(\max_{U \in \mathcal{M}_d^r(\mathbb{R}), \|U\|_F = 1} \langle \hat{Y}, U \rangle) / N$ (with $\hat{Y} = \sum \epsilon_i (Y_i - B)$, $\epsilon_i$ being i.i.d. Rademacher variables) instead of $\sigma \sqrt{kd}$, as mentioned in Section 3. Let us show a case where those two quantities have different behaviours.

We take for instance $N = 1$ and, independent identically distributed $(Y_i^{kl})_{1 \leq k,l \leq d}$ so that

$$Y_i^{kl} = \begin{cases} +\sigma d \text{ with probability } 1/(2d^2) \\ -\sigma d \text{ with probability } 1/(2d^2) \\ 0 \text{ with probability } 1 - 1/d^2, \end{cases}$$

If one of the $Y_i^{kl}$ is non-zero, then $\max_{U \in \mathcal{M}_d^r(\mathbb{R}), \|U\|_F = 1} \langle Y_i, U \rangle / N \geq \sigma d$. Given that $\mathbb{P}(\forall (k,l), Y_i^{kl} = 0) = (1 - 1/d^2)^d < e^{-1}$, we find that

$$\mathbb{E} \left( \max_{U \in \mathcal{M}_d^r(\mathbb{R}), \|U\|_F = 1} \langle \hat{Y}, U \rangle \right) \geq \sigma d (1 - e^{-1}).$$
In this case, the quantity we get from the Rademacher analysis scales as $d$ whereas our bound scales as $\sqrt{kd}$. Moreover, use of union bounds (as in Section 4.1, Remark 4 for the sparse case) is not possible here because $M^2_o r$ is not an union of linear subspaces.

4.4. Covariance estimation

This section examines the problem of robust covariance estimation. Consider Setting 1, and assume that $\mu$ is known, fixed to 0 without loss of generality. We want to estimate $\Sigma$. This problem has a number of applications: the bounds we present can for instance easily be transposed (with the Davis–Kahan Theorem) to the problem of robust Principal Component Analysis. It has already been studied in [60], or [26], but these estimators do not exhibit any decoupling between complexity and deviation. In [45], the authors propose a robust estimator for covariance using the MOM method, and get the optimal complexity-deviation decoupling. They also give interesting comments and insights about this estimation problem. However, they do not study the problem of low rank estimation that we present here.

For any matrix $A$, define its spectral norm by

$$\|A\| = \sup_x \frac{\|Ax\|_2}{\|x\|_2}.$$  

Let $\text{Sym}(d)$ denote the set of $d$ dimensional symmetric positive matrices. Assume that

$$\sigma^2 = \sup_{u \in B_2} \mathbb{E} \left( \langle u, (\Sigma - Y_1 Y_1^T) u \rangle^2 \right) < \infty.$$  

This quantity is sometimes called the weak variance of a random matrix [45]. While this is a condition on the fourth moment of the distribution, in this condition the fourth moment is not compared to the second moment of the distribution as in $L_4 - L_2$ norm equivalence condition, see for instance Def 1.5 in [45]. Our estimator is defined as follows:

$$\hat{\Sigma} = \arg\min_{M \in \text{Sym}(d)} \sup_{\|u\|_2 = 1} \text{Med}_k \langle u, (\frac{1}{m} \sum X_i X_i^T - M) u \rangle.$$  

It satisfies the following bound.

**Theorem 6.** There exists an absolute constant $C$ such that, if $K \geq C(d \vee |O|)$, then, with probability greater than $1 - \exp(-K/128)$,

$$\|\hat{\Sigma}_K - \Sigma\| \leq 8 \sigma \sqrt{\frac{K}{N}}.$$  

**Corollary 3.** Assume that $R = \sup_u \sqrt{\mathbb{E} \langle u, Y \rangle^4 / \mathbb{E} \langle u, Y \rangle^2} < \infty$, then, for $K \geq C(d \vee |O|)$

$$\|\hat{\Sigma}_K - \Sigma\| \leq 8 R \|\Sigma\| \sqrt{\frac{K}{N}}.$$  

The “bounded kurtosis assumption” $R < \infty$ similarly appears in [26]. In [26], the estimator achieves a bound of order $r(\Sigma) \|\Sigma\| \sqrt{K/N}$ where $r(\Sigma)$ is the effective rank of the covariance matrix; once again, the complexity $r(\Sigma) \|\Sigma\|$ is multiplied by the deviation term $K \propto \log(1/\delta)$ in this case, while here they are decoupled: the dimension does not multiply $K$ in our bound. In [45], the authors give a better rate (because they only need $K$ to be larger than $r(\Sigma)$ instead of $d$ for the bound to hold) but in their proof they use the $L_4 - L_2$ norm equivalence, which is a hypothesis we do not need to state Theorem 6 (even if we state Corollary 3 in term of this $L_4 - L_2$ norm equivalence).

5. Conclusion: concurrent work and discussion

This work is not the first to deal with robust estimation: many results and algorithms have been developed over the past few years for sparse estimation in the presence of outliers (see for instance the work of Li [37]) but most of these works assume that non-corrupted data are Gaussian. Chen, Gao and Ren [9], for example, already address mean and covariance
estimation using extensions of the Tukey-depth (and using the VC-dimension), but their methods rely on informative data being Gaussian.

Robustness to heavy tailed data has also been studied in various works (see [38] for a survey of recent developments). We already mentioned two articles that this work tries to complete and improve: [39] for mean estimation under any norm and [30] for sparse regression. Though the techniques involved are similar, this work illustrates that using VC-dimension can drastically improve risk bounds in various applications, in particular in the sparse setting.

Concurrent work: After the initial submission of this manuscript, we became aware of two concurrent works by Prasad, Balakrishnan and Ravikumar; [49] and [50]. Authors use an approach based on the 1/2-cover of the unit sphere to deal with mean and sparse-mean estimation for the euclidean norm ([49]), and with covariance estimation for spectral norms ([50]). They get close-to-optimal bounds for those two problems, with a remaining extra-logarithmic term. They do not tackle mean estimation in any norm, regression or low-rank covariance estimation.

There are still many interesting open questions. The different applications of Lemma 2 studied here show that neither the complexity term derived with VC-dimension techniques nor the one derived from the classical analysis are optimal in general. For example, for mean estimation in Euclidean norm, the complexity term reached by our estimator is proportional to \( \sqrt{\lambda_1(\Sigma) d / N} \), where the best rate would be \( \sqrt{\text{Tr}(\Sigma) / N} \). The quantity that is crucial in all the studies is

\[
\mathbb{E} \left( \sup \left\{ \sum_{k=0}^{K} f(Y_k) - K \mathbb{E}(f(Y_k)) \right\} \right)
\]

where \( f \) are Boolean functions. In mean estimation for instance,

\[
\mathbb{E} \left( \sup_{v \in V} \sum_{k=0}^{K} 1(\bar{Y}_k - \mu, v) \geq r - K \mathbb{E}(1(\bar{Y}_k - \mu, v) \geq r) \right)
\]

is the important quantity. Bounding this quantity using the VC-dimension of \( V \) yields a bound that does not depend on the covariance of \( \bar{Y} \). On the other hand, bounding that quantity by the Rademacher complexity of the \( Y_i \) (as in [30], [39] or here in Part 6.2) stating that

\[
\mathbb{E}(\sup_{v \in V} 1(\bar{Y}_k - \mu, v) \geq r - K \mathbb{E}(1(\bar{Y}_k - \mu, v))) < K \frac{\mathbb{E}(\|\bar{Y}\|)}{r \sqrt{N}}
\]

does not exploit the boundedness of the indicator function and thus needs an unnecessary stronger assumptions on data to achieve the right bounds. The ideal would be to reconcile both ideas, and to find a suitable in-between that would take into account both the boundedness and the dependency in the covariance structure.

The last point we make is about computational issues. The estimators presented can not be implemented as is. Nevertheless, encouraging recent works have shown that “relaxed” tractable estimators can be derived from this kind of work. For instance the pioneering work of [25], followed by [15] and [12], derived tractable estimators, computable in polynomial times, from the work of [41]. Even more recently, some new tractable estimators for regression and covariance estimation with heavy tailed data have emerged in [13]. We can hope for this work to be made tractable as well, which seems to be quite a challenge.

6. Main Proofs

6.1. Facts about the VC-dimension

For any Euclidean space \( E \), and any \( C \in E \)

**Lemma 3.** \( \text{VC}(\{ x \in E \rightarrow 1_{(x,v) \geq r}, v \in C \}) \leq \text{VC}(C - C) \leq c_0 \text{VC}(C) \) where \( c_0 \) is universal constant.

**Proof.** Assume that a set \( x_1, ..., x_d \in E \) is shattered by \( F = \{ x \in E \rightarrow 1_{(x,v) \geq r}, v \in C \} \). Then, for any \( I \subset \{1, 2, ..., d\} \), there is a vector \( v_1 \) so that \( \langle v_1, x_i \rangle \geq r \) if and only if \( i \in I \). There is a vector \( v_2 \) so that \( \langle v_2, x_i \rangle < r \) if and only if \( i \in I \). Then we have \( \langle v_1 - v_2, x_i \rangle \geq 0 \) if and only if \( i \in I \), so \( \{ x \in E \rightarrow 1_{(x,v) \geq 0}, v \in C-C \} \) shatters \( x_1, ..., x_d \), and
We want to prove that, with probability 

We begin by proving Lemma 2:

If \( C \geq 16, K \geq 16|O| \) and it is sufficient to show that \( \sup f f(X_k) \leq K/4 \).

By the bounded difference inequality [6, Theorem 6.2], with probability \( \geq 1 - \exp(-K/128) \), \( D \leq K/16 \).

For the magnitude term, we write

By hypothesis, \( \sup f K\mathbb{E}(f(Y_k)) \leq K/16 \). We just have to use a classical result of the Vapnik-Chervonenkis theory, either in the version of [56, Theorem 8.3.23], or of [55, Corollary 7.18]. There exists a universal constant \( C' \) such that

Hence, if \( K \geq 256 C'^2 \mathbb{VC}(F) \),

Putting everything together, we have the following. If \( C \geq 256 C'^2 \), with probability \( \geq 1 - \exp(-K/128) \), \( \sup f \sum_{k=0}^{K} f(Y_k) \leq K/16 + K/16 + K/16 \). Therefore, by Remark 1, for all \( f \in F \)

\[ \sum_{k=0}^{K} f(X_k) \leq K/4. \]
We state a technical lemma that appears in most proofs. Let $g$ be any measurable function $\mathbb{R}^d \to E$ so that $\mathbb{E}(g(Y_1))$ exists. We take
\[ \hat{a} = \text{argmin}_{a \in U} \text{Med} \left( \frac{1}{m} \sum_{i \in B_k} g(X_i) - a, v \right) \]
where $U, V$ are any sets of $E$. We have:

**Lemma 4.** If $K \geq C(\text{VC}(V) \lor |\mathcal{O}|)$ and if $\mathbb{E}(g(Y_1)) \in U$, then, with probability $\geq 1 - \exp(-K/128)$,
\[ \max_{v \in V} \mathbb{E}(g(Y_1)) - \hat{a}, v \] \[ \leq 8 \sup_{u \in V} \mathbb{E} \left( (g(Y_1) - \mathbb{E}(g(Y_1)), u)^2 \right)^{1/2} \sqrt{\frac{K}{N}} \]
where $C$ is a universal constant.

**Proof.** Let $K \geq C(\text{VC}(F) \lor |\mathcal{O}|)$ with $C$ the universal constant from Lemma 2, let $\bar{g} = \mathbb{E}(g(Y_1))$ and let
\[ r_K = 4 \sup_{u \in V} \mathbb{E} \left( (g(Y_1) - \bar{g}, u)^2 \right)^{1/2} \sqrt{\frac{K}{N}}. \]

Let $\mathcal{F} = \{ (x_i)_{i=1}^m \to 1(\sum_i g(x_i) - \mathbb{E}(g(Y_1)), v) \geq r_K, v \in V \}$. The function $f \in \mathcal{F}$ are compositions of the function $x \to \frac{1}{m} \sum_{i=1}^m g(x_i) - \mathbb{E}(g(Y_1))$ and of the functions $x \to 1(x,v) \geq r_K$ for $v \in V$. The VC-dimension of the set of these compositions is smaller than the VC-dimension of the set of indicator functions indexed by $V$, as recalled in the basic fact 2 at the beginning of Section 2.2. We just use fact 3 to remove the $r_K$ and we obtain $\text{VC}(\mathcal{F}) \leq c_0 \text{VC}(V)$ for some constant $c_0$.

By Markov’s inequality, for any $v \in V$,
\[ \mathbb{P}(|\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v| \geq r_K) \leq \frac{\mathbb{E} \left( \sum_{i \in B_k} (g(Y_i) - \bar{g}, u)^2 \right)^2}{m^2 r_K^2} \leq \frac{1}{16}. \]

By Lemma 2, applied with $\mathcal{F}$, the following event $\mathcal{E}$ has probability $\mathbb{P}(\mathcal{E}) \geq 1 - \exp(-K/128)$:
\[ \sup_{v \in V} \text{Med} |\frac{1}{m} \sum_{i \in B_k} g(X_i) - \bar{g}, v| \leq r_K. \]

For any $a \in U$ if there exists $v^* \in V$ such that $\langle \bar{g} - a, v^* \rangle > 2r_K$, then, on $\mathcal{E}$
\[ \text{Med} \langle \frac{1}{m} \sum_{i \in B_k} g(X_i) - a, v^* \rangle = \langle \bar{g} - a, v^* \rangle + \text{Med} \langle \frac{1}{m} \sum_{i \in B_k} g(X_i) - \bar{g}, v^* \rangle > r_K \geq \max_{v \in V} \text{Med} \langle \frac{1}{m} \sum_{i \in B_k} g(X_i) - \bar{g}, v \rangle. \]

Therefore $a \neq \hat{a}$. As this holds for any $a \in U$ such that $\sup_{v \in V} \langle \bar{g} - a, v \rangle > 2r_K$, it follows that, on $\mathcal{E}$,
\[ \sup_{v \in V} \langle \bar{g} - \hat{a}, v \rangle \leq 2r_K. \]

We can give a somewhat improved version of that lemma: let us denote
\[ \mathcal{R}(g, V) = \frac{1}{\sqrt{N}} \mathbb{E}(\sup_{v \in V} \left( \sum_i \epsilon_i g(Y_i), v \right)), \quad \sigma^2 = \sup_{u \in V} \left( \langle g(Y_1) - \mathbb{E}(g(Y_1)), u \rangle^2 \right). \]

\( \mathcal{R} \) is the Rademacher complexity associated with a given problem. The following lemma shows that we can take the best term between the one given by a rescaled Rademacher complexity and the one given by the VC-dimension.

**Lemma 5.** If \( \mathcal{K} \geq C((\text{VC}(V) \wedge (\mathcal{R}(g, V)/\sigma)^2) \vee |\mathcal{O}|) \) and if \( \mathbb{E}(g(Y_1)) \in U \), then, with probability \( \geq 1 - \exp(-K/128) \),

\[ \max_{v \in V} \mathbb{E}(g(Y_1)) - \hat{a}, v \leq 16\sigma \sqrt{\frac{K}{N}}, \]

where \( C \) is a universal constant.

**Proof.** We know that this holds when \( \mathcal{K} \geq C(\text{VC}(V) \vee |\mathcal{O}|) \).

If \( \mathcal{K} \geq C(\mathcal{R}(g, V)/\sigma)^2 \vee |\mathcal{O}| \), we only need to prove that, for \( r_K = 8\sigma \sqrt{K/N} \)

\[ \sup_{v \in V} \sum_k 1(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \mathbb{E}(g(Y_i)), v) \geq r_K \leq K/2 \]

and then we follow the path of the previous proof.

We do this in a way that can be found, for instance in [15] or the supplementary material of [43].

As \( K \geq 4|\mathcal{O}| \), we only need to show that

\[ \sup_{v \in V} \sum_k 1(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \mathbb{E}(g(Y_i)), v) \geq r_K \leq K/4. \]

We define \( \phi(t) = 0 \) if \( t \leq 1/2 \), \( \phi(t) = 2(t - 1/2) \) if \( 1/2 \leq t \leq 1 \) and \( \phi(t) = 1 \) if \( t \geq 1 \). We have \( 1(t \geq 1) \leq \phi(t) \leq 1(t \geq 1/2) \) for all \( t \in \mathbb{R} \) and so for \( v \in V \)

\[ \sum_k 1(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) \geq r_K) \]

\[ \leq \sum_k 1(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) > r_K / 2) \]

\[ + \mathbb{P}(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) > r_K / 2) \]

\[ \leq \sum_k \phi \left( \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) \right) - \mathbb{E}_\phi \left( \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) \right) \]

\[ + \mathbb{P}(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) > r_K / 2) . \]

For all \( v \in V \), we have

\[ \mathbb{P}(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) > r_K / 2) \leq \frac{\mathbb{E}(\frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v)^2}{(r_K / 2)^2} \leq \frac{1}{16} . \]

Next, using the bounded difference inequality (Theorem 6.2 in [6]), the symmetrization argument and the contraction principle (Chapter 4 in [34]) – we refer to the supplementary material of [43] for more details – with probability at least \( 1 - \exp(-K/128) \),

\[ \sup_{v \in V} \left( \sum_k \phi \left( \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) \right) - \mathbb{E}_\phi \left( \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v) \right) \right) \]
Proof of Theorem 3. We just use Lemma 5, this time with

\[ \sup_{v \in V} \left( \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v \right) \]

and, for any \( a \in \mathbb{R}^d \)

\[ \sup_{v \in B_0^d} \langle \mathbb{E}(Y_1) - a, v \rangle = \| \mu - a \| , \]

so by Lemma 5, we obtain that if \( K \geq C(VC(V) \lor |O|) \), then, with probability \( \geq 1 - \exp(-K/128) \)

\[ \| \hat{\mu} - \mu \| \leq 8\|\Sigma\| \sqrt{\frac{K}{N}}. \]

We continue with the proof of Theorem 3 for estimating sparse means.

Proof of Theorem 3. We just use Lemma 5, this time with \( g : x \rightarrow x, U = U_s \) and \( V = U_{2s} \cap B_2 \).

We have

\[ \sup_{u \in U_{2s} \cap B_2} \mathbb{E} \left( \langle g(Y_1) - \mathbb{E}(g(Y_1)), u \rangle \right)^2 = \sup_{u \in U_{2s} \cap B_2} \left\| \Sigma^{1/2} u \right\|_2^2 = \lambda_1(\Sigma) \]

and, for any \( a \in U_s \) (so a fortiori for \( \hat{\mu} \in U_s \)),

\[ \sup_{u \in U_{2s} \cap B_2} \langle \mathbb{E}(Y_1) - a, u \rangle = \| \mu - a \|_2. \]

because we assumed that \( \mu \in U_s \). So by Lemma 5, as \( \mu \in U_s \), we obtain that if \( K \geq C(VC(U_{2s}) \lor |O|) \), then, with probability \( \geq 1 - \exp(-K/128) \)

\[ \| \hat{\mu} - \mu \|_2 \leq 8\lambda_1(\Sigma) \sqrt{\frac{K}{N}}. \]

We recalled in part 2.2 that \( VC(U_{2s}) \leq 2s \log(d/s) \), which concludes the proof.
Proof of Theorem 5. Let \( V = \{ U \in \mathcal{M}_d^2(\mathbb{R}), \| U \|_F = 1 \} \). This is just Theorem 2, because we recalled in part \(2.2\) (Proposition 1) that \( \text{VC}(V) \leq c_0kd \), for some universal constant \( c_0 \), which concludes the proof.

We move to the proof of Theorem 6, for estimating covariance with respect to the canonical euclidean operator norm.

Proof of Theorem 6. This time, we take \( g : x \rightarrow xx^T \), \( U = \text{Sym}(d) \), and \( V = \{ uu^T | u \in B_2(\mathbb{R}^d) \} \). We notice that \( \mathbb{E}(g(Y_1)) = \Sigma \)

We have

\[
\sup_{M \in V} \mathbb{E} \left( (g(Y_1) - \mathbb{E}(g(Y_1)), M)^2 \right) = \sigma^2
\]

by definition of \( \sigma^2 \), and for any \( A \in \text{Sym}(d) \) (so a fortiori for \( \hat{\Sigma} \in \text{Sym}(d) \))

\[
\sup_{M \in V} \langle \Sigma - A, M \rangle = \| \Sigma - A \|.
\]

By Lemma 5, as \( \Sigma \in \text{Sym}(d) \), we obtain that if \( K \geq C(\text{VC}(V) \lor |O|) \), then, with probability \( \geq 1 - \exp(-K/128) \)

\[
\left\| \hat{\Sigma} - \Sigma \right\| \leq 8\beta \sqrt{\frac{K}{N}}.
\]

We recalled in part \(2.2\) (Proposition 1) that \( \text{VC}(V) \leq c_0d \), for some universal constant \( c_0 \), which concludes the proof.

Acknowledgements: I would like to thank Guillaume Lecué, Matthieu Lerasle and James Brian Mitchell for their precious comments. I also wish to thank anonymous referees for helping me change and improve this paper.

References

[9] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber’s contamination model.
Appendix

6.4. Proof of Theorem 4

This proof is a bit different from the rest because we will have to control two different events.

Proof. Let \( \mathcal{F} = \{ (\mathbf{x}, y) \in \mathbb{R}^{(d+1) \times m} : \mathbf{1}_{(u, \sum_i (y_i - \langle \beta^*, x_i \rangle) x_i) \geq 2r^2, u \in B_2} \}. \) This is not a set of indicators of half-spaces, but \( \mathcal{F} \) is the composition of \( g : (\mathbf{x}, y) \in \mathbb{R}^{(d+1) \times m} \to (u \to (u, \sum_i (y_i - \langle \beta^*, x_i \rangle) x_i)^2 - m^2r^2) \in \mathbb{R}_+^{d+1} \) and of \( \{ P \in \mathbb{R}_+^{d+1} : 1_{P(u) \geq 2}, u \in S - S \}. \) By Lemma 3 there exists an absolute constant \( \epsilon \) such that \( \text{VC}(\mathcal{F}) \leq \epsilon \text{VC}(S - S). \)

Let \( \mathcal{G} = \{ (\mathbf{x}_i) : 1_{\sum_i (x_i, u)^2 \geq 2r}, u \in S - S \}. \) The same way, by Lemma 3, there exists an absolute constant \( \epsilon \) such that \( \text{VC}(\mathcal{G}) \leq \epsilon \text{VC}(S). \) Assume that \( K \geq C(\text{VC}(\mathcal{F}) \lor \text{VC}(\mathcal{G}) \lor |O|) \) where \( C \) is the universal constant introduced in Lemma 2.

Multiplier process: Let

\[
\gamma = 4 \sqrt{\sup_{u \in B_2} \frac{\mathbb{E}(\xi_1^2 (u, Y_1)^2)}{N}}.
\]

For all \( u \in B_2, \)

\[
\mathbb{P}\left( \left| \frac{1}{m} \sum_{i \in B_1} (V_i - \langle \beta^*, Y_i \rangle) \langle u, Y_i \rangle \right| \geq \gamma \right) \leq \frac{\mathbb{E}(\xi_1^2 (u, Y_1)^2)}{mr^2} \leq \frac{1}{16}.
\]

By Lemma 2 applied with \( \mathcal{F}, \) it follows that the following event \( \mathcal{E} \) has probability \( \geq 1 - \exp(-K/128) \): for all \( u \in B_2, \) there exist more than \( 3/4K \) blocks \( k \) where

\[
\left| \sum_{i \in B_k} (Z_i - \langle a, X_i \rangle) \langle u, X_i \rangle \right| \leq mr.
\]

Quadratic process: From Chebyshev’s inequality, for any \( u \in S - S, \)

\[
\mathbb{P}\left( \left| \frac{1}{m} \sum_{i \in B_1} \langle u, Y_i \rangle \right| \leq \mathbb{E} \left| \langle u, Y_1 \rangle \right| - 4 \sqrt{\frac{\mathbb{E} \langle u, Y_1 \rangle^2}{m}} \right) \leq \frac{1}{16}.
\]

So, when \( K \leq \gamma^2 N/64, \) by the small ball hypothesis,

\[
\mathbb{P}\left( \left| \frac{1}{m} \sum_{i \in B_1} \langle u, Y_i \rangle \right| \leq \gamma / 2 \sqrt{\mathbb{E} \langle u, Y_1 \rangle^2} \right) \leq \frac{1}{16}.
\]
As \( \frac{1}{m} \sum_{i \in B_k} | \langle u, X_i \rangle | \leq \sqrt{\frac{1}{m} \sum_{i \in B_k} | \langle u, X_i \rangle |^2} \), by Lemma 2 applied with \( G \) and \( \bar{r} = m\gamma^2 / 4\mathbb{E} \langle u, Y_1 \rangle^2 \), the following event \( A \) has probability probability \( \geq 1 - \exp(-K/128) \): for all \( u \in \mathcal{S} - \mathcal{S} \), there exists more than \( 3/4 \) blocks \( k \) where

\[
\frac{1}{m} \sum_{i \in B_k} | \langle u, X_i \rangle | \geq \gamma^2 / 4 \langle u, \Sigma u \rangle.
\]

So we have \( Q^{K/4} \frac{1}{m} \sum_{i \in B_k} | \langle u, X_i \rangle |^2 \geq \gamma^2 / 4 \langle u, \Sigma u \rangle \).

**Conclusion of the proof.** The event \( E \cap A \) has probability at least \( 1 - 2 \exp(-K/128) \).

On \( A \), if \( u \in \hat{B}_{\Sigma} \), then \( \langle u, \Sigma u \rangle \leq 4/\gamma^2 \), so, on \( A \cap E \),

\[
\max_{u \in \hat{B}_{\Sigma}} \sum_{i \in B_k} (Z_i - \langle \beta^*, X_i \rangle) \langle u, X_i \rangle \leq 2/\gamma \max_{u \in \hat{B}_{\Sigma}} \sum_{i \in B_k} (Z_i - \langle \beta^*, X_i \rangle) \langle u, X_i \rangle \leq 2mr/\gamma.
\]

For any \( \beta \in \mathcal{S} \) such that \( \Sigma(\beta - \beta^*) \neq 0 \), let

\[
u^* = \frac{\beta - \beta^*}{\sqrt{Q^{K/4} \frac{1}{m} \sum_{i \in B_k} | \langle \beta - \beta^*, X_i \rangle |^2}}.
\]

By construction \( u^* \in \hat{B}_{\Sigma} \), so for 3/4 of the blocks, on \( E \cap A \),

\[
| \sum_{i \in B_k} (Z_i - \langle \beta^*, X_i \rangle) \langle u^*, X_i \rangle | \leq 2mr/\gamma.
\]

On the other hand, by definition, for 3/4 of the blocks,

\[
\frac{1}{m} \sum_{i \in B_k} | \langle \beta - \beta^*, X_i \rangle |^2 \geq Q^{K/4} \frac{1}{m} \sum_{i \in B_k} | \langle \beta - \beta^*, X_i \rangle |^2.
\]

Therefore, for at least half the blocks, both inequalities hold, so, on \( E \cap A \),

\[
\sum_{i \in B_k} (Z_i - \langle \beta - \beta^* + \beta^*, X_i \rangle) \langle u^*, X_i \rangle \geq -2mr/\gamma + m \sqrt{Q^{K/4} \frac{1}{m} \sum_{i \in B_k} | \langle \beta - \beta^*, X_i \rangle |^2}
\]

\[
\geq -2mr/\gamma + m\gamma^2 / 2 \sqrt{(\beta - \beta^*)\Sigma(\beta - \beta^*)}.
\]

It follows that, on \( E \cap A \), if \( \sqrt{(\beta - \beta^*)\Sigma(\beta - \beta^*)} \geq 8r/\gamma^2 \), then \( \sum_{i \in B_k} (Z_i - \langle \beta, X_i \rangle) \geq -2mr/\gamma + 4mr/\gamma \geq \sum_{i \in B_k} (Z_i - \langle \beta^*, X_i \rangle) \) and \( \beta \) can not be the chosen estimator. This concludes the proof. \( \square \)