On the robustness to adversarial corruption and to heavy-tailed data of the Stahel-Donoho median of means

Jules Depersin and Guillaume Lecué

email: jules.depersin@ensae.fr, email: guillaume.lecue@ensae.fr
CREST, ENSAE, IPParis. 5, avenue Henry Le Chatelier, 91120 Palaiseau, France.

Abstract

We consider median of means (MOM) versions of the Stahel-Donoho outlyingness (SDO) [63, 21] and of Median Absolute Deviation (MAD) [28] functions to construct subgaussian estimators of a mean vector under adversarial contamination and heavy-tailed data. We develop a single analysis of the MOM version of the SDO which covers all cases ranging from the Gaussian case to the L_2 case. It is based on isomorphic and almost isometric properties of the MOM versions of SDO and MAD. This analysis also covers cases where the mean does not even exist but a location parameter does; in those cases we still recover the same subgaussian rates and the same price for adversarial contamination even though there is not even a first moment. These properties are achieved by the classical SDO median and are therefore the first non-asymptotic statistical bounds on the Stahel-Donoho median complementing the $\sqrt{n}$-consistency [54] and asymptotic normality [71] of the Stahel-Donoho estimators. We also show that the MOM version of MAD can be used to construct an estimator of the covariance matrix under only a L_2-moment assumption or of a scale parameter if a second moment does not exist.

AMS subject classification: 62F35
Keywords: Robustness, adversarial contamination, heavy-tailed data, depth, d-dimensional median.

1 Introduction

Robust estimation of a mean vector has witnessed an important renewal during the last decade. Two communities have looked at this problem from their own perspective. In the statistics literature, several works have considered the problem of robustness with respect to heavy-tailed data. The aim here is to construct an estimator achieving statistical bounds with the same confidence as if all the data were i.i.d. Gaussian even though the data at hand are only assumed to have a second moment. Such estimators are called subgaussian estimators; they are said to be robust to heavy-tailed data. The first seminal result showing the existence of such an estimator may be found in [6]. It is also shown in [6] that the empirical mean does not achieve this goal: the rate achieved by the empirical mean in the L_2 setup cannot be better than $\sigma\sqrt{1/(\delta N)}$ with probability at least $1-\delta$ whereas it is of the order of $\sigma\sqrt{\log(1/\delta)/N}$ when the data are i.i.d. Gaussian. The rate $\sigma\sqrt{\log(1/\delta)/N}$ is called the subgaussian rate for the mean estimation problem in the one-dimensional case. It is this rate that has been first achieved in [6] by an M-estimator with a specific score function. This rate was then achieved using a median-of-means principle in several works such as [17, 4]. It was then extended to the d-dimensional case in many other works [51, 52, 8, 30, 10, 15, 39, 12] since then.

For the mean estimation problem in $\mathbb{R}^d$, most of the results have been given w.r.t. the Euclidean ℓ_2^d distance. There is however no statistical justification for this choice but that the ℓ_2^d metric is simply the most natural Hilbert metric in $\mathbb{R}^d$ and so it seems natural to use it as a way to measure the statistical performance of an estimator of a d-dimensional vector. The resulting confidence sets have therefore the form $\hat{\mu} + r_{N,\delta}^* B_2^d$ where $\hat{\mu}$ is an estimator, $B_2^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is the unit Euclidean ball and $r_{N,\delta}^*$ is the rate of convergence w.r.t. ℓ_2^d achieved by $\hat{\mu}$ with confidence $1-\delta$. When estimating w.r.t. the ℓ_2^d metric, confidence sets are therefore ℓ_2^d-balls. One may wonder if these confidence sets are the best from a statistical point of view, for instance, the one with smallest volume for a fixed confidence $1-\delta$. To answer this type of question, one usually go back to the ideal i.i.d. Gaussian case, and use results obtained in that framework as benchmark results. We may also consider this model to design optimal benchmark confidence sets, that could be used to define more appealing estimation metric of a mean vector in $\mathbb{R}^d$.

Let us now see what are the "best" (in some sense given later) confidence sets in the i.i.d. Gaussian case: let $X_1, \ldots, X_N$ be i.i.d. distributed like $\mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ is the mean and Σ is a symmetric definite positive matrix (we assume here that Σ is invertible). The MLE is the empirical mean $\bar{X}_N$ and $\sqrt{N}(\bar{X}_N - \mu) \sim \mathcal{N}(0, \Sigma)$. The

later result holds asymptotically if the data are only assumed to be in L_2 thanks to the CLT. The key observation here is that Σ is the inverse of the Fisher information in this model and thus there are no regular asymptotically normal M-estimator that can estimate the mean with an asymptotic covariance matrix better than Σ. Moreover, level sets of the standard Gaussian density function are Euclidean B_2^d balls centered at zero. As a consequence, the best confidence sets for μ with confidence $1-\delta$ are ellipsoids $\Sigma^{-1/2}B_2^d$ with radius given by the quantile of order $1-\delta$ of a chi-square variable with parameter d centered at the estimator. This type of confidence region are equivalently written as estimation results of μ with respect to the norm $x \in \mathbb{R}^d \to \left\|\Sigma^{-1/2}x\right\|_2$. It follows that the best metric – that is the one leading to minimal volume confidence sets for a given confidence in the benchmark i.i.d. Gaussian case – is the norm $\left\|\Sigma^{-1/2}\cdot\right\|_2$ whose unit ball is the ellipsoid $\Sigma^{1/2}B_2^d$.

Regarding our robust mean estimation problem, the two next natural questions are the following: is it possible to construct robust mean estimators w.r.t. the $\left\|\Sigma^{-1/2}\cdot\right\|_2$ metric? and what is the best convergence rate one can hope for? In the literature [50, 16], one may find estimators which can estimate in a robust way a mean vector w.r.t. any metric of the type $u \in \mathbb{R}^d \to \|u\|_S = \sup_{v\in S}\langle v, u\rangle$ where $S \subset \mathbb{R}^d$. In particular, for $S = \Sigma^{1/2}B_2^d$, this metric coincides with the one we want to use, i.e. $\left\|\Sigma^{-1/2}\cdot\right\|_2$. It has also been proved that the optimal deviation minimax rate (the one obtained in the benchmark i.i.d. Gaussian case) is for the mean estimation problem with respect to $\|\cdot\|_S$ given by (see [16])

$$\sqrt{\frac{\ell^*(\Sigma^{1/2}S)}{N}} + \sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\frac{\log(1/\delta)}{N}}. \tag{1}$$

For instance, for $S = B_2^d$ that is for $\|\cdot\|_S = \|\cdot\|_2$, the later rate is the classical $\sqrt{\text{Tr}(\Sigma)/N} + \sqrt{\|\Sigma\|_{op}\log(1/\delta)/N}$ rate. The case that is interesting to us is when $\|\cdot\|_S = \left\|\Sigma^{-1/2}\cdot\right\|_2$, that is for $S = \Sigma^{-1/2}B_2^d$. In that case, the subgaussian rate is

$$\sqrt{\frac{d}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}. \tag{2}$$

This is the rate we will try to reach from an adversarial corrupted and heavy-tailed dataset. We will also have to take into account the price for corruption. There are indeed known information theoretic lower bounds showing that there are no statistics that can do better than $(|\mathcal{O}|/N)^\alpha$ where $\alpha \in [1/2, 1]$ is some exponent depending on properties of the good data. For instance, $\alpha = 1$ for Gaussian variables and $\alpha = 1/2$ for some L_2 variables. However, we will see that the best possible cost $|\mathcal{O}|/N$ (i.e. for $\alpha = 1$) can be achieved even variables which do not have a first moment as long as the cdfs of all one-dimensional projections of the centered and normalized data are regular enough.

Unfortunately, all estimators known to achieve the subgaussian rate in (2) (the Le Cam test estimator in [50], the minmax MOM estimator with loss function $\ell(x, u) = \left\|\Sigma^{-1/2}(u - v)\right\|$ from [43] or the Fenchel-Legendre estimators from [16]) are using the set S in their construction. This is something we cannot do here because $S = \Sigma^{-1/2}B_2^d$ depends on Σ which is unknown in general. One therefore has to consider other type of estimators than the ones cited above. In this work, we will do it thanks to a notion of depth/outlyingness introduced at the beginning of the 80's which, unlike the last cited estimators, uses a normalization by a robust estimation of the scale.

There are several ways to measure how 'deep' is a vector with respect to a cloud of points, see for instance the half-space depth of Tukey [64, 58], the simplicial depth [44, 46], Mahalanobis depth or the projection depth [45]. Taking a point with maximal depth is usually seen as a way to define a median in $\mathbb{R}^d$ (see Radon points [2] or Fermat Points [27]). There are therefore several ways to define a median of a cloud of points in $\mathbb{R}^d$. One depth has received a particular attention both in theory and in practice and is known as the Stahel-Donoho outlyingness (SDO) [63, 23]. It can be used to construct estimators of multivariate location and scatter known as the Stahel-Donoho estimators (SDE) which were the first equivariant estimators with a high breakdown point. The aim of this work is to show that this notion of depth can be used to construct estimator of a mean vector in $\mathbb{R}^d$ which is robust to adversarial contamination and to heavy-tailed data with respect to $\left\|\Sigma^{-1/2}\cdot\right\|_2$. Let us now define this notion of depth[1] and recall some of its properties.

There is a common approach to many notion of depths for a general d-dimensional set of vectors: first, a definition of depth in $\mathbb{R}$ is given and second, this notion is extended to $\mathbb{R}^d$ simply by applying this one-dimensional definition to the set of one-dimensional projections of the data in all directions $v \in \mathbb{R}^d$ (or all $v \in S$ for some subset $S \subset \mathbb{R}^d$) and then by taking the supremum over all $v \in \mathbb{R}^d$ (or $v \in S$). This approach is based on the idea that if a point in $\mathbb{R}^d$ is an outlier then there must be some direction v such that it is an (univariate) outlier when projected into that direction.

The SDO of $z \in \mathbb{R}$ with respect to a dataset $\{a_1, \ldots, a_K\}$ in $\mathbb{R}$ is defined as

$$SDO(z; \{a_1, \ldots, a_K\}) = \frac{|z - \text{Med}(a_k)|}{\text{Med}(|a_k - \text{Med}(a_k)|)} \tag{3}$$

---

[1]We speak here about depth instead of outlyningness : these two concepts are expressing the same notion but in reverse order.

and a natural extension to $\mathbb{R}^d$ is using the previous one for all one-dimensional projections of the data and by taking the supremum over all directions: for any $\nu \in \mathbb{R}^d$ and a dataset $\{Z_1, \ldots, Z_K\}$ in $\mathbb{R}^d$, we set

$$SDO(\nu, \{Z_1, \ldots, Z_K\}) = \sup_{v \in \mathbb{R}^d} SDO(\langle \nu, v \rangle; \{\langle Z_1, v \rangle, \ldots, \langle Z_K, v \rangle\})$$

$$= \sup_{v \in \mathbb{R}^d} \frac{|\langle \nu, v \rangle - \text{Med}(\langle Z_k, v \rangle)|}{\text{Med}(|\langle Z_k, v \rangle - \text{Med}(\langle Z_k, v \rangle)|)}. \tag{4}$$

A natural way to define a median of the Z_k's is obtained by taking a point with minimal outlyingness (i.e. maximal depth):

$$\hat{\mu}^{SDO} \in \underset{\mu \in \mathbb{R}^d}{\text{argmin}}\, SDO(\mu, \{Z_1, \ldots, Z_K\}).$$

However $\hat{\mu}^{SDO}$ is not the most usual choice to estimate some location of the Z_k's when they are assumed to follow a statistical model. The Stahel-Donoho location estimator is rather defined as a convex sums of the data:

$$\hat{\mu}_K^{SDE} = \frac{\sum_{k=1}^K w_k Z_k}{\sum_{k=1}^K w_k} \tag{5}$$

where the weights are some function of the outlyingness of the data, i.e. $w_k = w(SDO(Z_k))$ for some (decreasing) weight function $w : \mathbb{R}^+ \to \mathbb{R}^+$. The weights can also be used to estimate the scatter of the points by

$$\hat{\Sigma}_K^{SDE} = \frac{\sum_k w_k (Z_k - \hat{\mu}^{SDE})(Z_k - \hat{\mu}^{SDE})^\top}{\sum_k w_k}. \tag{6}$$

Note that there is a more general definition of SDO than the one considered in (3) with general (one dimensional) definitions of location and scale statistics; in (3), we used the median $\text{Med}(a_k)$ and Median Absolute Deviation (MAD) $\text{Med}(|a_k - \text{Med}(a_k)|)$ for these statistics [29].

As mentioned previously several results on the Stahel-Donoho Estimator (SDE) have been established during the last forty years. They are *affine equivariant* meaning that for any affine transformation $x \in \mathbb{R}^d \to Ax + b$ of the dataset by a nonsingular matrix $A \in \mathbb{R}^{d \times d}$ and a vector $b \in \mathbb{R}^d$ the location estimator $\hat{\mu}_K^{SDE}$ is following the same transformation and the scatter estimator $\hat{\Sigma}_K^{SDE}$ is transformed via $M \in \mathbb{R}^{d \times d} \to AMA^\top$. SDE have been proved to have a *finite-sample breakdown point* [22] which is the "smallest amount of contamination necessary to upset an estimator entirely" from [24] in [21]. In [65], it is proved that the SDE with MAD replaced by the average of the k_1th and k_2th smallest absolute deviation about the median $\text{Med}(a_k)$ for $k_1 = d - 1 + [(K+1)/2]$ and $k_2 = d - 1 + [(K+2)/2]$ achieves the best finite-sample replacement breakdown point among all affine equivariant estimators obtained in [13] which is $[(K - d + 1)/2]/K$ (this result holds when the weight function w is continuous and there is an absolute constant c_0 such that $w(r) \leq c_0$, $w(r) \leq c_0/r^2$ for all $r \geq 0$). This result was later extended in Theorem 3.2 from [71]. This is to our knowledge the only established non-asymptotic properties of Stahel-Donoho estimators.

There are however several asymptotic results for SDE such as a $\sqrt{n}$-consistency in [54]: if the Z_k's are i.i.d. then $\sqrt{K}\left((\hat{\mu}_K^{SDE}, \hat{\Sigma}_K^{SDE}) - (\mathbf{t}, \mathbf{V})\right)$ tends to 0 in probability when $K \to +\infty$ where $\mathbf{t}$ and $\mathbf{V}$ are some location and scatter parameters of the distribution of Z_1. This result holds when the weight function w is such as $|w(r) - w(r')| \leq \gamma \min(1, 1/\min(r, r')^3)|r - r'|$ for all $r, r' \in \mathbb{R}$ and when for all $v \in \mathbb{R}^d$ the cumulative distribution function (cdf) of $\langle Z_1, v \rangle$ denoted by F_v satisfies the following assumption : there exists some absolute constants $c_0 > 0$ and $c_1 > 0$ such that for all $|\epsilon| \leq c_0$

$$|F_v(\text{Med}(F_v) + \epsilon) - F_v(\text{Med}(F_v))| \geq c_1|\epsilon| \quad \text{and} \quad |F_v(\text{Med}(F_v) \pm \sigma_v + \epsilon) - F_v(\text{Med}(F_v \pm \sigma_v))| \geq c_1|\epsilon| \tag{7}$$

where $\text{Med}(F_v) = \inf(x \in \mathbb{R} : F_v(x) \geq 1/2)$ is the median of F_v and $\sigma_v = \text{Med}(G_v)$ where G_v is the cumulative distribution of the random variable $MAD(\langle Z_1, v \rangle) := \text{Med}(|\langle Z_1, v \rangle - \text{Med}(\langle Z_1, v \rangle)|)$. A typical situation mentioned in [54] where (7) holds is when the cdf $F : \mathbb{R}^d \to [0, 1]$ of Z is such that $F = (1 - \eta)F_0 + \eta F^*$ where $\eta < 1$ and F^* is any cdf and F_0 is such that there exists $c_0 > 0$ and $c_1 > 0$ such that for all $v \in \mathbb{R}^d$, $\langle Z_1, v \rangle$ has a density denoted by f_v satisfying $f_v(t) \geq c_1$ for all $t \in [\text{Med}(F_v) \pm c_0] \cup [\text{Med}(F_v) - \sigma_v \pm c_0] \cup [\text{Med}(F_v) + \sigma_v \pm c_0]$. According to [54], the later holds when F is spherical with positive density in a neighborhood of 0 and $\sigma_{e_1} e_1$ where $e_1 = (1, 0, \cdots, 0) \in \mathbb{R}^d$. We will come back later on these conditions since we will encounter similar assumptions for our analysis. Finally, asymptotic normality of SDE location estimators have been obtained in [71] under great generality for the location and scatter estimators as well as for the weight function including the median and MAD estimators as in (3) and the projection depth obtained for the weight function $w : r \in \mathbb{R}^+ \to 1/(1 + r)$. From a stochastic point of view, asymptotic results for $\hat{\mu}_K^{SDE}$ hold when the cdf F is *elliptically symmetric around* μ which means that there exists a

symmetric definite positive matrix Σ such that for all $v \in \mathcal{S}_2^{d-1} := \{v \in \mathbb{R}^d : \|v\|_2 = 1\}$, $\langle \Sigma^{-1/2}(Z_1 - \mu), v \rangle$ has the same distribution as $\langle \Sigma^{-1/2}(Z_1 - \mu), e_1 \rangle$ which is a univariate symmetric variable with density function f. In that case, asymptotic normality was obtained when $f(0)f(\sigma) > 0$ where $\sigma = MAD(\langle \Sigma^{-1/2}(Z_1 - \mu), e_1 \rangle)$. Again we will meet this type of condition in our analysis.

On the practical side, SDEs have been used a lot in practice and implementation on various languages such as R exists; and that is one reason why the study of the SDO may be useful, maybe more than some other notions of depth. In the original paper [63], the author proposes a random algorithm where the supremum over all directions $v \in \mathbb{R}^d$ is approximated by subsampling orthogonal directions to $d-1$ hyperplanes generated by d randomly chosen points in the dataset. Other strategies mixing random and deterministic directions have been proposed for instance in [60]. Several adaptations and extensions of this algorithm may be found in [14] for an extension to an arbitrary kernel space or in [67, 66] for a "cell-wise weights" extension of the SDO where each coordinate of each data receives its own weight. However, only very little is known on the theoretical computational side. In Section 5 of [25], an algorithm running in time $\mathcal{O}(K^{d+1} \log K)$ is mentioned but its time complexity is making this approach impractical for dimensions larger than 5. There are to our knowledge no theoretical result of any kind on the convergence of some approximate algorithm for the computation of the SDO of a point in $\mathbb{R}^d$ that could be used in practice. As mentioned already in [25], "some sort of computational breakthrough is necessary to make the estimators, as defined here, really practical". This looks to be still the case. We will however not discuss about this issue in the present work and leave this question still opened.

The aim of this work is to construct mean vector estimators robust to adversarial outliers and heavy-tailed data achieving the deviation-minimax subgaussian rate from (2) with respect to the metric $\|\Sigma^{-1/2}\cdot\|_2$. On our way to our goal, we complement the results on the $\sqrt{n}$-consistency and the asymptotic normality of SDE, by deriving the first non-asymptotic convergence rate for the original SDO median (as well as its median of means version) as a robust mean estimator in $\mathbb{R}^d$ under the following assumption.

Assumption 1. *[Adversarial contamination and L_2 inliers] There exists N random vectors $(\tilde{X}_i)_{i=1}^N$ in $\mathbb{R}^d$ which are independent with mean μ and covariance matrix Σ. The N random vectors $(\tilde{X}_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives back the modified dataset $(X_i)_{i=1}^N$ to the statistician. Hence, the statistician receives an "adversarially" contaminated dataset of N vectors in $\mathbb{R}^d$ which can be partitioned into two groups: the modified data $(X_i)_{i \in \mathcal{O}}$, which can be seen as outliers and the "good data" or inliers $(X_i)_{i \in \mathcal{I}}$ such that $\forall i \in \mathcal{I}, X_i = \tilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O} \cup \mathcal{I} = \{1, \ldots, N\}$ is unknown to the statistician.*

The contamination model defined in Assumption 1 covers the Huber ϵ-contamination model from [31] and also the $\mathcal{O} \cup \mathcal{I}$ contamination framework from [37]. It has been popularized by the Computer Science community in particular in [19, 20]. In the adversarial contamination model from Assumption 1, the set $\mathcal{O}$ can depend arbitrarily on the initial data $(\tilde{X}_i)_{i=1}^N$; the corrupted data $(X_i)_{i \in \mathcal{O}}$ can have any arbitrary dependence structure; and the informative data $(X_i)_{i \in \mathcal{I}}$ may also be correlated (for instance, it is the case, in general, when the $|\mathcal{O}|$ data $\tilde{X}_i$ with largest ℓ_2^d-norm are modified by the adversary).

In the setup defined by Assumption 1, we will use the SDO as one of our building block to achieve our goal as well as the Median-of-means principle [59, 1, 32]. This principle has been extensively used during the last decades in particular for the problem of robust mean estimation [42, 18, 56, 50, 49, 52, 16, 30, 10]. The starting point of MOM estimator is to chose an integer $K \in [N]$, split the dataset into K equal size blocks $B_1 \sqcup \cdots \sqcup B_K = [N]$ (w.l.o.g. we assume that N can be divided by K) and construct K empirical means $\bar{X}_k = |B_k|^{-1} \sum_{i \in B_k} X_i$, one over each block. The Stahel-Donoho Median-of-Means that will be used to achieve the subgaussian rate (2) with respect to $\|\Sigma^{-1/2}\cdot\|_2$ under in the adversarial and heavy-tailed setup from Assumption 1 is

$$\hat{\mu}_{MOM,K}^{SDO} \in \underset{\mu \in \mathbb{R}^d}{\operatorname{argmin}} \sup_{\|v\|_2 = 1} \frac{|\langle \mu, v \rangle - \operatorname{Med}(\langle \bar{X}_k, v \rangle)|}{\operatorname{Med}(|\langle \bar{X}_k, v \rangle - \operatorname{Med}(\langle \bar{X}_k, v \rangle)|)}.$$

Unlike recently introduced robust mean estimators, $\hat{\mu}_{MOM,K}^{SDO}$ is using a robust scatter estimator for normalization. Here its is a MOM version of MAD which is used to construct $\hat{\mu}_{MOM,K}^{SDO}$, i.e. $v \to \operatorname{Med}(|\langle \bar{X}_k, v \rangle - \operatorname{Med}(\langle \bar{X}_k, v \rangle)|)$. We will show that this normalization plays a central role in the analysis when one wants results w.r.t. the $\|\Sigma^{-1/2}\cdot\|_2$-norm. But beyond this observation, we will show that MAD and its MOM version satisfy isomorphic and almost-isometric properties that can be used for other task such as to construct estimator of the covariance under only a L_2 assumption as in Section 4 below.

The paper is organized as follows. In the next section, we consider the case where the good data have a Gaussian distribution and the dataset has been adversarially corrupted. In that case, no need to construct bucketed means and

the original Stahel-Donoho median is proved to achieve the subgaussian rate (2). The Section 3 considers the general adversarial corrupted and heavy-tailed framework from Assumption 1 where the MOM version of the SDO is proved to achieve the subgaussian rate. We also exhibit in this section a family of cdfs denoted here by $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$ which plays a key role in our analysis. In particular, when the behavior of these function around 0 is similar to the one described above in (7) then the same result as in the Gaussian case can be obtained and that may hold without anymore than 2 moments (see Section 3.3). In Section 4, we show how to use the MOM version of MAD to construct an estimator of the covariance matrix under L_2. In Section 5, we explore the properties of the family of functions $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$. A conclusion and open questions are provided in Section 6 that are followed by the proofs of all the results in Section 7.

Notations. We denote by $x \in \mathbb{R}^d \to \|x\|_2 = \left(\sum_j x_j^2\right)^{1/2}$ the Euclidean norm with associated unit sphere $\mathcal{S}_2^{d-1}$ and ball B_2^d. We also denote by $g \sim \mathcal{N}(0,1)$ a standard one-dimensional Gaussian variable and its associated standard Gaussian cdf by $\Phi : t \in \mathbb{R} \to \mathbb{P}[g \leq t] = \int_{-\infty}^t \phi(u)du$ where $\phi : u \in \mathbb{R} \to (2\pi)^{-1/2}\exp(-u^2/2)$ is the one dimensional Gaussian density function. We also set $H_G : t \to 1 - \Phi(t)$ and $W_G : p \in (0,1) \to H_G^{(-1)}(p)$ the inverse function of H_G so that $W(p) = \Phi^{-1}(1-p)$.

2 The Gaussian case

In this section, we prove that the original SDO median achieves the (non-asymptotic) subgaussian rate (2) when the dataset may have been corrupted by an adversary and when the good data have a Gaussian distribution; our main model assumption is the following.

Assumption 2. *[Adversarial contamination and Gaussian inliers] There exists N i.i.d. Gaussian vectors $(G_i)_{i=1}^N$ in $\mathbb{R}^d$ with mean μ and (unknown) covariance matrix Σ. We assume that Σ is invertible. The N random vectors $(G_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives the modified dataset $(X_i)_{i=1}^N$ to the statistician.*

We use the Gaussian case again as a benchmark case for the more involved heavy-tailed situation which requires to bucket the data and some assumption on the distribution of the good data. When the "good" data are Gaussian there is no need to bucket the data and the elliptically symmetric property of the Gaussian variables is simplifying the analysis. The mean estimator we use in this section is therefore the median of the original Stahel-Donoho outlyingness function

$$SDO : \mu \in \mathbb{R}^d \to \sup_{v \in \mathbb{R}^d} \frac{|\langle \mu, v \rangle - \mathrm{Med}(\langle X_i, v \rangle)|}{\mathrm{Med}(|\langle X_i, v \rangle - \mathrm{Med}(\langle X_i, v \rangle)|)} \tag{8}$$

and the associated median is a point minimizing this outlyingness function:

$$\hat{\mu}^{SDO} \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} SDO(\mu).$$

Our main result in the adversarial corruption setup with Gaussian inliers is the following:

Theorem 1. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that the adversarial contamination with Gaussian inliers model Assumption 2 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Let $0 < \epsilon < \Phi^{-1}(3/4)/c_1$. We assume that $|\mathcal{O}| \leq \epsilon N$ and $N \geq c_0 \epsilon^{-2} d$. For all $0 < u < c_1 \epsilon^2 N$, with probability at least $1 - 2\exp(-u)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO} - \mu)\right\|_2 \leq 2(1 + c_1\epsilon)\left(C_0\sqrt{\frac{d+1}{N}} + \sqrt{\frac{u}{N}} + \frac{|\mathcal{O}|}{N}\right).$$

Let us first remark that if $N < d$ then the N data $X_1, \ldots, X_N$ cannot span the entire $\mathbb{R}^d$ space and so there exists a non zero vector $v \in \mathbb{R}^d$ which is orthogonal to all the data points. Hence, $MAD(v) := \mathrm{Med}(|\langle X_i, v \rangle - \mathrm{Med}(\langle X_i, v \rangle)|) = 0$ a.s. and so $SDO(\mu) = +\infty$ for all $\mu \in \mathbb{R}^d$. Therefore, assuming that $N \geq d$ is a minimal assumption when we work with the SDO function. We also note that the factor $\Phi^{-1}(3/4)$ is sometimes used as a renormalization factor in the definition of the Stahel-Donoho outlyingness function [62].

Theorem 1 shows that the SD median $\hat{\mu}^{SDO}$ is robust to adversarial contamination up to a proportion of N and that the rate achieved remains the same as if there was no contamination when $|\mathcal{O}| \lesssim \sqrt{N}\min(\sqrt{u}, \sqrt{d})$. If we put this result with regard to the finite-sample replacement breakdown point (RBP) achieved by the SDE with a slight modification of MAD at the denominator as recalled in the Introduction, we see that the order of magnitude are

the same: SDE and $\hat{\mu}^{SDO}$ can handle both a proportion of N adversarial outliers. The constant in the RBP (which is close to $1/2$ when $N >> d$) is certainly better than the one obtained in Theorem 1 but the result in the later theorem shows that the estimator still achieve the deviation minimax rate (2) even up to ϵN outliers whereas RBP can only insure that the estimator does not go to infinity; RBP does not ensure any statistical convergence rate after data corruption unlike Theorem 1 does.

The rate of convergence obtained in Theorem 1 has been obtained by several other procedures. For instance, it has been proved that the Tukey median achieves this rate in [8] when the covariance is proportional to the identity and for the Huber-contamination setup. The same bound was also obtained by a polynomial time algorithm in [11] when the covariance matrix Σ is known.

The proof of Theorem 1 (which may be found in Section 7) is based on two isomorphic principles of the MAD and SDO functions. We will extend these two properties to the MOM versions of MAD and SDO in the next section. For the moment, let us recall their definitions and write these two properties that are interesting beyond the proof of Theorem 1.

The normalization factor in the SDO function (8) is called the MAD (median absolute deviation) [28]

$$MAD : v \in \mathbb{R}^d \to \mathrm{Med}(|\langle X_i, v\rangle - \mathrm{Med}(\langle X_i, v\rangle)|).$$

It plays a key role to get estimation result w.r.t. the $\left\|\Sigma^{-1/2}\cdot\right\|_2$ norm whereas Σ is unknown. However, this normalization factor requires some more work than for the analysis of classical robust estimators that are only focused on the estimation of the mean. Indeed, $MAD(v)$ is actually a robust estimator of the scatter of $\langle G, v\rangle$ which is $\Phi^{-1}(3/4)\left\|\Sigma^{1/2}v\right\|_2$ (note that if $g \sim \mathcal{N}(0,1)$ then $MAD(g) = \Phi^{-1}(3/4)$). It is therefore a 'second order' robust estimator but since it appears in the denominator of the SDO function, we cannot only prove an upper estimate for this quantity and we need an isomorphic result – that is upper and lower matching (up to constants) bounds – on the MAD. This result is of independent interest and we are therefore stating it here. The proof is given in Section 7.1. We also state a similar isomophic result for SDO which can be use to prove Theorem 1. We will see later in Section 5 that these metric properties of SDO and MAD can be extended to cases where the mean does not even exists (in that case μ is a *location* parameter) showing that these properties have actually more to do with elliptical symmetry than they have to do with concentration.

Proposition 1. *There are absolute constants c_1, c_2 and c_3 such that the following holds. Let $0 < \epsilon < \Phi^{-1}(3/4)/c_1$. We assume that the adversarial model with Gaussian inliers Assumption 2 holds with a number of adversarial outliers $|\mathcal{O}| \leq \epsilon N$. We assume that $N \geq c_2\epsilon^{-2}d$. With probability at least $1 - \exp(-c_3\epsilon^2 N)$, for all $v \in \mathbb{R}^d$,*

$$(\Phi^{-1}(3/4) - c_1\epsilon)\left\|\Sigma^{1/2}v\right\|_2 \leq MAD(v) \leq (\Phi^{-1}(3/4) + c_1\epsilon)\left\|\Sigma^{1/2}v\right\|_2.$$

Moreover, for all $0 < u < c_1\epsilon^2 N$, with probability at least $1 - 2\exp(-u)$, for all $v \in \mathbb{R}^d$, if $\left\|\Sigma^{-1/2}(v - \mu)\right\|_2 \geq 2r^*$ then

$$\frac{\left\|\Sigma^{-1/2}(v - \mu)\right\|_2}{2(\Phi^{-1}(3/4) + c_1\epsilon)\sqrt{K/N}} \leq SDO_K(v) \leq \frac{3\left\|\Sigma^{-1/2}(v - \mu)\right\|_2}{2(\Phi^{-1}(3/4) - c_1\epsilon)\sqrt{K/N}}$$

and if $\left\|\Sigma^{-1/2}(v - \mu)\right\|_2 \leq 2r^*$ then $SDO_K(v) \leq 3r^*(\Phi^{-1}(3/4) - c_1\epsilon)^{-1}$ where Φ^{-1} is the quantile function of a standard Gaussian cdf and $r^* = \left(C_0\sqrt{(d+1)/N} + \sqrt{u/N} + |\mathcal{O}|/N\right)$ is the subgaussian rate from (2) with the additive adversarial contamination term $|\mathcal{O}|/N$.

The isomorphic properties of the MAD and SDO functions uniformly over $\mathbb{R}^d$ imply the robustness and subgaussian properties of the SDO median in Theorem 1. Similar results for other depths may be found in the literature on robust mean estimation such as the isomorphic property of the Tukey depth proved in [8].

3   The L_2 case

In this section, we do not anymore assume that the good data follow a Gaussian distribution but we only assume that they have a second moment (and that the dataset may still be contaminated by an adversary following Assumption 1). Nevertheless, even though we are in the heavy tailed setup with adversarially corrupted data we still want to achieve the subgaussian rate for the $\left\|\Sigma^{-1/2}\cdot\right\|_2$-norm. To achieve such a result the median-of-means principle has been proved to perform well. We will therefore use this principle together with the Stahel-Donoho concept of outlyingness. We introduce now an estimator constructed according to these two principles.

Let $K \in [N]$ be the number of blocks and let $\bar{X}_k = (1/|B_k|) \sum_{i \in B_k} X_i, k \in [K]$ be the bucketed means. Outlyingness / depth of a point $\mu \in \mathbb{R}^d$ is measured with respect to the bucketed means:

$$SDO_K(\mu) = \sup_{v \in \mathbb{R}^d} \frac{|\langle \mu, v \rangle - \mathrm{Med}(\langle \bar{X}_k, v \rangle)|}{\mathrm{Med}(|\langle \bar{X}_k, v \rangle - \mathrm{Med}(\langle \bar{X}_k, v \rangle)|)}$$

and the Stahel-Donoho Median of means is defined as

$$\hat{\mu}_{MOM,K}^{SDO} \in \underset{\mu \in \mathbb{R}^d}{\mathrm{argmin}} \, SDO_K(\mu).$$

As for the Gaussian case, the isomorphic and nearly-isometric properties of SDO_K and its denominator, called $MOMAD_K$, play a key role in our analysis. The $MOMAD_K$ is a Median of means version of the Median Absolute Deviation function. We denote it as MOMAD for Median Of Means Absolute Deviation:

$$MOMAD_K : v \in \mathbb{R}^d \rightarrow \mathrm{Med}\left(|\langle \bar{X}_k, v \rangle - \mathrm{Med}(\langle \bar{X}_k, v \rangle)|\right). \tag{9}$$

In the next section, we study metric properties of $MOMAD_K$ and for SDO_K that will useful for our analysis of $\hat{\mu}_{MOM,K}^{SDO}$. Then, we will turn to the statistical bounds obtained for the median $\hat{\mu}_{MOM,K}^{SDO}$ in the general heavy-tailed L_2 setup in Section 3.2 and then we will study some extra regularity assumption of the cdfs $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$ at 0 that allows to get better rates in Section 3.3.

3.1 Some isomorphic and almost isometric properties of $MOMAD_K$ and SDO_K

In this section, we show that the MOM versions of the SDO and MAD operators (called SDO_K and $MOMAD_K$) satisfy an isomorphic and almost-isometry properties as the MAD and SDO do in Proposition 1 that holds only under a L_2 moment assumption.

We introduce two families of functions which play a central role in our analysis. They involve the non-corrupted random variables $\tilde{X}_i, i \in [N]$ (and not the corrupted data $X_i, i \in [N]$).

Definition 1. *For all $v \in \mathcal{S}_2^{d-1}$,*

$$H_v := H_{N,K,v} : r \in \mathbb{R} \rightarrow \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle \geq r\right] \quad and \quad W_v := W_{N,K,v} : p \in (0,1) \rightarrow H_v^{(-1)}(p), \tag{10}$$

where $H_v^{(-1)}(p) = \max(r \in \mathbb{R} : H_v(r) \geq p)$ is the generalized inverse of H_v.

As already observed in the proof of the $\sqrt{n}$-consistency of SDE from [54] as well as its asymptotic normality in [71], the behavior of the one-dimensional projection cdfs at the median and the two $1/4$ and $3/4$ quartiles play a central role in the analysis of SDO based estimators. This will also be the case for the MOM version of the SD median. It will appear in Section 5 that taking bucketed mean may force toward the Gaussian case for which all these conditions are naturally satisfied because of the elliptical symmetry of Gaussian variables. Let us now state our main assumption on the behavior of the one-dimensional quantile functions $(W_v : v \in \mathcal{S}_2^{d-1})$.

Assumption 3. *There exists some $0 < \epsilon < 1/8$ and some absolute constants $0 < \varphi_l(\epsilon) < \varphi_u(\epsilon)$ such that for all $v \in \mathcal{S}_2^{d-1}$,*

$$\max\left(W_v\left(\frac{1}{4} - 2\epsilon\right) - W_v\left(\frac{1}{2} + 2\epsilon\right), W_v\left(\frac{1}{2} - 2\epsilon\right) - W_v\left(\frac{3}{4} + 2\epsilon\right)\right) \leq \varphi_u(\epsilon)$$

and

$$\min\left(W_v\left(\frac{1}{4} + 2\epsilon\right) - W_v\left(\frac{1}{2} - 2\epsilon\right), W_v\left(\frac{1}{2} + 2\epsilon\right) - W_v\left(\frac{3}{4} - 2\epsilon\right)\right) \geq \varphi_l(\epsilon).$$

Assumption 3 is a pretty weak assumption since, intuitively, it requires that the distribution of the centered and variance one real-valued random variables $\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle$ have their $1/4$-quartiles and medians constant far away as well as for their $3/4$-quartiles and medians, and this has to hold uniformly in all directions $v \in \mathcal{S}_2^{d-1}$. For instance, in the Gaussian case, Assumption 3 holds for $\varphi_u(\epsilon) = \Phi^{-1}(3/4) + c_0\epsilon$ and $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - c_0\epsilon$ for some absolute constant c_0 and for all $0 < \epsilon < 1/10$ (where we recall that $\Phi : t \rightarrow \mathbb{P}[g \leq t]$ where $g \sim \mathcal{N}(0,1)$).

Proposition 2. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 3 holds for some $0 < \epsilon < 1/8$ and constants $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$. We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers $|\mathcal{O}| \leq \epsilon K$. We assume that $K \geq c_0 \epsilon^{-2} d$. With probability at least $1 - \exp(-c_1 \epsilon^2 K)$, for all $v \in \mathbb{R}^d$,*

$$\varphi_l(\epsilon) \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \leq MOMAD_K(v) \leq \varphi_u(\epsilon) \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2.$$

Proposition 2 shows that $MOMAD_K$ is equivalent to $v \to \sqrt{K/N} \left\| \Sigma^{1/2} v \right\|_2$ up to the two constants $\varphi_u(\epsilon)$ and $\varphi_l(\epsilon)$. We will be interested in two situations regarding these constants. The first one is when their ratio is upper bounded by some absolute constant: there exists some absolute constant c_0 such that

$$\frac{\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \leq c_0. \tag{11}$$

This condition will be enough to obtain robust optimal subgaussian bounds for $\hat{\mu}_{MOM,K}^{SDO}$ in the two following theorems. If condition (11) holds we say that $MOMAD_K$ is isomorphic to $v \to \sqrt{K/N} \left\| \Sigma^{1/2} v \right\|_2$. The second condition, that will be of interest to us is when we will estimate Σ using $MOMAD_K$ in Section 4, is when the two constants $\varphi_u(\epsilon)$ and $\varphi_l(\epsilon)$ can be made arbitrarily close to the same constant by taking ϵ small enough, that is when there exists some absolute constants ϕ_0 and $c_1 > 0$ such that for all $0 < \epsilon < \phi_0/c_1$,

$$\varphi_l(\epsilon) = \phi_0 - c_1 \epsilon \text{ and } \varphi_u(\epsilon) = \phi_0 + c_1 \epsilon. \tag{12}$$

In that case, we speak about an *almost-isometric property* of $MOMAD_K$. The later condition is stronger than an isomorphic property but it allows to solve a higher order moment problem. In Section 5, we provide several examples where these conditions hold as well as other properties of the family of cdfs $(H_v : v \in \mathcal{S}_2^{d-1})$ even when there is not even a first moment.

We finish this section with an isomorphic result for SDO_K. The rate of convergence appears in this result: it is the level r^* above which SDO_K is isomorphic to $\nu \in \mathbb{R}^d \to \left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2 / \sqrt{K/N}$. One can define it as a solution to

$$C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + \sup_{\|v\|_2 = 1} H_{N,K,v}(r^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{13}$$

where u is a confidence parameter and C_0 a constant appearing in (28).

Proposition 3. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 3 holds for some $0 < \epsilon < 1/8$ and constants $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$. We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. We assume that $|\mathcal{O}| \leq \epsilon K$ and $K \geq c_0 \epsilon^{-2} d$. Let $u > 0$ and r^* be such that (13) holds. Then, with probability at least $1 - \exp(-u) - \exp(-c_1 \epsilon^2 K)$, for all $\nu \in \mathbb{R}^d$, if $\left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2 \geq 2\sqrt{K/N} r^*$ then*

$$\frac{\left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2}{2\varphi_u(\epsilon)\sqrt{K/N}} \leq SDO_K(\nu) \leq \frac{3 \left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2}{2\varphi_l(\epsilon)\sqrt{K/N}}$$

and if $\left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2 \leq 2\sqrt{K/N} r^*$ then $SDO_K(\nu) \leq (3/\varphi_l(\epsilon)) r^*$.

Proposition 3 may be seen as a MOM version holding in the heavy-tailed case of the Proposition 1 obtained in the Gaussian case. Such an extension from the Gaussian case to the L_2 heavy-tail case is made possible only thanks to the median-of-means principle and the use of the bucketed means instead of the data themselves. However, we will identify situations where condition (11) and (13) with an optimal choice of rate r^* (that is for the subgaussian rate (2)) hold for $K = N$ even when a first moment do not exist. In that case, one can get a contamination price down to $|\mathcal{O}|/N$ instead of the information theoretic lower bound in the general L_2 case given by $\sqrt{|\mathcal{O}|/N}$ (see Section 3.3). We start with the general L_2 case and then we will consider an extra assumption that allows for such better bounds.

3.2 The general L_2 case

Unlike in Section 2 or Section 3.3 below where, we demand that for all $v \in \mathcal{S}_2^{d-1}$ and for values of $0 < r < c_0$ the deviation function $H_{N,K,v}(r)$ is less than $1/2 - c_1 r$ here we simply use Markov inequality to control the function $H_{N,K,v}$ around 0. The price we pay by using this approach is that we will not have anymore estimation results for

the SDO MOM over K blocks which hold for all deviation parameter u up to K but only for $u \sim K$. The other price we pay here is for the adversarial contamination cost that will be of the order of $\sqrt{|\mathcal{O}/N}$ whereas as proved by Theorem 4 below it can be better up to $|\mathcal{O}|/N$ (as in the Gaussian case from Theorem 1). We will be able to achieve this result thanks to an extra regularity assumption of the cdfs H_v of all one-dimensional projections around 0 (see Assumption 4 below). But, for the moment, we do not grant this type of assumption in this section and obtain a general result under only the existence of a second moment as well as Assumption 3. Subgaussian rates can be derived out of this result when condition (11) holds (we refer to Section 5 where this condition is studied).

In this section, the bound we use is simply the one deduced from Markov's inequality that is for all $r > 0$ and $K \in [N]$:

$$H_{N,K,v}(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}}\sum_{i=1}^{N/K}\left\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\right\rangle \geq r\right] \leq \frac{1}{1+r^2}. \tag{14}$$

(Note that we used a slightly modification of Markov's inequality: if Z is a centered variance one real-valued random variable then $\mathbb{P}[Z \geq r] = \min_{a \in \mathbb{R}} \mathbb{P}[Z + a \geq r + a] \leq (1 + r^2)^{-1}$). Our main result in the general L_2 setup will follow from this bound and a general result stated in Section 7. It is now stated in the following theorem.

Theorem 2. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 3 holds for some $0 < \epsilon < 1/8$. We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers $|\mathcal{O}| \leq c_0 \epsilon K$. We assume that $c_0 K \geq \epsilon^{-2} d$. With probability at least $1 - 2\exp(-c_0 \epsilon^2 K)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}_{MOM,K}^{SDO} - \mu)\right\|_2 \leq \frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{K}{N}}.$$

The rate of convergence in Theorem 2 can be written like the one in Theorem 1 and Theorem 4 below where the three terms: complexity, deviation and price for adversarial corruption appear. Indeed, one should notice here that the deviation probability in Theorem 2 is fixed equal to $1 - 2\exp(-c_0 \epsilon^2 K)$ because we had to take the deviation parameter u equal to K because of the approach based on Markov's inequality (14). It is however, equivalent to replace $\sqrt{K/N}$ by $\sqrt{d/(\epsilon^2 N)} + \sqrt{u/N} + \sqrt{|\mathcal{O}|/(\epsilon N)}$ for $u = K$ since the two quantities are equivalent under the assumptions of Theorem 2. In that case, one may recognize the complexity term $\sqrt{d/N}$, the deviation term $\sqrt{u/N}$ as well as the price for adversarial corruption $\sqrt{|\mathcal{O}|/N}$. In particular, we see that the price we pay for the corruption is of the order of $\sqrt{|\mathcal{O}|/N}$ which is larger than the $|\mathcal{O}|/N$ term in the Gaussian case from Theorem 1 and it is the worst case of Theorem 4 below. Indeed, in Theorem 2 we did not exploit any other property than the existence of a second moment whereas the two other two Theorems 1 and Theorem 4 exploit some regularity assumption around 0 of the family of functions $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$.

Adaptation to K via Lepski's method. It follows from Theorem 2 that $\hat{\mu}_{MOM,K}^{SDO}$ is an estimator which depends on the deviation parameter. Therefore, we need to construct an adaptive to K version of this estimator to disentangle the estimator from the deviation parameter. The classical way to do it is via Lepski's method [40, 41]. Usually, the price we pay to make this approach work is some extra knowledge on Σ such as its trace and operator norm. But here for the SDO type estimator we are using together with the isomorphic property of the SDO_K we only need knowledge on $\varphi_u(\epsilon)$ and $\varphi_l(\epsilon)$. Let us now construct this adaptive scheme: the number of blocks is chosen via

$$\hat{K} = \min\left(K \in [N] : SDO_k(\hat{\mu}_{MOM,K}^{SDO} - \hat{\mu}_{MOM,k}^{SDO}) \leq \max\left(\frac{9}{\varphi_l(\epsilon)}, \frac{6\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)}\left(1 + \sqrt{\frac{K}{k}}\right)\right), \forall k = N, \cdots, K\right). \tag{15}$$

Theorem 3. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 3 holds for some $0 < \epsilon < 1/8$ and all $K \in [N]$. We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Then, for all $K \geq \max(c_0 \epsilon^{-2} d, c_0 |\mathcal{O}|)$ with probability at least $1 - 2\exp(-c_0 \epsilon^2 K)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}_{MOM,\hat{K}}^{SDO} - \mu)\right\|_2 \leq \frac{28\varphi_u^2(\epsilon)}{\varphi_l^2(\epsilon)}\sqrt{\frac{K}{N}}.$$

where $\hat{K}$ is the adaptive choice of number of blocks from (15).

9

3.3 The L_2 case under an extra regularity condition around 0 of the H_v's

In this section, we obtain estimation bound for the MOM version of the SDO median in the adversarial corruption with heavy-tailed L_2 inliers model under an extra assumption on the regularity at 0 of the family of functions $H_v, v \in \mathcal{S}_2^{d-1}$ that is stated now.

Assumption 4. *There exists some absolute constants $c_0, c_1 > 0$ and $c_2 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $(2C_0/c_1)\sqrt{(d+1)/K} \leq r \leq c_0$*

$$H_{N,K,v}(r) = H_v(r) := \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \left\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \right\rangle \geq r \right] \leq \frac{1}{2} - c_2 r.$$

This assumption is about the behavior around the origin of the cdf of all one-dimensional projections of the random vectors $(N/K)^{-1/2} \sum_{i=1}^{N/K} \Sigma^{-1/2}(\tilde{X}_i - \mu)$ where the $\tilde{X}_i$ are the non-corrupted data. The term $\frac{1}{2} - c_2 r$ in the bound above is the behavior of regular in 0 cdfs such as in the Gaussian case (see Section 5 for more details and more examples).

Our main result in the adversarial corruption and heavy-tailed L_2 model under Assumption 4 is the following theorem. The proof may be found in Section 7.

Theorem 4. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 3 holds for some $0 < \epsilon < 1/4$ and that Assumption 4 holds as well. We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers $|\mathcal{O}| \leq c_0 \epsilon K$. We assume that $c_0 K \geq \epsilon^{-2} d$. For all $0 < u \leq c_0 \epsilon^2 K$, with probability at least $1 - 2\exp(-u)$,*

$$\left\| \Sigma^{-1/2}(\hat{\mu}_{MOM,K}^{SDO} - \mu) \right\|_2 \leq \frac{c_1 \varphi_u(\epsilon)}{\varphi_l(\epsilon)} \left(\sqrt{\frac{d}{N}} + \sqrt{\frac{u}{N}} + \frac{|\mathcal{O}|}{\sqrt{NK}} \right). \tag{16}$$

We recover the optimal subgaussian rate (2) in Theorem 4 when for some $0 < \epsilon < 1/8$, condition (11) holds and $|\mathcal{O}| \lesssim \sqrt{Kd}$. The term $|\mathcal{O}|/\sqrt{KN}$ appearing in the convergence rate of Theorem 4 is the price we pay for the adversarial contamination. It is between $\sqrt{|\mathcal{O}|/N}$ when $K \sim |\mathcal{O}|$ and $|\mathcal{O}|/N$ when $K \sim N$. Usually when the inliers are only in L_2, the information theoretic lower bound is known to be of the order of $\sqrt{|\mathcal{O}|/N}$ and not like $|\mathcal{O}|/N$. We get a better rate in Theorem 4 thanks to Assumption 4 which is using in some more efficient way the regularity of the H_v functions at 0.

Unlike typical results in the MOM literature except for the one obtained in [55], the deviation rate in Theorem 4 is $1 - 2\exp(-u)$ for all $u \lesssim K$, in particular it does not have to depend on parameter K. As a consequence, the estimator $\hat{\mu}_{MOM,K}^{SDO}$ does not depend on the deviation parameter. Usually, results for MOM estimators constructed on K blocks are given with probability at least $1 - \exp(-c_0 K)$ and then a Lepski's method is used to construct an adaptive to K procedure. This is not the case here nor it is for the Gaussian case in Section 2 (as we did in the previous section). This is again because Assumption 4 is using more efficiently the behavior of $H_{N,K,v}$ around 0.

4 Estimation of Σ under a L_2-moment assumption with MOMAD

In this section, we show that it is possible to estimate the covariance matrix Σ using the MOMAD estimator. In particular, given that the isomorphic property of MOMAD hold under Assumption 3 which does not require more moment than L_2 moment, we show that it is possible to estimate Σ without requiring more moment than 2 that is just under the assumption that Σ exists. This differs from approaches based on the empirical covariance matrix where at best a $L_{2+\delta}$-moment assumption for some positive δ is granted for the estimation of the covariance matrix [47, 5, 48].

We show that for the estimation of Σ via the MOMAD, the properties of $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ introduced in Assumption 3 play a key role. Let us first have a look at these quantities in the Gaussian case. In that case, there are some absolute constants ϕ_0 and $c_1 > 0$ such that for all $0 < \epsilon < \phi_0/c_1$,

$$\varphi_l(\epsilon) = \phi_0 - c_1 \epsilon \quad \text{and} \quad \varphi_u(\epsilon) = \phi_0 + c_1 \epsilon \tag{17}$$

where $\phi_0 = \Phi^{-1}(3/4)$ (see Section 5 or the proof of Proposition 1 for more details). This later result holds in the Gaussian case first because the two interquartile intervals have the same length: $\Phi^{-1}(0) - \Phi^{-1}(1/4) = \Phi^{-1}(3/4) - \Phi(0) = \phi_0$ and, second, because the Gaussian density function is uniformly lower bounded by an absolute positive constant locally around the two $1/4$ and $3/4$ quartiles $\Phi^{-1}(1/4)$ and $\Phi^{-1}(3/4)$ as well as around the median $\Phi^{-1}(1/2) = 0$.

If this last condition were not true at some $q \in \{W(1/4), W(1/2), W(3/4)\}$ where $W = W_{N,K,v}$ for some direction $v \in \mathcal{S}_2^{d-1}$ then there will be some plateau of the cdf $r \in \mathbb{R} \to 1 - H_{N,K,v}(r)$ starting at q and thus there would be a constant factor gap between $W(\ell/4 - 2\epsilon)$ and $W(\ell/4 + 2\epsilon)$ for some $\ell \in \{1,2,3\}$. In that case, there would be some absolute constants $c_0 > 0$ and $c_1 > 0$ such that $|\varphi_l(\epsilon) - \varphi_l(\epsilon)| \geq c_0$ for all $0 < \epsilon \leq c_1$. In particular, we would only have an isomorphic property for the MOMAD and thus its is not clear how to estimate Σ using MOMAD at a better rate than a constant rate. Typical values of ϕ_0 in (17) will be $\phi_0 = W(1/4) - W(1/2) = W(1/2) - W(3/4)$. In particular, the interquartile interval lengths have to be equal in all directions $v \in \mathcal{S}_2^{d-1}$; this will hold, in particular, under a spherical symmetry assumption of the $\Sigma^{-1/2}(\tilde{X}_i - \mu)$ (see Section 5 for more formal statement).

Assumption 5. *For the same choice of K as in Assumption 3 where $\epsilon > 0 \to \varphi_l(\epsilon), \varphi_u(\epsilon)$ are defined, there are absolute constants ϕ_0, $c_0 > 0$ and $c_1 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $0 < \epsilon < c_0$, $\varphi_l(\epsilon) = \phi_0 - c_1\epsilon$ and $\varphi_u(\epsilon) = \phi_0 + c_1\epsilon$.*

Let us now turn to the construction of an estimator of the covariance matrix Σ using MOMAD under Assumption 5 (as well as Assumption 3). Because of the constant factor ϕ_0 in Assumption 5 we will provide an estimator of the *scatter matrix* $\phi_0^2 \Sigma$ (according to [53], a scatter matrix is any matrix proportional to the covariance matrix).

It follows from Proposition 2 that $MOMAD_K$ is isomorphic to $v \in \mathbb{R}^d \to \phi_0 \sqrt{K/N} \left\| \Sigma^{1/2} v \right\|_2$ and that under Assumption 5 it becomes an almost isometry, that is, with probability at least $1 - \exp(-c_0\epsilon^2 K)$, for all $v \in \mathbb{R}^d$,

$$\left| MOMAD_K(v) - \phi_0 \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \right| \leq c_1 \epsilon \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \tag{18}$$

as long as $|\mathcal{O}| \leq \epsilon K$ and $K \geq c_2 \epsilon^{-2} d$. In the Gaussian case and other spherical cases as in Section 5, this almost isometric property holds for $K = N$ (and $MOMAD_N = MAD$) and any $0 < \epsilon < 1/4$: it follows from Proposition 1 that with probability at least $1 - \exp(-c_0\epsilon^2 N)$, for all $v \in \mathbb{R}^d$,

$$\left| MAD(v) - \Phi^{-1}(3/4) \left\| \Sigma^{1/2} v \right\|_2 \right| \leq c_1 \epsilon \left\| \Sigma^{1/2} v \right\|_2. \tag{19}$$

We may use (18) to estimate directly the entries of Σ following an idea from [26]. Let $(e_j)_{j=1}^d$ denote the canonical basis of $\mathbb{R}^d$. We have, for all $i, j \in [d]$,

$$4\Sigma_{ij} = 4\langle e_i, \Sigma e_j \rangle = \left\| \Sigma^{1/2}(e_i + e_j) \right\|_2^2 - \left\| \Sigma^{1/2}(e_i - e_j) \right\|_2^2.$$

As a consequence, a natural estimator of $\phi_0^2 \Sigma$ based on $MOMAD_K$ is the matrix $\hat{\Sigma}$ whose entries are defined for all $i, j \in [d]$ by

$$\hat{\Sigma}_{ij} = \frac{N}{4K} \left(MOMAD_K^2(e_i + e_j) - MOMAD_K^2(e_i - e_j) \right).$$

Note that $\hat{\Sigma}$ is symmetric but it may not be a SDP. To overcome this issue, a projection method has been introduced in [48] which may also be used as well for $\hat{\Sigma}$. Our main statistical bound for $\hat{\Sigma}$ is the following.

Proposition 4. *Assume that Assumption 1 holds. Let $K \in [N]$, φ_l and φ_u be such that Assumption 3 and Assumption 5 hold. Then, for all $0 < \epsilon < c_0$ such that $|\mathcal{O}| \leq \epsilon K$ and $K \geq c_2 \epsilon^{-2} d$, with probability at least $1 - \exp(-c_3\epsilon^2 K)$,*

$$\max_{i,j \in [d]} \left| \frac{\phi_0^2 \Sigma_{ij} - \hat{\Sigma}_{ij}}{\Sigma_{ii} + \Sigma_{jj}} \right| \leq \sup_{\|u\|_1 = \|v\|_1 = 1} \left| \frac{\langle u, (\phi_0^2 \Sigma - \hat{\Sigma})v \rangle}{\sum_i (|u_i| + |v_i|)\Sigma_{ii}} \right| \leq c_1 \epsilon (c_1 \epsilon + \phi_0)/2.$$

In particular, if one can choose $K = N$ so that Assumption 3 and Assumption 5 hold – for instance, in the Gaussian case or for other spherical variables as in Section 5 – then the $MOMAD_N$ estimator becomes the classical MAD one and for $\epsilon^2 = c_2 d/N$ we have that with probability at least $1 - \exp(-c_4 d)$,

$$\max_{i,j \in [d]} \left| \frac{\phi_0^2 \Sigma_{ij} - \hat{\Sigma}_{ij}}{\Sigma_{ii} + \Sigma_{jj}} \right| \leq \sup_{\|u\|_1 = \|v\|_1 = 1} \left| \frac{\langle u, (\phi_0^2 \Sigma - \hat{\Sigma})v \rangle}{\sum_i (|u_i| + |v_i|)\Sigma_{ii}} \right| \leq c_5 \sqrt{\frac{d}{N}}.$$

as long as $|\mathcal{O}| \leq c_6 d$.

5 Study of the $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ functions

The functions $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ play a key role in our analysis. Their behavior in a neighborhood of their $1/4$ and $3/4$ quartiles and medians should be controlled so that Assumption 3 may hold: they are driving the isomorphic properties and almost isometric properties of the $MOMAD_K$ and SDO_K functions and so of the statistical performance of the Stahel Donoho Median and Median of Means. Their behavior around 0 also drives the improved rates obtained under Assumption 4. From our perspective, it is of the utmost importance to understand the behavior of these functions at these particular points.

Let us first settle down the properties of the $H_{N,K,v}$ functions desirable for our analysis. We set $Z_i = \Sigma^{-1/2}(\tilde{X}_i - \mu)$ for all $i \in [N]$ so that the Z_i's are independent centered isotropic vectors in $\mathbb{R}^d$ and $n = N/K$. We want to identify conditions on the distributions of the Z_i's such that

- for Assumption 4: there exists some absolute constants $c_0, c_1 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $0 < r < c_0$,

$$H_{n,v}(r) := \mathbb{P}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle Z_i, v \rangle \geq r\right] \leq \frac{1}{2} - c_1 r. \tag{20}$$

- for Assumption 3: there exists some absolute constants $c_0 > 0$ and $0 < \epsilon < 1/8$ such that $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ exist and are such that

$$\frac{\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \leq c_0 \tag{21}$$

or there are absolute constants ϕ_0 and $c_1 > 0$ such that for all $0 < \epsilon < \phi_0/c_1$,

$$\varphi_l(\epsilon) = \phi_0 - c_1\epsilon \text{ and } \varphi_u(\epsilon) = \phi_0 + c_1\epsilon \tag{22}$$

which are respectively Condition 11 (insuring an isomorphic property of $MOMAD_K$ and SDO_K as well as optimal subgaussian rates for SD median and median of means) and Condition 17 (insuring almost isometric property of $MOMAD_K$ as well as estimation properties for $\hat{\Sigma}$ in Section 4).

Let us first study the Gaussian case which is our benchmark situation. We will then study other cases where the family of functions $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ satisfies these conditions.

The Gaussian case. We recall that $\Phi : t \in \mathbb{R} \to \mathbb{P}[g \leq t] = \int_{-\infty}^t \phi(u)du$ where $\phi : u \in \mathbb{R} \to (2\pi)^{-1/2}\exp(-u^2/2)$ is the Gaussian density function. We also denote $H_G : t \to 1 - \Phi(t)$ and $W_G : p \in (0,1) \to H_G^{(-1)}(p)$ the inverse function of H_G so that $W_G(p) = \Phi^{-1}(1-p)$. It follows from the mean value theorem that for all $t, \epsilon \in \mathbb{R}$, $|H_G(t+\epsilon) - H_G(t)| \leq \max(-\phi(t), -\phi(t+\epsilon))\epsilon$ so that around 0 we have for all $c_0 > 0$ and $0 < r < c_0$, $H_G(r) \leq 1/2 - \phi(c_0)r$. As a consequence, (20) holds in the Gaussian case for instance with $c_0 = 1$ and $c_1 = \phi(1)$. Let us now look at the two other conditions in the Gaussian case. From the mean value theorem, we have for all $p \in [1/2, 1)$ and $\epsilon \geq 0$ such that $p + \epsilon \in [1/2, 1)$, $\epsilon/\phi(W_G(p)) \leq W_G(p) - W_G(p+\epsilon) \leq \epsilon/\phi(W_G(p+\epsilon))$ and for all $p \in (0, 1/2]$ and $\epsilon \geq 0$ such that $p - \epsilon \in (0, 1/2]$, $\epsilon/\phi(W_G(p)) \leq W_G(p-\epsilon) - W_G(p) \leq \epsilon/\phi(W_G(p+\epsilon))$. We conclude that there are absolute constants $c_0, c_1 > 0$ such that for all $0 \leq \epsilon \leq c_0$,

$$\varphi_u(\epsilon) = \Phi^{-1}(3/4) + 2\epsilon\left(\frac{1}{\phi(\Phi^{-1}(3/4))} + \frac{1}{\phi(0)}\right) + c_1\epsilon^2$$

and

$$\varphi_l(\epsilon) = \Phi^{-1}(3/4) - 2\epsilon\left(\frac{1}{\phi(\Phi^{-1}(3/4))} + \frac{1}{\phi(0)}\right) - c_1\epsilon^2.$$

So that both conditions (20) and (22) hold with $\phi_0 = \Phi^{-1}(3/4)$. In particular, we see that the values of the density function ϕ at the $1/4$ and $3/4$ quartiles (here we used that $\phi(\Phi^{-1}(3/4)) = \phi(\Phi^{-1}(1/4))$) and at the median $\phi(0)$ (here we used that $\Phi^{-1}(1/2) = 0$) play a key role.

In the following, we identify situations where the $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ functions and their pseudo inverses mimic the H_G and W_G functions from the Gaussian case. There are at least two reasons: the first one is that we are projecting random vectors leaving in $\mathbb{R}^d$ onto one dimensional subspaces; the second reason is that we are averaging random variables having a second moment. We will explore these two observations in the two following paragraphs.

One dimensional projections and elliptically contoured distributions. The fact that the H_v functions deal only with one-dimensional marginals is making these functions likely to behave as in the Gaussian case since one-dimensional projections of sufficiently spherically symmetric random vectors in $\mathbb{R}^d$ are expected to behave like one-dimensional Gaussian variables and this phenomenon is even more accentuated when d is large (this is one particular situation where large dimension d may help in Statistics). Indeed, one may have in mind an observation – sometimes attributed to H. Poincaré – that the density function of the one-dimensional projection $\langle \sqrt{d}U, e_1 \rangle$ – where $\sqrt{d}U$ is uniformly distributed over $\sqrt{d}\mathcal{S}_2^{d-1}$ and $(e_j)_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$ – converges to the density of a $\mathcal{N}(0,1)$ when $d \to \infty$ (see page 16 in [38] or Chapter 4 in [3]). One may also have in mind that there are directions such as $v = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ which are mixing the coordinates of $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ when projected onto v and therefore may have the tendency to mimic a standard Gaussian variable because of the CLT. Note that all these observations hold for $N = K$ that is even for $n = 1$: because of the one-dimensional projections we may not even have to average the Z_i's to mimic the Gaussian case. Therefore, Theorem 5 can be extended beyond the Gaussian case when this phenomenon occurs.

Let us now consider an example of elliptically contoured distributions where this happens to be true. Our aim is to show that Condition (11) and Assumption 4 (and so Theorem 4) may hold for $K = N$ (i.e. $n = 1$) even when the $\tilde{X}_i$'s do not have a first moment.

We assume that the $\tilde{X}_i$'s are i.i.d. and that $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ has a spherically symmetric distribution; in that case, $\tilde{X}_1 - \mu$ is sometimes said to have an *elliptically contoured distribution*. Then, there exists a non-negative random variable R such that $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ is distributed according to RU where U is uniformly distributed on $\mathcal{S}_2^{d-1}$ and is independent of R (see Chapter 4 in [3]). In that case, all the $\langle \Sigma^{-1/2}(\tilde{X}_1 - \mu), v \rangle$ for $v \in \mathcal{S}_2^{d-1}$ have the same distribution as $\langle \Sigma^{-1/2}(\tilde{X}_1 - \mu), e_1 \rangle$ (where $(e_j)_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$) which is distributed according to $R\langle U, e_1 \rangle$. Now, using that $\langle U, e_1 \rangle$ is absolutely continuous w.r.t. the Lebesgue measure with density function given by, when $d \geq 2$,

$$t \in \mathbb{R} \to C_d(1 - t^2)^{\frac{d-3}{2}} I(|t| \leq 1) \text{ where } C_d = \left(\int_{-1}^1 (1 - t^2)^{\frac{d-3}{2}} dt \right)^{-1} = \frac{2\Gamma(d/2)}{\Gamma((d-1)/2)\sqrt{\pi}}$$

and Γ is the Gamma function, we can deduce that (even for $K = N$), H_v is independent of $v \in \mathcal{S}_2^{d-1}$ and is such that for all $r \geq 0$, $H_v(-r) = 1 - H_v(r)$ and

$$H_v(r) = H(r) := C_d \int_0^1 \mathbb{P}[R \geq r/x] \left(1 - x^2\right)^{\frac{d-3}{2}} dx.$$

In particular, we recover that $H(0) = 1/2$ since $R \geq 0$ a.s.. Let us now consider a simple example for the distribution of R. In that example, R takes values $r_1 < r_2 < \cdots$ such that $\alpha_j = \mathbb{P}[R = r_j]$ for all $j \in \mathbb{N}^*$ so that for all $q > 0$, $\mathbb{E}R^q = \sum_j r_j^q \alpha_j$ which may be infinite even for $q = 1$ (that is when there is not even a first moment). For this example, we have for all $r \geq 0$,

$$H(r) = C_d \sum_{j=1}^\infty \alpha_j \int_{r/r_j}^1 (1 - x^2)^{\frac{d-3}{2}} dx I(r \leq r_j).$$

In particular, H is differentiable and $R\langle U, e_1 \rangle$ is absolutely continuous w.r.t. the Lebesgue measure with a density function given by

$$f : r \in \mathbb{R} \to -H'(r) = C_d \sum_{j=1}^\infty \frac{\alpha_j}{r_j} \left[1 - \left(\frac{r}{r_j} \right)^2 \right]^{\frac{d-3}{2}} I(r \leq r_j).$$

In particular, for $r_\infty = \lim_{j \to \infty} r_j$, H is strictly decreasing on $[0, r_\infty)$ from $H(0) = 1/2$ to $H(r_\infty) = 0$ and beyond r_∞ it is constant equal to 0. Therefore, for all $v \in \mathcal{S}_2^{d-1}$, the generalized inverse W_v of H_v is independent of v and it is the inverse of H: for all $p \in (0, 1/2]$ there is a unique element $W(p)(= W_v(p))$ in $[0, r_\infty)$ such that $H(W(p)) = p$ and $W(1 - p) = -W(p)$.

Now, let us choose $r_j = 2^j C_d$ and $\alpha_j = 2^{-j}$ for all $j \in \mathbb{N}$. We also assume $d \geq 4$ to make the presentation simpler (the cases $d = 1, 2, 3$ can be treated separately). In that case, $\mathbb{E}R = +\infty$ and so the mean and covariance matrix do not exist. Nevertheless, one may still assume that there exists $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ definite positive such that $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ is spherically symmetric (without having μ to be a mean vector and Σ to be a covariance matrix). Then, Theorem 4 still applies.

Let us first check Condition (20). We have $H(0) = 1/2$ and for all $0 \leq r \leq C_d/\sqrt{d - 3}$,

$$f(r) \geq \sum_{j=1}^\infty \frac{1}{2^{2j}} \left[1 - \frac{d-3}{2C_d^2} \left(\frac{r}{2^j} \right)^2 \right] \geq \frac{1}{3}. \tag{23}$$

Moreover, we see that $\sqrt{d} \leq C_d \leq 6\sqrt{d}$, hence, (23) holds for all $0 \leq r \leq 1$. Which is according to the means value theorem enough to show that Condition (20) holds (see (25) below for more details).

Let us now check conditions (21) and (22). It follows from Proposition 5 below that, it is enough to lower bound the density function f in a neighborhood of p for $p \in \{W(1/4), W(1/2), W(3/4)\}$ and that $W(1/4) - W(3/4)$ is an absolute constant. But, given that $W(1/2) = 0$ and (23) holds, that f is symmetric about 0 and that $W(1/4) = -W(3/4)$, we only have to checked that $f(q) \geq c_0$ for all $q \in [W(1/4) - 2\epsilon, W(1/4) + 2\epsilon]$ for some $0 < \epsilon < 1/8$ and an absolute constant c_0 and that $W(1/4)$ is an absolute constant. We first have to find $W(1/4)$ which is the unique solution r such that $H(r) = 1/4$. We see that f is symmetric unimodal with maximal value at 0 given by $f(0) = 4/3$ and we showed that $f(r) \geq 1/3$ for all $0 \leq r \leq 1$ in (23). Therefore, $H(1/8) \geq 1/3$ and $H(1 - 1/10) \leq 3/15 < 1/4$, hence, $W(1/4) \in [1/8, 1 - 1/10]$. It follows from (23) that $f(q) \geq 1/3$ for all $q \in [W(1/4) - 1/10, W(1/4) + 1/10]$. We conclude that both conditions (21) and (22) hold thanks to Proposition 5 below.

For this example, one can take $\varphi_u(\epsilon) = W(1/4) - (4/3)\epsilon$ and $\varphi_l(\epsilon) = W(1/4) + (4/3)\epsilon$ for all $0 < \epsilon < 1/16$. In that case, $MOMAD_K$ is an almost isometry and we can state a result like Theorem 1 where $\Phi^{-1}(3/4)$ is replaced by $W(1/4)$ and μ and Σ are not anymore the mean and covariance matrix since they do not exist but 'location' and 'scale' parameters defined such that $\Sigma^{-1/2}(\tilde{X}_i - \mu)$ are spherically symmetric. As a consequence, the phenomenon underlying the Gaussian case from Section 2 has nothing to do with concentration but it is more about elliptical symmetry.

Gaussian approximation. In cases where there is some lack of spherical symmetry of $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ one may study the H_v functions for a smaller number K of blocks so that $n = N/K$ may be large enough to see some averaging effect. In that case and because Gaussian variables satisfy all the properties we need, it is tempting to use a Gaussian approximation result such as a Berry-Esseen bound (see [61, 9, 7]) to approximate the H_v functions by $1 - \Phi$ for $n = N/K$ large enough. This strategy has been used several times in Minsker and co-authors works on Median-of-means and Catoni's type of estimators (see for instance [55, 57]).

For instance, when for all $v \in \mathcal{S}_2^{d-1}$, $\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle, i \in [n]$ are (independent, centered and variance one) real-valued random variables in $L_{2+\delta}$ such that $\left\| \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle \right\|_{2+\delta} \leq \kappa$ (uniformly in $v \in \mathcal{S}_2^{d-1}$) for some $\delta > 0$ then, it follows from Theorem 5.7 in [61] that there is an absolute constant $c_0 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $r \in \mathbb{R}$,

$$|H_{N,K,v}(r) - \mathbb{P}[g \geq r]| \leq \frac{c_1 \kappa^{2+\delta}}{n^{\delta/2}} := c_n \qquad (24)$$

It follows that for all $p \in (0,1)$ and $\epsilon \in \mathbb{R}$ satisfying $p + \epsilon \in (0,1)$ that

$$\Phi^{-1}\left(1 - p - \epsilon - c_n\right) \leq W(p + \epsilon) \leq \Phi^{-1}\left(1 - p - \epsilon + c_n\right).$$

In particular, for all $0 < \epsilon < 1/16$, if n is large enough so that $c_n \leq \epsilon$ then one can take $\varphi_u(\epsilon) = \Phi^{-1}(3/4) - c_0\epsilon$ and $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - c_0\epsilon$. So that the ratio $\varphi_u(\epsilon)/\varphi_l(\epsilon)$ is constant; in that case, the $MOMAD_K$ and SDO_k are isomorphism (see Proposition 2) and we recover a subgaussian rate in Theorem 4.

However, a Gaussian approximation result such as the one in (24) is not enough for Assumption 4. Indeed, it follows from (24) that for all $0 \leq r \leq c_0$, $H_v(r) \leq H_G(r) + c_n \leq 1/2 - c_1 r + c_n$ for some absolute constants $c_0 > 0$ and $c_1 > 0$. It appears that our analysis used to prove Theorem 4 does not stand this extra error term c_n compare with Assumption 4. Gaussian approximation does not help in this case: indeed Assumption 4 is more about the existence of a uniform lower bound around 0 of the density functions of the one-dimensional projections $\langle n^{-1/2} \sum_i Z_i, v \rangle$ as we are considering now.

Beyond the Gaussian behavior. In the later two paragraphs, we identified situations where the $n^{-1/2} \sum_{i=1}^n \langle Z_i, v \rangle$ for $v \in \mathcal{S}_2^{d-1}$ behave like Gaussian variables. We saw that this may be the case because we are considering one-dimensional projections of d-dimensional vectors and/or we are taking empirical means over n variables. But properties we are looking for the $H_{n,v}, v \in \mathcal{S}_2^{d-1}$ functions (see (20), (21) and (22)) are all dealing only with their behavior around 3 (or 4 when the median is not 0) points. So that only the behavior of these functions at these points play a role and there is no need to mimic the Gaussian case for all values of r in $\mathbb{R}$. We now state a general result going in this direction. In particular, we recover the conditions from [54] and [71] recalled in the Introduction section.

Let us assume that the $n^{-1/2} \sum_{i=1}^n \langle Z_i, v \rangle$ for $v \in \mathcal{S}_2^{d-1}$ are absolutely continuous w.r.t. the Lebesgue measure with a density function denoted by f_v. By the mean value theorem, we have for all $r \geq 0$, all $p \in (0,1)$ and $\epsilon \geq 0$ such that $p + \epsilon \in (0,1)$,

$$H_v(r) \leq H_v(0) - \min_{0 \leq t \leq r} f_v(t) r \text{ and } \frac{\epsilon}{\max_{q \in [p, p+\epsilon]} f_v(W_v(q))} \leq W_v(p) - W_v(p + \epsilon) \leq \frac{\epsilon}{\min_{q \in [p, p+\epsilon]} f_v(W_v(q))}. \qquad (25)$$

In particular, the values of the density functions $f_v, v \in \mathcal{S}_2^{d-1}$ at $0, W(1/4), W(1/2)$ and $W(3/4)$ drives the quality of inequalities from (25) and, as noted in previous works on the Stahel-Donoho outlyingness function, are enough to insure all the conditions we need on H_v and W_v recalled in (20), (21) and (22).

Proposition 5. *Let $K \in [N]$ be such that $N/K \in \mathbb{N}$. We assume that the original non-corrupted data $\tilde{X}_i, i \in [N]$ are independent and that there exists $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ definite positive so that for all $v \in \mathcal{S}_2^{d-1}$, $\sqrt{K/N} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle$ are absolutely continuous real valued random variables with a density denoted by f_v.*

If there exists $0 < \epsilon < 1/8$ and $c_0 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$, all $p \in \{W_v(1/4), W_v(1/2), W_v(3/4)\}$ and all $q \in [p - 2\epsilon, p + 2\epsilon]$, $f_v(q) \geq c_0$ then for $\mathcal{I}_v = \max(W_v(1/4) - W_v(1/2), W_v(1/2) - W_v(3/4))$

$$\varphi_u(\epsilon) = \max_{v \in \mathcal{S}_2^{d-1}} (\mathcal{I}_v + 4\epsilon/c_0) \ \text{and} \ \varphi_l(\epsilon) = \min_{v \in \mathcal{S}_2^{d-1}} (\mathcal{I}_v - 4\epsilon/c_0).$$

We also have

$$\frac{\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \leq \frac{\max_{v \in \mathcal{S}_2^{d-1}} \mathcal{I}_v}{\min_{v \in \mathcal{S}_2^{d-1}} \mathcal{I}_v} \left(1 + \frac{16\epsilon}{c_0 \min_{v \in \mathcal{S}_2^{d-1}} \mathcal{I}_v} \right)$$

when $4\epsilon \leq c_0 \min_v \mathcal{I}_v$. Moreover, if $(c_0/4) \max_v \mathcal{I}_v < 1/8$ and $\min_v \mathcal{I}_v \geq c_1$, for some absolute constant $c_1 > 0$, then condition (21) holds (and so we recover the optimal subgaussian rates in Theorem 2 and Theorem 3) and if for all $v \in \mathcal{S}_2^{d-1}$, $\mathcal{I}_v := \phi_0$ then condition (22) holds and so does Proposition 4.

If for all $v \in \mathcal{S}_2^{d-1}$, $H_v(0) \leq 1/2$ and there are absolute constants $c_0 > 0$ and $c_1 > 0$ so that for all $0 < r < c_0$, $f_v(v) \geq c_1$ then Assumption 4 holds (that is (22) holds) and so does Theorem 4.

Note that in Proposition 5, μ and Σ do not have to be the mean and covariance matrix of the $\tilde{X}_i$'s. In that case, μ and Σ are sometimes called location and scale and so Theorem 4 still apply for the robust to adversarial contamination and heavy-tail estimation of location, even in situations where there is not even a first moment.

Proposition 5 gives an alternative to Gaussian approximation which does not, in general, allow to check Assumption 4 because of the residual terms in Esseen or Berry-Esseen type inequalities. The assumptions in Proposition 5 are all granting that the density functions f_v are locally lower bounded around the 'critical' 1/4 and 3/4 quartiles and medians. They are natural assumptions that already appeared in several studies of estimators based on the SDO. In Proposition 5 we show that by using the median-of-means principle these assumptions are dealing with the density functions on the bucketed means and not the data themselves. However, Proposition 5 may also be applied in the $K = N$ case as for elliptically contoured distributions.

6 Conclusion

We showed that it is possible to estimate a mean vector in $\mathbb{R}^d$ w.r.t. the metric $\left\| \Sigma^{-1/2} \cdot \right\|_2$ even though Σ is unknown, the data set is corrupted by an adversary and the data are heavy-tailed. The rate obtained are the (deviation) minmax one in the ideal i.i.d. Gaussian case. The estimator used to achieve this rate is a deepest point with respect a median-of-means version of the Stahel-Donoho outlyingness functional. When the data are spherical enough there is no need to bucket the data and then the estimator is using the classical Stahel-Donoho outlyingness. Our analysis shows that the two cases can be handled using the same methodology and that the family of cdfs $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$ plays a key role in this analysis, in particular, their behavior around 0, the median and the 1/4 and 3/4 quartiles.

In this work, we have not deal with several research opportunities opened by the SDO. We now list some of them that may be considered in future works. a) It may look possible to use the isomorphic properties of the $MOMAD_K$ and SDO_K to study the Stahel-Donoho estimator (SDE) or a median-of-means version of the SDE defined as

$$\tilde{\mu}_{MOM,K}^{SDE} = \frac{\sum_{k=1}^{K} \hat{w}_k \bar{X}_k}{\sum_{k=1}^{K} \hat{w}_k} \tag{26}$$

where $(\hat{w}_k)_{k=1}^{K}$ are non-negative weights such that $\hat{w}_k$ depends on the outlyingness of the k-th bucketed mean $\bar{X}_k$. For instance,

$$\hat{w}_k = \begin{cases} 1 & \text{if } SDO_K(\bar{X}_k) \leq \hat{\alpha}_K \\ 0 & \text{otherwise.} \end{cases} \quad \text{where } \hat{\alpha}_k = \text{Med}(SDO_K(\bar{X}_k)). \tag{27}$$

b) Similarly, the isomorphic or almost-isometry properties of $MOMAD_K$ and SDO_K may also be used to study the properties of a MOM version of the SDE of the covariance matrix:

$$\hat{\Sigma} = \frac{2}{K} \sum_{k=1}^{K} \hat{w}_k (\bar{X}_k - \tilde{\mu}_{MOM,K}^{SDE})(\bar{X}_k - \tilde{\mu}_{MOM,K}^{SDE})^\top.$$

c) From a computational point of view, it is still an open question to construct an approximate solution to the SDO. The original or MOM version of the Stahel-Donoho median could be approximated via a robust gradient descent algorithms such as the one introduced in [10, 15, 39] with some extra normalization step required by the MAD denominator. We expect this algorithm to be more efficient than the classical weighted SDE because we expect to do only $\log d$ iterations to achieve a subgaussian estimator using a robust gradient descent algorithm whereas the SDE would require to approximate the K depths $SDO_K(\bar{X}_k), k \in [K]$ and should therefore require more computational time (note that, in practice the SDE has been reported to be more efficient than the deepest data that is the data $\bar{X}_k$ with the smallest $SDO_K(\bar{X}_k)$ but the SDE was not compared with an approximate solution of $\hat{\mu}_{MOM,K}^{SDO}$).

7 Proofs

In this section, we provide some proofs of all the results from the preceding sections. The only complexity measure we are using in this work is the Vapnik and Chervonenkis (VC) dimension [69, 70] of a class $\mathcal{F}$ of Boolean functions, i.e. of functions from $\mathbb{R}^d$ to $\{0,1\}$ in our case. We recall that $VC(\mathcal{F})$ is the maximal integer n such that there exists $x_1, \ldots, x_n \in \mathbb{R}^d$ for which the set $\{(f(x_1), \cdots, f(x_n)) : f \in \mathcal{F}\}$ is of maximal cardinality that is 2^n. The only VC-dimension we will use is the one of the set of all indicators of half affine spaces in $\mathbb{R}^d$: $VC(\{x \in \mathbb{R}^d \to I(\langle \cdot, v \rangle \geq r) : v \in \mathbb{R}^d, r \in \mathbb{R}\}) = d + 1$ (see Example 2.6.1 in [68]). The main technical tool (see Chapter 3 in [36]) we will be using is the following one: let $Y_1, \ldots, Y_n$ be independent random vectors in $\mathbb{R}^d$, there exists an absolute constant C_0 such that for all $u > 0$, with probability at least $1 - \exp(-u)$,

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(Y_i) - \mathbb{E}f(Y_i) \right) \leq C_0 \left(\sqrt{\frac{VC(\mathcal{F})}{n}} + \sqrt{\frac{u}{n}} \right). \tag{28}$$

We recall that for all $v \in \mathcal{S}_2^{d-1}$, $K \in [N]$ and $r > 0$,

$$H_{N,K,v}(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle \geq r \right].$$

The rate of convergence we will obtained is any r^* satisfying

$$C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + \sup_{\|v\|_2 = 1} H_{N,K,v}(r^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{29}$$

where C_0 is the constant from (28) and for some choice of K and u specified in each result depending on the set of assumptions.

7.1 Proof of Proposition 2 and 1 (first part): isomorphic property of MOMAD

We first prove Proposition 2 – the proof of Proposition 1 is a straighforward application of Proposition 2.

Proof of Proposition 2. We first observe that by renormalization, it is enough to show that for all $v \in \mathcal{S}_2^{d-1}$,

$$\varphi_l(\epsilon) \leq \text{Med}(|\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle - \text{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)|) \leq \varphi_u(\epsilon). \tag{30}$$

Moreover, for all $i \in [N]$, $\Sigma^{-1/2}(\tilde{X}_i - \mu)$ has mean zero and covariance I_d. Hence, without loss of generality we assume that $\mu = 0$ and $\Sigma = I_d$.

The strategy we are using to prove (30) is the following one. Let K real numbers $a_1, \ldots, a_K$ be given and denote by $a_{(1)} \leq \cdots \leq a_{(K)}$ the non-decreasing rearrangement of the $(a_k)_k$ (this is the rearrangement of the a_k's and not of their absolute values). To prove a result like $\varphi_l(\epsilon) \leq \text{Med}(|a_k - \text{Med}(a_k)|) \leq \varphi_u(\epsilon)$, it is enough to show that $\varphi_l(\epsilon) \leq a_{(3(K+1)/4)} - a_{((K+1)/2)} \leq \varphi_u(\epsilon)$ and $\varphi_l(\epsilon) \leq a_{((K+1)/2)} - a_{((K+1)/4)} \leq \varphi_u(\epsilon)$. As a consequence, to prove a result like (30), we should study the rearrangement (the two quartiles and the median) of the $\langle \bar{X}_k, v \rangle, k \in [K]$ uniformly over all $v \in \mathcal{S}_2^{d-1}$. But, $|\mathcal{O}|$ elements among the X_i's come from the adversary and we do not have any control on their behavior. We therefore have to consider the worst possible case which is when $|\mathcal{O}|$ bucketed means $\bar{X}_k$ are corrupted by one outliers from $\{X_i : i \in \mathcal{O}\}$. However, one may check that if we change $|\mathcal{O}|$ points in a set $\{a_k : k \in [K]\}$ to get a new set $\{A_k : k \in [K]\}$ then $\varphi_l(\epsilon) \leq a_{(3(K+1)/4)} - a_{((K+1)/2)} \leq \varphi_u(\epsilon)$ will be true if we show that $\varphi_l(\epsilon) \leq A_{(3(K+1)/4-|\mathcal{O}|)} - A_{((K+1)/2+|\mathcal{O}|)}$ and $A_{(3(K+1)/4+|\mathcal{O}|)} - A_{((K+1)/2-|\mathcal{O}|)} \leq \varphi_u(\epsilon)$ – and a similar observation holds for the other $(1/4)$-quartile. We will therefore first study the rearrangement of the original (i.e.

non corrupted) bucketed means (later denoted by $\overline{X}_k, k \in [K]$) projected on all one dimensional directions uniformly over these directions to deduce the result from (30) on the corrupted bucketed means $\bar{X}_k$.

We denote by $\overline{\tilde{X}}_k, k \in [K]$ the bucketed means of the original (non corrupted) dataset, i.e. $\overline{\tilde{X}}_k = (1/|B_k|) \sum_{i \in B_k} \tilde{X}_i$ for $k \in [K]$. To prove (30) we first study the rearrangements of vectors $(\langle \overline{\tilde{X}}_k, v \rangle)_{k \in [K]}$ uniformly over all $v \in \mathcal{S}_2^{d-1}$. We will then deal with the adversarial corruption to get (30).

We introduce the following supremum of empirical process:

$$Z = \sup_{\ell \in [K-1]} \sup_{\|v\|_2 = 1} \left| \frac{1}{K} \sum_{k=1}^{K} I \left(\langle \overline{\tilde{X}}_k, v \rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}} \right) - \mathbb{P} \left[\langle \overline{\tilde{X}}_k, v \rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}} \right] \right|$$

where W_v has been defined in Definition 1. It follows from (28) that for all $u > 0$, with probability at least $1 - \exp(-u)$, $Z \leq C_0 \left(\sqrt{(d+1)/K} + \sqrt{u/K} \right)$ (note that even though the function W_v depends on v, the boolean function $x \to I(\langle x, v \rangle \geq W_v(\ell/N))$ is still the indicator of a affine half-space of $\mathbb{R}^d$ for all $v \in \mathbb{R}^d$ and all $\ell \in [K-1]$ and thus the VC dimension of the set of Boolean functions $\{x \to I(\langle x, v \rangle \geq W_v(\ell/K)) : v \in \mathbb{R}^d, \ell \in [K-1]\}$ is less or equal to $d+1$). As a consequence, for some choice of $0 < \epsilon < 1/8$ such that Assumption 3 holds, if $K \geq (2C_0)(d+1)\epsilon^{-2}$ then with probability at least $1 - \exp(-\epsilon^2 K/(2C_0)^2)$, $Z \leq \epsilon$. Let us denote by Ω_ϵ the event onto which $Z \leq \epsilon$; we proved that $\mathbb{P}[\Omega_\epsilon] \geq 1 - \exp(-\epsilon^2 K/(2C_0)^2)$.

Let us place ourselves on the event Ω_ϵ up to the end of the proof. Since for all $v \in \mathcal{S}_2^{d-1}$,

$$\mathbb{P} \left[\langle \overline{\tilde{X}}_k, v \rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}} \right] = H_v(W_v(\ell/K)) = \ell/K,$$

(by left continuity of H_v we have $H_v(W_v(p)) = p$ for all $p \in (0,1)$), we have for all $\ell \in [K]$ and $v \in \mathcal{S}_2^{d-1}$, that

$$\left| \left\{ k \in [K] : \langle \overline{\tilde{X}}_k, v \rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}} \right\} \right| \in [\ell - \epsilon K, \ell + \epsilon K]. \tag{31}$$

This last result on the uniform in $v \in \mathcal{S}_2^{d-1}$ rearrangement of $(\langle \overline{\tilde{X}}_k, v \rangle)_k$ will be used to get the desire result on the rearrangement for $(\langle \overline{X}_k, v \rangle)_k$ (uniformly in v). To go from the $\overline{\tilde{X}}_k$'s to the $\overline{X}_k$'s we now have to deal with the adversarial corruption.

Since, there are $|\mathcal{O}|$ original data that may have been modified by the adversary, in the worse case $|\mathcal{O}|$ bucketed means $\overline{\tilde{X}}_k$ may be considered as corrupted and so, from the above cardinality estimation result (31), we may only certify (on Ω_ϵ) that

$$\left| \left\{ k \in [K] : \langle \bar{X}_k, v \rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}} \right\} \right| \in [\ell - \epsilon K - |\mathcal{O}|, \ell + \epsilon K + |\mathcal{O}|] \subset [\ell - 2\epsilon K, \ell + 2\epsilon K]$$

on the K bucketed means $\bar{X}_k$ constructed from the adversarialy corrupted dataset $\{X_i : i \in [N]\}$. We used here the assumption that $|\mathcal{O}| \leq \epsilon K$. If follows from the later result that if we denote by $q_{K,v}^{1/4}$ the $1/4$ quartile of vector $(\langle \bar{X}_k, v \rangle : k \in [K])$, by $q_{K,v}^{1/2}$ its median and by $q_{K,v}^{3/4}$ its $3/4$ quartile then,

$$\sqrt{\frac{K}{N}} W_v \left(\frac{3}{4} + 2\epsilon \right) \leq q_{K,v}^{1/4} \leq \sqrt{\frac{K}{N}} W_v \left(\frac{3}{4} - 2\epsilon \right); \sqrt{\frac{K}{N}} W_v \left(\frac{1}{2} + 2\epsilon \right) \leq q_{K,v}^{1/2} \leq \sqrt{\frac{K}{N}} W_v \left(\frac{1}{2} - 2\epsilon \right)$$

and

$$\sqrt{\frac{K}{N}} W_v \left(\frac{1}{4} + 2\epsilon \right) \leq q_{K,v}^{3/4} \leq \sqrt{\frac{K}{N}} W_v \left(\frac{1}{4} - 2\epsilon \right).$$

It follows from these inequalities that on the event Ω_ϵ, we have for all $v \in \mathcal{S}_2^{d_1}$,

$$\mathrm{Med}(|\langle \bar{X}_k, v \rangle - \mathrm{Med}(\langle \bar{X}_k, v \rangle)|) \leq \sqrt{\frac{K}{N}} \max \left(W_v \left(\frac{1}{4} - 2\epsilon \right) - W_v \left(\frac{1}{2} + 2\epsilon \right), W_v \left(\frac{1}{2} - 2\epsilon \right) - W_v \left(\frac{3}{4} + 2\epsilon \right) \right)$$

and

$$\mathrm{Med}(|\langle \bar{X}_k, v \rangle - \mathrm{Med}(\langle \bar{X}_k, v \rangle)|) \geq \sqrt{\frac{K}{N}} \min \left(W_v \left(\frac{1}{4} + 2\epsilon \right) - W_v \left(\frac{1}{2} - 2\epsilon \right), W_v \left(\frac{1}{2} + 2\epsilon \right) - W_v \left(\frac{3}{4} - 2\epsilon \right) \right).$$

The result follows from the definition of $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ in Assumption 3. ∎

7.2 Proof of Proposition 3 and 1 (second part): isomorphic property of SDO_K.

The proof of Proposition 3 and 1 (second part) relies on the next result.

Proposition 6. *We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Let $K \in [N]$, $u > 0$ and r^* be such that (29) holds. Then, with probability at least $1 - \exp(-u)$,*

$$\sup_{v \in \mathcal{S}_2^{d-1}} |\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)| \leq \sqrt{\frac{K}{N}} r^*.$$

Proof of Proposition 6. Denote by $\mathcal{K} = \{k : B_k \cap \mathcal{O} = \emptyset\}$ the set of indices of non-corrupted blocks of data. It follows from (28) and the definition of r^* that with probability at least $1 - \exp(-u)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\frac{1}{K} \sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right)$$

$$= \frac{1}{K} \sum_{k \in \mathcal{K}} I\left(\langle \Sigma^{-1/2}(\widetilde{\bar{X}}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right) + \frac{1}{K} \sum_{k \in \mathcal{K}^c} I\left(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right)$$

$$\leq \frac{1}{K} \sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\widetilde{\bar{X}}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right) + \frac{|\mathcal{O}|}{K}$$

$$\leq \sup_{\|v\|_2 = 1} \left(\frac{1}{K} \sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\widetilde{\bar{X}}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right) - P\left(\langle \Sigma^{-1/2}(\widetilde{\bar{X}}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right) \right)$$

$$+ P\left(\langle \Sigma^{-1/2}(\widetilde{\bar{X}}_1 - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right) + \frac{|\mathcal{O}|}{K}$$

$$\leq C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + H_{N,K,v}(r^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}.$$

As a consequence, with probability at least $1 - \exp(-u)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle \geq \frac{r^*}{\sqrt{N/K}} \right) < \frac{K}{2}$$

and so

$$\sup_{v \in \mathcal{S}_2^{d-1}} |\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)| \leq \sqrt{\frac{K}{N}} r^*. \tag{32}$$

Remark 1. *It is also possible to consider a "directional version" of Proposition 6 if one defines a "directional version" of r^*, that is for all directions $v \in \mathcal{S}_2^{d-1}$, define $r_v^* > 0$ satisfying*

$$C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + H_{N,K,v}(r_v^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}.$$

Then, under the same conditions as in Proposition 6, we have with probability at least $1 - \exp(-u)$,

$$\sup_{v \in \mathcal{S}_2^{d-1}} \frac{|\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)|}{r_v^*} \leq \sqrt{\frac{K}{N}}.$$

Hence, Proposition 6 holds as well for $r^* = \sup_{\|v\|_2 = 1} r_v^*$. Note that for most of the $v \in \mathcal{S}_2^{d-1}$ the values of r_v^* is expected to be much smaller than r^*. For instance, for vectors v well-spread, we expect them to have a strong "mixing" power (see for instance "super-Gaussian directions" in [35] or [33, 34]).

Proof of Proposition 3 and 1. It follows from Proposition 6 and Proposition 2 that, with probability at least $1 - \exp(-u) - \exp(-c_1\epsilon^2 K)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$|\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)| \leq \sqrt{\frac{K}{N}} r^*$$

and

$$\varphi_l(\epsilon)\sqrt{\frac{K}{N}} \left\| \Sigma^{1/2}v \right\|_2 \leq MOMAD_K(v) \leq \varphi_u(\epsilon)\sqrt{\frac{K}{N}} \left\| \Sigma^{1/2}v \right\|_2.$$

We denote by Ω_0 the event onto which the last two properties hold. On the even Ω_0, for all $\nu \in \mathbb{R}^d$, we have

$$SDO_K(\nu) = \sup_{v \in \mathbb{R}^d} \frac{|\operatorname{Med}(\langle \bar{X}_k - \nu, v \rangle)|}{MOMAD_K(v)} \leq \sup_{v \in \mathbb{R}^d} \frac{|\operatorname{Med}(\langle \bar{X}_k - \nu, v \rangle)|}{\varphi_l(\epsilon)\sqrt{K/N}\left\|\Sigma^{1/2}v\right\|_2} = \sup_{v \in \mathcal{S}_2^{d-1}} \frac{|\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \nu), v \rangle)|}{\varphi_l(\epsilon)\sqrt{K/N}}$$

$$\leq \sup_{v \in \mathbb{R}^d} \frac{|\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)| + |\langle \Sigma^{-1/2}(\nu - \mu), v \rangle|}{\varphi_l(\epsilon)\sqrt{K/N}} \leq \sup_{v \in \mathbb{R}^d} \frac{\sqrt{K/N}r^* + |\langle \Sigma^{-1/2}(\nu - \mu), v \rangle|}{\varphi_l(\epsilon)\sqrt{K/N}}$$

$$\leq \begin{cases} \frac{3\left\|\Sigma^{-1/2}(\nu-\mu)\right\|_2}{2\varphi_l(\epsilon)\sqrt{K/N}} & \text{if } \left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 \geq 2\sqrt{K/N}r^* \\ 3r^*/\varphi_l(\epsilon) & \text{otherwise} \end{cases}$$

and when $\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 \geq 2\sqrt{K/N}r^*$, we have

$$SDO_K(\nu) \geq \sup_{v \in \mathcal{S}_2^{d-1}} \frac{|\langle \Sigma^{-1/2}(\nu - \mu), v \rangle| - |\operatorname{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)|}{\varphi_u(\epsilon)\sqrt{K/N}} \geq \frac{\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2}{2\varphi_u(\epsilon)\sqrt{K/N}}.$$

∎

7.3 Proof of the statistical bounds

Proof of Proposition 1. Proposition 1 is a corollary of Proposition 2 for $K = N$. For this choice of K, there are N blocks, each containing only one data and so $MOMAD_N(v) = MAD(v)$ for all $v \in \mathbb{R}^d$. The only thing that remains to be checked is the validity of Assumption 3 in the Gaussian case and the dependency of the $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ in terms of ϵ.

When the original data $\tilde{X}_i, i \in [N]$ are N i.i.d. Gaussian vectors $G_1, \ldots, G_N$ with mean μ and covariance matrix Σ then for all $K \in [N]$, $(1/\sqrt{N/K})\sum_{i=1}^{N/K} \Sigma^{-1/2}(\tilde{X}_i - \mu)$ is a standard Gaussian vector in $\mathbb{R}^d$. Therefore the $H := H_{N,K,v}$ function from Assumption 3 is equal to the function $x \in \mathbb{R} \to 1 - \Phi(x)$ where $\Phi : x \in \mathbb{R} \to \mathbb{P}[g \leq x]$ is the cdf of a standard Gaussian variable $g \sim \mathcal{N}(0,1)$ in $\mathbb{R}$. This holds for all N, K and $v \in \mathcal{S}_2^{d-1}$, that is $H_{N,K,v}$ is independent of N, K and $v \in \mathcal{S}_2^{d-1}$. Since $W := W_{N,K,v}$ is the generalized inverse of H, in the Gaussian case, we obtain that $W(p) = \Phi^{-1}(1 - p)$ for all $p \in (0,1)$. It follows from Lemma 5.2 in [61] that there exists some absolute constant $C_1 > 0$ such that

$$\min\left(W\left(\frac{1}{4} + 2\epsilon\right) - W\left(\frac{1}{2} - 2\epsilon\right), W\left(\frac{1}{2} + 2\epsilon\right) - W\left(\frac{3}{4} - 2\epsilon\right) \right) \geq \Phi^{-1}(3/4) - C_1\epsilon := \varphi_l(\epsilon)$$

and

$$\max\left(W\left(\frac{1}{4} - 2\epsilon\right) - W\left(\frac{1}{2} + 2\epsilon\right), W\left(\frac{1}{2} - 2\epsilon\right) - W\left(\frac{3}{4} + 2\epsilon\right) \right) \leq \Phi^{-1}(1/4) \leq \Phi^{-1}(3/4) + C_1\epsilon := \varphi_u(\epsilon).$$

As a consequence, Assumption 3 holds in the Gaussian case for all $0 < \epsilon < \Phi^{-1}(3/4)/C_1$ with $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - C_1\epsilon$ and $\varphi_u(\epsilon) = \Phi^{-1}(3/4) + C_1\epsilon$. ∎

Proofs of theorems 1, 2 and 4 Theorems 1, 2 and 4 are corollaries of a general result that we are stating now.

Theorem 5. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 3 holds for some $0 < \epsilon < 1/4$ and constants $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$. We assume that the adversarial contamination with L_2 inliers model from Assumption 1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Let $K \geq \max(\epsilon^{-1}|\mathcal{O}|, c_0\epsilon^{-2}d)$, $0 < u < c_0\epsilon^2 K$ and r^* be such that (29) holds. Then, with probability at least $1 - 2\exp(-u)$,*

$$\left\| \Sigma^{-1/2}(\hat{\mu}_{MOM,K}^{SDO} - \mu) \right\|_2 \leq \frac{2\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{K}{N}}r^*.$$

Proof of Theorem 1. There exists an absolute constant c_0 such that for all $0 \leq r \leq c_0, \mathbb{P}[g \geq r] \leq 1/2 - 2r$ where $g \sim \mathcal{N}(0,1)$. Moreover, for all $K \in [N], v \in \mathcal{S}_2^{d-1}$ and $r > 0$, we have $H_{N,K,v}(r) = \mathbb{P}[g \geq r]$. As a consequence, one can chose r^*, u and K such that

$$r^* = C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + \frac{|\mathcal{O}|}{K}$$

as long as this later quantity is less or equal to c_0. Finally, we apply Theorem 5 for $K = N$ and the result follows since $\hat{\mu}_{MOM,N}^{SDO} = \hat{\mu}^{SDO}$. ∎

Proof of Theorem 2. It follows from Markov's inequality (14) that we can chose u, r^* and K such that

$$C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + \frac{1}{1 + (r^*)^2} + \frac{|\mathcal{O}|}{K} < \frac{1}{2}$$

for instance by taking $r^* = 2$, $K \geq 4|\mathcal{O}|$, $K > 16C_0^2(d+1)$ and $K > 16C_0 u$. Note however, that because r^* is constant, the convergence rate is proportional to $\sqrt{K/N}$, in particular it does not depend on u. Hence there is no interest to consider values of u smaller than K (up to constant). We therefore apply Theorem 5 for this choice of K, $u = c_0\epsilon^2 K$ and $r^* = 2$. ∎

Proof of Theorem 4. Thanks to Assumption 4, there exists absolute constants $c_0 > 0$ and $c_1 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and $(2C_0/c_1)\sqrt{(d+1)/K} \leq r \leq c_0, H_{N,K,v} \leq 1/2 - c_1 r$. As a consequence, one can chose r^*, u and K such that

$$r^* = \frac{2}{c_1} \left(C_0 \left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}} \right) + \frac{|\mathcal{O}|}{K} \right)$$

as long as this later quantity is less or equal to c_0. Finally, we apply Theorem 5 for this choice of K, u and r^*. ∎

Proof of Theorem 5. We first note that a proof of Theorem 5 may follow from the isomorphic property of SDO_K from Proposition 3. However, it is possible to improve constants by using the following strategy.

Let us place ourselves on the intersection of the two events where the results of both Proposition 2 and Proposition 6 hold. We set $f : v \in \mathbb{R}^d \to \text{Med}(\langle \bar{X}_k, v \rangle)$. Since f is symmetric we have

$$\left\| \Sigma^{-1/2}(\hat{\mu}_{MOM,K}^{SDO} - \mu) \right\|_2 = \sup_{\|v\|_2 = 1} \langle \Sigma^{-1/2}(\hat{\mu}_{MOM,K}^{SDO} - \mu), v \rangle = \sup_{v \in \mathbb{R}^d} \langle \hat{\mu}_{MOM,K}^{SDO} - \mu, \frac{v}{\left\| \Sigma^{1/2}v \right\|_2} \rangle$$

$$= \sup_{v \in \mathbb{R}^d} \frac{\langle \hat{\mu}_{MOM,K}^{SDO}, v \rangle - f(v) + f(v) - \langle \mu, v \rangle}{MOMAD_K(v)} \frac{MOMAD_K(v)}{\left\| \Sigma^{1/2}v \right\|_2}$$

$$\leq \left(\sup_{v \in \mathbb{R}^d} \frac{\langle \hat{\mu}_{MOM,K}^{SDO}, v \rangle - f(v)}{MOMAD_K(v)} + \sup_{v \in \mathbb{R}^d} \frac{f(v) - \langle \mu, v \rangle}{MOMAD_K(v)} \right) \sup_{v \in \mathbb{R}^d} \frac{MOMAD_K(v)}{\left\| \Sigma^{1/2}v \right\|_2}$$

$$\leq \left(SDO_K(\hat{\mu}_{MOM,K}^{SDO}) + SDO_K(\mu) \right) \sup_{v \in \mathbb{R}^d} \frac{MOMAD_K(v)}{\left\| \Sigma^{1/2}v \right\|_2} \leq 2SDO_K(\mu) \sup_{v \in \mathbb{R}^d} \frac{MOMAD_K(v)}{\left\| \Sigma^{1/2}v \right\|_2}.$$

where we used that $SDO_K(\hat{\mu}_{MOM,K}^{SDO}) \leq SDO_K(\mu)$ by definition of $\hat{\mu}_{MOM,K}^{SDO}$.

We know how to control $\sup_{v \in \mathbb{R}^d} MOMAD_K(v)/\left\| \Sigma^{1/2}v \right\|_2$ by $\sqrt{K/N}\varphi_u(\epsilon)$ using Proposition 2. It remains to control the term $SDO_K(\mu)$. We have

$$SDO_K(\mu) = \sup_{v \in \mathbb{R}^d} \frac{|\langle \mu, v \rangle - \text{Med}(\langle \bar{X}_k, v \rangle)|}{\text{Med}(|\langle \bar{X}_k, v \rangle - \text{Med}(\langle \bar{X}_k, v \rangle)|)} = \sup_{v \in \mathbb{R}^d} \frac{|\text{Med}(\langle \mu - \bar{X}_k, v \rangle)|}{\left\| \Sigma^{1/2}v \right\|_2} \frac{\left\| \Sigma^{1/2}v \right\|_2}{MOMAD_K(v)}$$

$$\leq \sup_{\|v\|_2 = 1} |\text{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)| \sup_{v \in \mathbb{R}^d} \frac{\left\| \Sigma^{1/2}v \right\|_2}{MOMAD_K(v)}.$$

The term $\sup_{v \in \mathbb{R}^d} \left\| \Sigma^{1/2}v \right\|_2 / MOMAD_K(v)$ is smaller than $\sqrt{N/K}/\varphi_l(\epsilon)$ thanks to Proposition 2. Finally, to finish the proof, we upper bound the term $\sup_{\|v\|_2 = 1} |\text{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)|$ by $\sqrt{K/N}r^*$ thanks to Proposition 6. ∎

Proof of Theorem 3 For all $k \in [N]$, we set $\hat{\mu}_k = \hat{\mu}_{MOM,k}^{SDO}$ we denote by Ω_k the event onto which

$$\left\| \Sigma^{-1/2}(\hat{\mu}_k - \mu) \right\|_2 \leq \frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \sqrt{\frac{k}{N}}$$

and, for all $\nu \in \mathbb{R}^d$, if $\left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2 \geq 6\sqrt{k/N}$ then

$$\frac{\left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2}{2\varphi_u(\epsilon)\sqrt{k/N}} \leq SDO_k(\nu) \leq \frac{3 \left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2}{2\varphi_l(\epsilon)\sqrt{k/N}}$$

and if $\left\| \Sigma^{-1/2}(\nu - \mu) \right\|_2 \leq 6\sqrt{k/N}$ then

$$SDO_k(\nu) \leq \frac{9}{\varphi_l(\epsilon)}.$$

It follows from Proposition 3 for $r^* = 3$ and $u = K/(16C_0^2)$ and Theorem 2 that $\mathbb{P}[\Omega_k] \geq 1 - 3\exp(-c_1\epsilon^2 k)$ when $k \geq \max(|\mathcal{O}|/\epsilon, c_0 d/\epsilon^2)$.

Let $K \geq \max(|\mathcal{O}|/\epsilon, c_0 d/\epsilon^2)$. On the even $\cap_{k=K}^N \Omega_k$, we have for all $K \leq k \leq N$,

$$SDO_k(\hat{\mu}_K - \hat{\mu}_k) \leq \max\left(\frac{9}{\varphi_l(\epsilon)}, \frac{3 \left\| \Sigma^{-1/2}(\hat{\mu}_K - \hat{\mu}_k) \right\|_2}{2\varphi_l(\epsilon)\sqrt{k/N}} \right) \leq \max\left(\frac{9}{\varphi_l(\epsilon)}, \frac{6\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)} \left(1 + \sqrt{\frac{K}{k}} \right) \right)$$

and so, by definition of $\hat{K}$, we have $\hat{K} \leq K$. We also have by definition of $\hat{K}$ and because $\hat{K} \leq K$ that

$$SDO_K(\hat{\mu}_{\hat{K}} - \hat{\mu}_K) \leq \max\left(\frac{9}{\varphi_l(\epsilon)}, \frac{6\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)} \left(1 + \sqrt{\frac{\hat{K}}{K}} \right) \right) \leq \frac{12\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)}.$$

We conclude that either $\left\| \Sigma^{-1/2}(\hat{\mu}_{\hat{K}} - \hat{\mu}_K) \right\|_2 \leq 6\sqrt{K/N}$ and so

$$\left\| \Sigma^{-1/2}(\hat{\mu}_{\hat{K}} - \mu) \right\|_2 \leq \left\| \Sigma^{-1/2}(\hat{\mu}_{\hat{K}} - \hat{\mu}_K) \right\|_2 + \left\| \Sigma^{-1/2}(\mu - \hat{\mu}_K) \right\|_2 \leq \left(6 + \frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \right) \sqrt{\frac{K}{N}}$$

or $\left\| \Sigma^{-1/2}(\hat{\mu}_{\hat{K}} - \hat{\mu}_K) \right\|_2 \geq 6\sqrt{K/N}$ and so

$$\left\| \Sigma^{-1/2}(\hat{\mu}_{\hat{K}} - \mu) \right\|_2 \leq \left\| \Sigma^{-1/2}(\hat{\mu}_{\hat{K}} - \hat{\mu}_K) \right\|_2 + \left\| \Sigma^{-1/2}(\hat{\mu}_K - \mu) \right\|_2$$

$$\leq SDO_K(\hat{\mu}_{\hat{K}} - \hat{\mu}_K)2\varphi_u(\epsilon)\sqrt{\frac{K}{N}} + \frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{K}{N}} \leq \frac{28\varphi_u^2(\epsilon)}{\varphi_l^2(\epsilon)}\sqrt{\frac{K}{N}}.$$

∎

Proof of Proposition 4. We have for all $i, j \in [d]$, $\left| \phi_0^2 \Sigma_{ij} - \hat{\Sigma}_{ij} \right| \leq c_1\epsilon(c_1\epsilon + \phi_0)(\Sigma_{ii} + \Sigma_{ij})$ because, it follows from (18) that for all $v \in \mathbb{R}^d$,

$$\left| MOMAD_K^2(v) - \phi_0^2 \frac{K}{N} \left\| \Sigma^{1/2}v \right\|_2^2 \right| = \left| MOMAD_K(v) - \phi_0\sqrt{\frac{K}{N}} \left\| \Sigma^{1/2}v \right\|_2 \right| \left(MOMAD_K(v) + \phi_0\sqrt{\frac{K}{N}} \left\| \Sigma^{1/2}v \right\|_2 \right)$$

$$\leq c_1\epsilon\frac{K}{N} \left\| \Sigma^{1/2}v \right\|_2^2 (c_1\epsilon + \phi_0).$$

Next, we have for all $u, v \in \mathbb{R}^d$ such that $\|u\|_1 = \|v\|_1 = 1$

$$|\langle u, (\phi_0^2 \Sigma - \hat{\Sigma})v \rangle|$$

$$= \frac{N}{4K} \left| \sum_{i,j} u_i v_j \left(\phi_0^2 \frac{K}{N} \left\| \Sigma^{1/2}(e_i + e_j) \right\|_2^2 - MOMAD_K^2(e_i + e_j) + \phi_0^2 \frac{K}{N} \left\| \Sigma^{1/2}(e_i - e_j) \right\|_2^2 - MOMAD_K^2(e_i - e_j) \right) \right|$$

$$\leq \frac{c_1\epsilon(c_1\epsilon + \phi_0)}{4} \sum_{i,j} |u_i||v_j| \left(\left\| \Sigma^{1/2}(e_i + e_j) \right\|_2^2 + \left\| \Sigma^{1/2}(e_i - e_j) \right\|_2^2 \right) = \frac{c_1\epsilon(c_1\epsilon + \phi_0)}{2} \sum_{i,j} |u_i||v_j| (\Sigma_{ii} + \Sigma_{jj}).$$

∎

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147, 1999. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

[2] Imre Bárány and Nabil H Mustafa. An application of the universality theorem for tverberg partitions to data depth and hitting convex sets. *Computational Geometry*, page 101649, 2020.

[3] Włodzimierz Bryc. *The normal distribution*, volume 100 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995. Characterizations with applications.

[4] Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans. Inform. Theory*, 59(11):7711–7717, 2013.

[5] T. Tony Cai, Weidong Liu, and Harrison H. Zhou. Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *Ann. Statist.*, 44(2):455–488, 2016.

[6] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[7] Louis H. Y. Chen and Qi-Man Shao. A non-uniform Berry-Esseen bound via Stein's method. *Probab. Theory Related Fields*, 120(2):236–254, 2001.

[8] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 2018.

[9] Yanchu Chen and Qi-Man Shao. Berry-Esseen inequality for unbounded exchangeable pairs. In *Probability approximations and beyond*, volume 205 of *Lect. Notes Stat.*, pages 13–30. Springer, New York, 2012.

[10] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates, 2019.

[11] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using l1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

[12] Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *arXiv preprint arXiv:2002.01432*, 2020.

[13] P. L. Davies. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, 15(3):1269–1292, 1987.

[14] Michiel Debruyne. An outlier map for support vector machine classification. *Ann. Appl. Stat.*, 3(4):1566–1580, 2009.

[15] Jules Depersin and Guillaume Lecué. Fast algorithms for robust estimation of a mean vector. 2019.

[16] Jules Depersin and Guillaume Lecué. Convex programs and algorithms for robust subgaussian estimation of a mean vector with respect to any norm. Technical report, IPParis, Crest, ENSAE, 2020.

[17] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 12 2016.

[18] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016.

[19] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 655–664. IEEE Computer Soc., Los Alamitos, CA, 2016.

[20] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*, 2017.

[21] D Donoho. *L.(1982) Breakdown properties of multivariate location estimtors*. PhD thesis, Ph. D. Qualifying Paper, Dept. of Statistics, Harvard Univ.

[22] David Donoho and Peter J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA, 1983.

[23] David L Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford . . . , 1982.

[24] David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

[25] David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

[26] Ramanathan Gnanadesikan and John R Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pages 81–124, 1972.

[27] JBS Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 1948.

[28] Frank R. Hampel. Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 27:87–104, 1973.

[29] Frank R. Hampel. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393, 1974.

[30] Samuel B Hopkins. Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

[31] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[32] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

[33] B. Klartag and S. Sodin. Variations on the Berry-Esseen theorem. *Teor. Veroyatn. Primen.*, 56(3):514–533, 2011.

[34] Bo'az Klartag. A Berry-Esseen type inequality for convex bodies with an unconditional basis. *Probab. Theory Related Fields*, 145(1-2):1–33, 2009.

[35] Bo'az Klartag. Super-Gaussian directions of random vectors. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 187–211. Springer, Cham, 2017.

[36] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, Berlin, 2011.

[37] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *Ann. Statist.*, 48(2):906–931, 2020.

[38] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[39] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 2598–2612. PMLR, 2020.

[40] O. V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990.

[41] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991.

[42] M. Lerasle and R. Oliveira. Robust empirical mean estimators. Technical report, IMPA and CNRS, 2011.

[43] Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, and Guillaume Lecué. Monk outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, pages 3782–3793. PMLR, 2019.

[44] Regina Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18(1):405–414, 1990.

[45] Regina Y. Liu. Data depth and multivariate rank tests. In *L_1-statistical analysis and related methods (Neuchâtel, 1992)*, pages 279–294. North-Holland, Amsterdam, 1992.

[46] Regina Y. Liu and Kesar Singh. Ordering directional data: concepts of data depth on circles and spheres. *Ann. Statist.*, 20(3):1468–1484, 1992.

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

[48] Junwei Lu, Fang Han, and Han Liu. Robust scatter matrix estimation for high dimensional distributions with heavy tail. *IEEE transactions on information theory*, 2020.

[49] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019.

[50] Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields*, 175(3-4):957–973, 2019.

[51] Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

[52] Z. Szabo M. Lerasle, T. Matthieu and G. Lecué. Monk – outliers-robust mean embedding estimation by median-of-means. Technical report, CNRS, University of Paris 11, Ecole Polytechnique and CREST, 2017.

[53] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2006. Theory and methods.

[54] Ricardo A Maronna and Victor J Yohai. The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341, 1995.

[55] S Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. Technical report, arXiv: 1704.02658, 2017.

[56] Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[57] Stanislav Minsker. Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*, 2018.

[58] Stanislav Nagy, Carsten Schütt, Elisabeth M Werner, et al. Halfspace depth and floating body. *Statistics Surveys*, 13:52–118, 2019.

[59] A. S. Nemirovsky and D. B. and Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[60] Daniel Peña and Francisco J Prieto. Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, 16(1):228–254, 2007.

[61] Valentin V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

[62] Peter J Rousseeuw, Jakob Raymaekers, and Mia Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27(2):345–359, 2018.

[63] Werner A Stahel. *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, ETH Zurich, 1981.

[64] John W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531, 1975.

[65] David E. Tyler. Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.*, 22(2):1024–1044, 1994.

[66] Stefan Van Aelst. Stahel–donoho estimation for high-dimensional data. *International Journal of Computer Mathematics*, 93(4):628–639, 2016.

[67] Stefan Van Aelst, E Vandervieren, and Gert Willems. Stahel-donoho estimators with cellwise weights. *Journal of Statistical Computation and Simulation*, 81(1):1–27, 2011.

[68] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[69] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, Cham, 2015. Reprint of Theor. Probability Appl. **16** (1971), 264–280.

[70] Vladimir N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.

[71] Yijun Zuo, Hengjian Cui, and Xuming He. On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Ann. Statist.*, 32(1):167–188, 2004.