Geometric median and robust estimation in Banach spaces

STANISLAV MINSKER

Mathematics Department, Duke University, Box 90320, Durham, NC 27708-0320, USA.
E-mail: sminsker@math.duke.edu

In many real-world applications, collected data are contaminated by noise with heavy-tailed distribution and might contain outliers of large magnitude. In this situation, it is necessary to apply methods which produce reliable outcomes even if the input contains corrupted measurements. We describe a general method which allows one to obtain estimators with tight concentration around the true parameter of interest taking values in a Banach space. Suggested construction relies on the fact that the geometric median of a collection of independent "weakly concentrated" estimators satisfies a much stronger deviation bound than each individual element in the collection. Our approach is illustrated through several examples, including sparse linear regression and low-rank matrix recovery problems.

Keywords: distributed computing; heavy-tailed noise; large deviations; linear models; low-rank matrix estimation; principal component analysis; robust estimation

1. Introduction

Given an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}$ from a distribution Π with $\mathrm{Var}(X_1) < \infty$ and $t > 0$, is it possible to construct an estimator $\hat{\mu}$ of the mean $\mu = \mathbb{E}X_1$ which would satisfy

$$\Pr\left(|\hat{\mu} - \mu| > C\sqrt{\mathrm{Var}(X_1)\frac{t}{n}} \right) \leq \mathrm{e}^{-t} \tag{1.1}$$

for some absolute constant C without *any* extra assumptions on Π? What happens if the sample contains a fixed number of outliers of arbitrary nature? Does the estimator still exist?

A (somewhat surprising) answer is yes, and several ways to construct $\hat{\mu}$ are known. The earliest reference that we are aware of is the book by Nemirovski and Yudin [34], where related question was investigated in the context of stochastic optimization. We learned about problem (1.1) and its solution from the work of Oliveira and Lerasle [28] who used the ideas in spirit of [34] to develop the theory of "robust empirical mean estimators". Method described in [28] consists of the following steps: divide the given sample into $V \approx t$ blocks, compute the sample mean within each block and then take the median of these sample means. A relatively simple analysis shows that the resulting estimator indeed satisfies (1.1). Similar idea was employed earlier in the work of Alon, Matias and Szegedy [1] to construct randomized algorithms for approximating the so-called "frequency moments" of a sequence. Recently, the aforementioned "median of the means" construction appeared in [7] in the context of multi-armed bandit problem under weak assumptions on the reward distribution. A different approach to the question (1.1) (based on PAC-

Bayesian truncation) was given in [14]. A closely related independent recent work [18] applies original ideas of Nemirovski and Yudin to general convex loss minimization.

The main goal of this work is to design a general technique that allows construction of estimators satisfying a suitable version of (1.1) for Banach space-valued μ. To achieve this goal, we show that a collection of independent estimators of a Banach space-valued parameter can be transformed into a new estimator which preserves the rate and admits much tighter concentration bounds. The method we propose is based on the properties of a *geometric median*, which is one of the possible extensions of a univariate median to higher dimensions.

Many popular estimators (e.g., Lasso [42] in the context of high-dimensional linear regression) admit strong theoretical guarantees if the distribution of the noise satisfies restrictive assumptions (such as sub-Gaussian tails). An important question that we attempt to answer is: can one design algorithms which preserve nice properties of existing techniques and at the same time:

(1) admit strong performance guarantees under weak assumptions on the noise;
(2) are not affected by the presence of a fixed number of outliers of arbitrary nature and size;
(3) can be implemented in parallel for faster computation with large data sets.

Our results imply that in many important applications the answer is positive. In Section 4, we illustrate this assertion with several classical examples, including principal component analysis, sparse linear regression and low-rank matrix recovery. In each case, we present non-asymptotic probabilistic bounds describing performance of proposed methods.

For an overview of classical and modern results in robust statistics, see [19,20], and references therein. Existing literature contains several approaches to estimation in the presence of heavy-tailed noise based on the aforementioned estimators satisfying (1.1). However, most of the previous work concentrated on one-dimensional versions of (1.1) and used it as a tool to solve intrinsically high-dimensional problems. For example, in [28] authors develop robust estimator selection procedures based on the medians of empirical risks with respect to disjoint subsets of the sample. While this approach admits strong theoretical guarantees, it requires several technical assumptions that are not always easy to check it practice. Another related work [2] discusses robust estimation in the context of ridge regression. Proposed method is based on a "min–max" estimator which has good theoretical properties but can only be evaluated approximately based on heuristic methods. It is also not immediately clear if this technique can be extended to robust estimation in other frameworks. An exception is the approach described in [34] and further explored in [18], where authors use a version of the multidimensional median for estimator selection. However, this method has several weaknesses in statistical applications when compared to our technique; see Section 3 for more details and discussion. The main results of our work require minimal assumptions, apply to a wide range of models, and allow to use many existing algorithms as a subroutine to produce robust estimators which can be evaluated exactly via a simple iterative scheme.

2. Geometric median

Let $\mathbb{X}$ be a Banach space with norm $\|\cdot\|$, and let μ be a probability measure on $(\mathbb{X}, \|\cdot\|)$. Define the *geometric median* (also called the spatial median, Fermat–Weber point [45] or Haldane's

median [17]) of μ as

$$x_* = \arg\min_{y \in \mathbb{X}} \int_{\mathbb{X}} \big(\|y - x\| - \|x\| \big) \mu(\mathrm{d}x).$$

For other notions of the multidimensional median and a nice survey of the topic, see [41]. In this paper, we will only be interested in a special case when μ is the empirical measure corresponding to a finite collection of points $x_1, \ldots, x_k \in \mathbb{X}$, so that

$$x_* = \mathrm{med}(x_1, \ldots, x_k) := \arg\min_{y \in \mathbb{X}} \sum_{j=1}^{k} \|y - x_j\|. \tag{2.1}$$

Geometric median exists under rather general conditions. For example, it is enough to assume that $\mathbb{X} = \mathbb{Y}^*$, where $\mathbb{Y}$ is a separable Banach space and $\mathbb{Y}^*$ is its dual – the space of all continuous linear functionals on $\mathbb{Y}$. This includes the case when $\mathbb{X}$ is separable and reflexive, that is, $\mathbb{X} = (\mathbb{X}^*)^*$. Moreover, if the Banach space $\mathbb{X}$ is strictly convex (i.e., $\|x_1 + x_2\| < \|x_1\| + \|x_2\|$ whenever x_1 and x_2 are not proportional), then x_* is unique unless all the points $x_1, \ldots, x_n$ are on the same line. For proofs of these results, see [22]. Throughout the paper, it will be assumed that $\mathbb{X}$ is separable and reflexive.

In applications, we are often interested in the situation when $\mathbb{X}$ is a Hilbert space (in particular, it is reflexive and strictly convex) and $\| \cdot \|$ is induced by the inner product $\langle \cdot, \cdot \rangle$. In such cases, we will denote the ambient Hilbert space by $\mathbb{H}$.

The cornerstone of our subsequent presentation is the following lemma, which states that if a given point z is "far" from the geometric median $x_* = \mathrm{med}(x_1, \ldots, x_k)$, then it is also "far" from a constant fraction of the points $x_1, \ldots, x_k$. We will denote $F(y) := \sum_{j=1}^{k} \|y - x_j\|$.

Lemma 2.1.

(a) *Let $x_1, \ldots, x_k \in \mathbb{H}$ and let x_* be their geometric median. Fix $\alpha \in (0, \frac{1}{2})$ and assume that $z \in \mathbb{H}$ is such that $\|x_* - z\| > C_\alpha r$, where*

$$C_\alpha = (1 - \alpha)\sqrt{\frac{1}{1 - 2\alpha}} \tag{2.2}$$

and $r > 0$. Then there exists a subset $J \subseteq \{1, \ldots, k\}$ of cardinality $|J| > \alpha k$ such that for all $j \in J$, $\|x_j - z\| > r$.

(b) *For general Banach spaces, the claim holds with a constant $C_\alpha = \frac{2(1-\alpha)}{1-2\alpha}$.*

Proof. (a) Assume that the implication is not true. Without loss of generality, it means that $\|x_i - z\| \le r$, $i = 1, \ldots, \lfloor (1 - \alpha)k \rfloor + 1$.

Consider the directional derivative

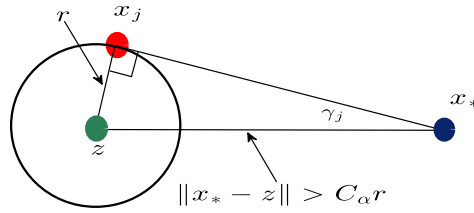$$DF(x_*; z - x_*) := \lim_{t \searrow 0} \frac{F(x_* + t(z - x_*)) - F(x_*)}{t}$$

Figure 1. Geometric illustration.

at the point x_* in direction $z - x_*$. Since x_* minimizes F over $\mathbb{H}$, $DF(x_*; z - x_*) \geq 0$. On the other hand, it is easy to see that

$$\frac{DF(x_*; z - x_*)}{\|z - x_*\|} = -\sum_{j: x_j \neq x_*} \frac{\langle x_j - x_*, z - x_* \rangle}{\|x_j - x_*\| \|z - x_*\|} + \sum_{j=1}^{k} I\{x_j = x_*\}. \tag{2.3}$$

For $j = 1, \ldots, \lfloor (1 - \alpha)k \rfloor + 1$ and $\gamma_j = \arccos(\frac{\langle x_j - x_*, z - x_* \rangle}{\|x_j - x_*\| \|z - x_*\|})$, we clearly have (see Figure 1)

$$\frac{\langle x_j - x_*, z - x_* \rangle}{\|x_j - x_*\| \|z - x_*\|} = \cos(\gamma_j) > \sqrt{1 - \frac{1}{C_\alpha^2}},$$

while $\frac{\langle x_j - x_*, z - x_* \rangle}{\|x_j - x_*\| \|z - x_*\|} \geq -1$ for $j > \lfloor (1 - \alpha)k \rfloor + 1$. This yields

$$\frac{DF(x_*; z - x_*)}{\|z - x_*\|} < -(1 - \alpha)k \sqrt{1 - \frac{1}{C_\alpha^2}} + \alpha k \leq 0$$

whenever $C_\alpha \geq (1 - \alpha)\sqrt{\frac{1}{1 - 2\alpha}}$, which leads to a contradiction.

(b) See Appendix. □

Remark 2.1.

(1) Notice that in a Hilbert space, the geometric median $x_* = \text{med}(x_1, \ldots, x_k)$ always belongs to the convex hull of $\{x_1, \ldots, x_k\}$. Indeed, if x_* coincides with one of x_j's, there is nothing to prove. Otherwise, since for any $v \in \mathbb{H}$ we have $DF(x_*; v) \geq 0$ and $DF(x_*; -v) \geq 0$, it follows from (2.3) that $\sum_{j=1}^{k} \frac{x_j - x_*}{\|x_j - x_*\|} = 0$, which yields the result.

(2) In general Banach spaces, it might be convenient to consider

$$\hat{x}_* := \underset{y \in \text{co}(x_1, \ldots, x_k)}{\arg\min} \sum_{j=1}^{k} \|y - x_j\|,$$

where $\text{co}(x_1, \ldots, x_k)$ is the convex hull of $\{x_1, \ldots, x_k\}$. The claim of Lemma 2.1 remains valid for $\hat{x}_*$ whenever $z \in \text{co}(x_1, \ldots, x_k)$.

3. "Boosting the confidence" by taking the geometric median of independent estimators

A useful property of the geometric median is that it transforms a collection of independent estimators that are "weakly" concentrated around the true parameter of interest into a single estimator which admits significantly tighter deviation bounds. For $0 < p < \alpha < \frac{1}{2}$, define

$$\psi(\alpha; p) = (1 - \alpha) \log \frac{1 - \alpha}{1 - p} + \alpha \log \frac{\alpha}{p}.$$

Theorem 3.1. *Assume that $\mu \in \mathbb{X}$ is a parameter of interest, and let $\hat{\mu}_1, \ldots, \hat{\mu}_k \in \mathbb{X}$ be a collection of independent estimators of μ. Fix $\alpha \in (0, \frac{1}{2})$. Let $0 < p < \alpha$ and $\varepsilon > 0$ be such that for all j, $1 \leq j \leq k$,*

$$\Pr(\|\hat{\mu}_j - \mu\| > \varepsilon) \leq p. \tag{3.1}$$

Set

$$\hat{\mu} := \mathrm{med}(\hat{\mu}_1, \ldots, \hat{\mu}_k). \tag{3.2}$$

Then

$$\Pr(\|\hat{\mu} - \mu\| > C_\alpha \varepsilon) \leq \mathrm{e}^{-k\psi(\alpha; p)}, \tag{3.3}$$

where C_α is a constant defined in Lemma 2.1 above.

Proof. Assume that event $\mathcal{E} := \{\|\hat{\mu} - \mu\| > C_\alpha \varepsilon\}$ occurs. Lemma 2.1 implies that there exists a subset $J \subseteq \{1, \ldots, k\}$ of cardinality $|J| \geq \alpha k$ such that $\|\mu_j - \mu\| > \varepsilon$ for all $j \in J$, hence

$$\Pr(\mathcal{E}) \leq \Pr\left(\sum_{j=1}^{k} I\{\|\hat{\mu}_j - \mu\| > \varepsilon\} > \alpha k\right).$$

If W has Binomial distribution $W \sim B(k, p)$, then

$$\Pr\left(\sum_{j=1}^{k} I\{\|\hat{\mu}_j - \mu\| > \varepsilon\} > \alpha k\right) \leq \Pr(W > \alpha k)$$

(see Lemma 23 in [28] for a rigorous proof of this fact). Chernoff bound (e.g., Proposition A.6.1 in [43]) implies that

$$\Pr(W > \alpha k) \leq \exp(-k\psi(\alpha; p)). \qquad \square$$

Remark 3.1.

(a) If (3.1) is replaced by a weaker condition assuming that

$$\Pr(\|\hat{\mu}_j - \mu\| > \varepsilon) \leq p < \alpha$$

is satisfied only for $\hat{\mu}_j$, $j \in J \subset \{1, \ldots, k\}$, where $|J| = (1 - \tau)k$ for $0 \leq \tau < \frac{\alpha - p}{1 - p}$, then the previous argument implies

$$\Pr\left(\|\hat{\mu} - \mu\| > C_\alpha \varepsilon\right) \leq \exp\left(-k(1 - \tau)\psi\left(\frac{\alpha - \tau}{1 - \tau}, p\right)\right).$$

In particular, this version is useful in addressing the situation when the sample contains a subset of cardinality at most τk consisting of "outliers" of arbitrary nature.

(b) It is also clear that results of Theorem 3.1 can be used to positively answer question (3) posed in the Introduction. Indeed, if several autonomous computational resources (e.g., processors) are available, one can evaluate estimators $\hat{\mu}_j$, $j = 1, \ldots, k$ in parallel and combine them via the geometric median as a final step. In many situations, the improvement in computational cost will be significant.

Note that it is often easy to obtain an estimator satisfying (3.1) with the correct rate ε under minimal assumptions on the underlying distribution. In particular, if μ is the mean and $\hat{\mu}_k$ is the sample mean, then (3.1) can be deduced from Chebyshev's inequality, see Section 4.1 below for more details.

Next, we describe the method proposed in [34] which is based on a different notion of the median. Let $\hat{\mu}_1, \ldots, \hat{\mu}_k$ be a collection of independent estimators of μ and assume that $\varepsilon > 0$ is chosen to satisfy

$$\Pr\left(\|\hat{\mu}_j - \mu\| > \varepsilon\right) \leq p < \tfrac{1}{2}, \qquad 1 \leq j \leq k. \tag{3.4}$$

Define $\tilde{\mu} := \hat{\mu}_{j_*}$, where

$$j_* = j_*(\varepsilon) := \min\left\{ j \in \{1, \ldots, k\} : \exists I \subset \{1, \ldots, k\}\right.$$
$$\left. \text{such that } |I| > \frac{k}{2} \text{ and } \forall i \in I, \|\hat{\mu}_i - \hat{\mu}_j\| \leq 2\varepsilon\right\}, \tag{3.5}$$

and $j_* = 1$ if none of $\hat{\mu}_j$'s satisfy the condition in braces. It is not hard to show that

$$\Pr\left(\|\tilde{\mu} - \mu\| > 3\varepsilon\right) \leq e^{-k\psi(1/2; p)}, \tag{3.6}$$

which is similar to (3.3).

However, it is important to note that $\tilde{\mu}$ defined by (3.5) explicitly depends on ε which is often unknown in practice, while the "geometric median" estimator $\hat{\mu}$ (3.2) does not require any additional information.

Remark 3.2. It is possible to modify $\tilde{\mu}$ by choosing ε_* to be the smallest $\varepsilon > 0$ for which (3.5) defines a non-empty set, and setting $j_* := j_*(\varepsilon_*)$. The resulting construction does not assume that ε satisfying condition (3.4) is known a priori, while (3.6) remains valid. See [18] for discussion and applications of this method.

It is important to mention the fact that (3.6) and the inequality (3.3) of Theorem 3.1 have different constants in front of ε: it is equal to C_α in (3.3) and to 3 in (3.6). Note that in the Hilbert space case, $C_\alpha = (1-\alpha)\sqrt{\frac{1}{1-2\alpha}} \to 1$ as $\alpha \to 0$, while for general Banach spaces $C_\alpha = \frac{2(1-\alpha)}{1-2\alpha} \to$ 2 as $\alpha \to 0$. In particular, $C_\alpha < 3$ for all sufficiently small α (e.g., for $\alpha < -8 + 6\sqrt{2} \approx 0.485$ in Hilbert space framework). This difference becomes substantial when ε is of the form

$$\varepsilon = \text{approximation error} + \text{random error},$$

where the first term in the sum is a constant and the second term decreases with the growth of the sample size. This is a typical situation when the model is misspecified, see Section 4.4 below for a concrete example related to matrix regression. Our method allows to keep the constant in front of the approximation error term arbitrary close to 1 (and often leads to noticeably better constants in general).

4. Examples

In this section, we discuss applications of Theorem 3.1 to several classical problems, namely, estimation of the mean, principal component analysis, sparse linear regression and low-rank matrix recovery.

Our priority was simplicity and clarity of exposition of the main ideas which could affect optimality of some constants and generality of obtained results.

4.1. Estimation of the mean in a Hilbert space

Assume that $\mathbb{H}$ is a separable Hilbert space with norm $\|\cdot\|$. Let $X, X_1, \ldots, X_n \in \mathbb{H}$, $n \geq 2$, be an i.i.d. sample from a distribution Π such that $\mathbb{E}X = \mu$, $\mathbb{E}[(X - \mu) \otimes (X - \mu)] = \Sigma$ is the covariance operator, and $\mathbb{E}\|X - \mu\|^2 = \text{tr}(\Sigma) < \infty$. We will apply result of Theorem 3.1 to construct a "robust estimator" of μ. Let us point out that a simple alternative estimator of μ can be obtained by applying the univariate "median of the means" construction (explained in Section 1) coordinatewise. When $\dim(\mathbb{H}) < \infty$, this method leads to dimension-dependent bounds that can be even better than the result for the geometric median-based approach (when $\dim(\mathbb{H})$ is small). However, when $\dim(\mathbb{H})$ is large or infinite, dimension-dependent estimates become uninformative; see Remark 4.1 below for more details.

Set $\alpha_* := \frac{7}{18}$ and $p_* := 0.1$ (these numerical values allow to optimize the constants in Corollary 4.1 below). Let $0 < \delta < 1$ be the confidence parameter, and set

$$k =: \left\lfloor \frac{\log(1/\delta)}{\psi(\alpha_*; p_*)} \right\rfloor + 1 \leq \left\lfloor 3.5\log\left(\frac{1}{\delta}\right) \right\rfloor + 1$$

(we will assume that δ is such that $k \leq \frac{n}{2}$). Divide the sample $X_1, \ldots, X_n$ into k disjoint groups $G_1, \ldots, G_k$ of size $\lfloor \frac{n}{k} \rfloor$ each, and define

$$\hat{\mu}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i, \qquad j = 1, \ldots, k,$$

$$\hat{\mu} := \mathrm{med}(\hat{\mu}_1, \ldots, \hat{\mu}_k). \tag{4.1}$$

Corollary 4.1. *Under the aforementioned assumptions,*

$$\Pr\left(\|\hat{\mu} - \mu\| \geq 11 \sqrt{\frac{\mathrm{tr}(\Sigma) \log(1.4/\delta)}{n}} \right) \leq \delta. \tag{4.2}$$

Proof. We will apply Theorem 3.1 to the independent estimators $\hat{\mu}_1, \ldots, \hat{\mu}_k$.
To this end, we need to find ε satisfying (3.1). Since for all $1 \leq j \leq k \leq \frac{n}{2}$

$$\mathbb{E} \|\hat{\mu}_j - \mu\|^2 \leq \frac{\mathbb{E}\|X - \mu\|^2}{|G_j|} \leq \frac{2k}{n} \mathrm{tr}(\Sigma),$$

Chebyshev's inequality gives

$$\Pr\left(\|\hat{\mu}_j - \mu\| \geq \varepsilon \right) \leq \frac{2k}{n\varepsilon^2} \mathrm{tr}(\Sigma),$$

which is further bounded by p_* whenever $\varepsilon^2 \geq \frac{2k}{p_* n} \mathrm{tr}(\Sigma)$. The claim now follows from Theorem 3.1 and the bounds $C_{\alpha_*} \sqrt{\frac{2}{p_* \psi(\alpha_*; p_*)}} \leq 11$ and $\log(1/\delta) + \psi(\alpha_*; p_*) \leq \log(\frac{1.4}{\delta})$. $\qquad \square$

Remark 4.1.

(a) It is easy to see that the proof of Corollary 4.1 actually yields a better bound

$$\Pr\left(\|\hat{\mu} - \mu\| \geq \frac{C_{\alpha_*}}{\sqrt{p_* \psi(\alpha_*; p_*)}} \sqrt{\mathrm{tr}(\Sigma) \frac{\log(1.4/\delta)}{n - 3.5 \log(1.4/\delta)}} \right) \leq \delta, \tag{4.3}$$

with $\frac{C_{\alpha_*}}{\sqrt{p_* \psi(\alpha_*; p_*)}} \leq 7.6$.

(b) For the estimator $\tilde{\mu}$ defined in (3.5), it follows from (3.6) (with $p = 0.12$) that

$$\Pr\left(\|\tilde{\mu} - \mu\| \geq 13.2 \sqrt{\mathrm{tr}(\Sigma) \frac{\log(1.6/\delta)}{n - 2.4 \log(1.6/\delta)}} \right) \leq \delta,$$

which yields a noticeably larger constant (13.2 versus 7.6).

(c) When $\mathbb{H}$ is a D-dimensional Euclidean space, it is interesting to compare $\hat{\mu}$ with another natural estimator $\hat{\mu}_*$ obtained by taking the median *coordinate-wise*, that is, if

$\hat{\mu}_j = (\hat{\mu}_j^{(1)}, \ldots, \hat{\mu}_j^{(D)})$, $j = 1, \ldots, k$, then

$$\hat{\mu}_* := \left(\text{med}(\hat{\mu}_1^{(1)}, \ldots, \hat{\mu}_k^{(1)}), \ldots, \text{med}(\hat{\mu}_1^{(D)}, \ldots, \hat{\mu}_k^{(D)}) \right).$$

It is easy to see that for the univariate median, inequality (3.3) holds with $\alpha = 1/2$ and $C_\alpha = 1$, hence the union bound over $i = 1, \ldots, D$ implies that

$$\Pr\left(\|\hat{\mu}_* - \mu\| \geq 4.4 \sqrt{\text{tr}(\Sigma) \frac{\log(1.6D/\delta)}{n - 2.4 \log(1.6D/\delta)}} \right) \leq \delta \tag{4.4}$$

(here, p was set to be 0.12). This bound should be compared to (4.3) – the latter becomes better only when D is sufficiently large (e.g., $D \geq 165$ for $\delta = 0.1$ and $D \geq 15\,806$ for $\delta = 0.01$).[1] Note that the constant in (4.4) can be further improved in a situation when tight upper bounds on the true variances or kurtoses of coordinates of X are known by using a univariate estimator of [14] to construct $\hat{\mu}_*$.

Our estimation technique naturally extends to the problem of constructing the confidence sets for the mean. Indeed, when faced with the task of obtaining the non-asymptotic confidence interval, one usually fixes the desired coverage probability in advance, which is exactly how we build our estimator. To obtain a parameter-free confidence ball from (4.2), one has to estimate $\text{tr}(\Sigma)$. To this end, we will apply Theorem 3.1 to a collection of independent statistics $\hat{T}_1, \ldots, \hat{T}_k$, where

$$\hat{T}_j = \frac{1}{|G_j|} \sum_{i \in G_j} \|X_i - \hat{\mu}_j\|^2, \qquad j = 1, \ldots, k,$$

and $\hat{\mu}_j$ are the sample means defined in (4.1). Let $\hat{T} := \text{med}(\hat{T}_1, \ldots, \hat{T}_k)$ (if k is even, the median is not unique, so we pick an arbitrary representative).

Proposition 4.1. *Assume that*

$$15.2 \sqrt{\frac{\mathbb{E}\|X - \mu\|^4 - (\text{tr}(\Sigma))^2}{(\text{tr}(\Sigma))^2}} \leq \left(\frac{1}{2} - 178 \frac{\log(1.4/\delta)}{n} \right) \sqrt{\frac{n}{\log(1.4/\delta)}}. \tag{4.5}$$

Then

$$\Pr\left(\text{tr}(\Sigma) \leq 2\hat{T} \right) \geq 1 - \delta. \tag{4.6}$$

Proof. Note that

$$\hat{T}_j = \frac{1}{|G_j|} \sum_{i \in G_j} \|X_i - \mu\|^2 - \|\hat{\mu}_j - \mu\|^2.$$

---

[1] We want to thank the anonymous reviewer for pointing this out.

Chebyshev's inequality gives (assuming that $k \leq n/2$)

$$\Pr\left(\|\hat{\mu}_j - \mu\|^2 \geq \underbrace{4\frac{\mathrm{tr}(\Sigma)\log(1.4/\delta)}{p_*\psi(\alpha_*;\, p_*)n}}_{\varepsilon_1} \right) \leq \frac{p_*}{2},$$

$$\Pr\left(\left| \frac{1}{|G_j|}\sum_{i \in G_j} \|X_i - \mu\|^2 - \mathrm{tr}(\Sigma) \right| \geq \underbrace{2\sqrt{\mathbb{E}\|X - \mu\|^4 - \left(\mathrm{tr}(\Sigma)\right)^2}\sqrt{\frac{\log(1.4/\delta)}{np_*\psi(\alpha_*;\, p_*)}}}_{\varepsilon_2} \right) \leq \frac{p_*}{2},$$

hence $\Pr(|\hat{T}_j - \mathrm{tr}(\Sigma)| \geq \varepsilon_1 + \varepsilon_2) \leq p_*$. Theorem 3.1 implies that

$$\Pr\left(\hat{T} \leq \mathrm{tr}(\Sigma) - C_{\alpha_*}(\varepsilon_1 + \varepsilon_2) \right) \leq \Pr\left(\left| \hat{T} - \mathrm{tr}(\Sigma) \right| \geq C_{\alpha_*}(\varepsilon_1 + \varepsilon_2) \right) \leq \delta.$$

Since $\Pr(\hat{T} \leq \frac{\mathrm{tr}(\Sigma)}{2}) \leq \Pr(\hat{T} \leq \mathrm{tr}(\Sigma) - C_{\alpha_*}(\varepsilon_1 + \varepsilon_2))$ whenever (4.5) is satisfied, the result follows. $\qquad\square$

Combining (4.6) with Corollary 4.1, we immediately get the following statement.

Corollary 4.2. *Let $B(h, r)$ be the ball of radius r centered at $h \in \mathbb{H}$. Define the random radius*

$$r_n := 11\sqrt{2}\sqrt{\hat{T}\frac{\log(1.4/\delta)}{n}}$$

and let $\hat{\mu}$ be the estimator defined by (4.1). If (4.5) holds, then

$$\Pr\left(B(\hat{\mu}, r_n) \text{ contains } \mu \right) \geq 1 - 2\delta.$$

4.2. Robust Principal Component Analysis

It is well known that classical Principal Component Analysis (PCA) [38] is very sensitive to the presence of the outliers in a sample. The literature on robust PCA suggests several computationally efficient and theoretically sound methods to recover the linear structure in the data. For instance, if part of the observations is contained in a low-dimensional subspace while the rest are corrupted by noise, the low-dimensional subspace can often be recovered exactly, see [9,48] and references therein.

However, for the case when no additional geometric structure in the data can be assumed, we suggest a simple and easy-to-implement alternative which uses the geometric median to obtain a robust estimator of the covariance matrix. In this section, we study the simplest case when the geometric median is combined with the sample covariance estimator. However, it is possible to use various alternatives in place of the sample covariance, such as the shrinkage estimator [27], banding/tapering estimator [3], hard thresholding estimator [4] or the nuclear norm-penalized estimator [29], to name a few.

Let $X, X_1, \ldots, X_n \in \mathbb{R}^D$ be i.i.d. random vectors such that $\mathbb{E}X = \mu$, $\mathbb{E}[(X - \mu)(X - \mu)^T] = \Sigma$ and $\mathbb{E}\|X\|^4 < \infty$, where $\|\cdot\|$ is the usual Euclidean norm. We are interested in estimating the covariance matrix Σ and the linear subspace generated by its eigenvectors associated to "large" eigenvalues. For simplicity, suppose that all positive eigenvalues of Σ have algebraic multiplicity 1. We will enumerate $\lambda_i := \lambda_i(\Sigma)$ in the decreasing order, so that $\lambda_1 > \lambda_2 \geq \cdots \geq 0$.

Assume first that the data is centered (so that $\mu = 0$). As before, set $\alpha_* := \frac{7}{18}$, $p_* := 0.1$, divide the sample $X_1, \ldots, X_n$ into $k = \lfloor \frac{\log(1/\delta)}{\psi(\alpha_*; p_*)} \rfloor + 1$ disjoint groups $G_1, \ldots, G_k$ of size $\lfloor \frac{n}{k} \rfloor$ each, and let

$$\hat{\Sigma}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i X_i^T, \qquad j = 1, \ldots, k,$$

$$\hat{\Sigma} := \mathrm{med}(\hat{\Sigma}_1, \ldots, \hat{\Sigma}_k),$$

(4.7)

where the median is taken with respect to Frobenius norm $\|A\|_F := \sqrt{\mathrm{tr}(A^T A)}$.

Remark 4.2. Note that $\hat{\Sigma}$ is positive semidefinite as a convex combination of positive semidefinite matrices.

Let Proj_m be the orthogonal projector on a subspace corresponding to the m largest positive eigenvalues of Σ. Let $\widehat{\mathrm{Proj}_m}$ be the orthogonal projector of the same rank as Proj_m corresponding to the $m \leq \lfloor \frac{n}{k} \rfloor$ largest eigenvalues of $\hat{\Sigma}$. In this case, the following bound holds.

Corollary 4.3. *Let $\Delta_m := \lambda_m - \lambda_{m+1}$ and assume that*

$$\Delta_m > 44\sqrt{\frac{(\mathbb{E}\|X\|^4 - \mathrm{tr}(\Sigma^2)) \log(1.4/\delta)}{n}}.$$

(4.8)

Then

$$\Pr\left(\|\widehat{\mathrm{Proj}_m} - \mathrm{Proj}_m\|_F \geq \frac{22}{\Delta_m} \sqrt{\frac{(\mathbb{E}\|X\|^4 - \mathrm{tr}(\Sigma^2)) \log(1.4/\delta)}{n}} \right) \leq \delta.$$

Proof. It follows from Davis–Kahan perturbation theorem [16] (see also Theorem 3 in [49]) that, whenever $\|\hat{\Sigma} - \Sigma\|_F < \frac{1}{4}\Delta_m$,

$$\|\widehat{\mathrm{Proj}_m} - \mathrm{Proj}_m\|_F \leq \frac{2\|\hat{\Sigma} - \Sigma\|_F}{\Delta_m}.$$

(4.9)

Define $Y_j := X_j X_j^T$, $j = 1, \ldots, n$ and note that $\mathbb{E}\|Y - \mathbb{E}Y\|_F^2 = \mathbb{E}\|X\|^4 - \mathrm{tr}(\Sigma^2)$. Applying Corollary 4.1 to Y_j, $j = 1, \ldots, n$, we get

$$\Pr\left(\|\hat{\Sigma} - \Sigma\|_F \geq 11\sqrt{\frac{(\mathbb{E}\|X\|^4 - \mathrm{tr}(\Sigma^2)) \log(1.4/\delta)}{n}} \right) \leq \delta.$$

Whenever (4.8) is satisfied, inequality $11\sqrt{\frac{(\mathbb{E}\|X\|^4 - \mathrm{tr}(\Sigma^2))\log(1.4/\delta)}{n}} < \frac{\Delta_m}{4}$ holds, and (4.9) yields the result. $\qquad\square$

Similar bounds can be obtained in a more general situation when X is not necessarily centered. To this end, let

$$\hat{\mu}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i, \qquad j = 1, \ldots, k,$$

$$\hat{\Sigma}_j := \frac{1}{|G_j|} \sum_{i \in G_j} (X_i - \hat{\mu}_j)(X_i - \hat{\mu}_j)^T, \qquad j = 1, \ldots, k, \tag{4.10}$$

$$\hat{\Sigma} := \mathrm{med}(\hat{\Sigma}_1, \ldots, \hat{\Sigma}_k).$$

Note that $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_k$ are independent. Then, using the fact that for any $1 \le j \le k$

$$\hat{\Sigma}_j = \frac{1}{|G_j|} \sum_{i \in G_j} (X_i - \mu)(X_i - \mu)^T - (\mu - \hat{\mu}_j)(\mu - \hat{\mu}_j)^T,$$

it is easy to prove the following bound.

Corollary 4.4. *Let*

$$\varepsilon_n(\delta) := 15.2\sqrt{\frac{(\mathbb{E}\|X - \mu\|^4 - \mathrm{tr}(\Sigma^2))\log(1.4/\delta)}{n}} + 178\frac{\mathrm{tr}(\Sigma)\log(1.4/\delta)}{n}$$

and assume that $\Delta_m > 4\varepsilon_n(\delta)$. Then

$$\Pr\left(\|\widehat{\mathrm{Proj}}_m - \mathrm{Proj}_m\|_F \ge \frac{2\varepsilon_n(\delta)}{\Delta_m}\right) \le \delta.$$

4.3. High-dimensional sparse linear regression

Everywhere in this subsection, $\|\cdot\|$ stands for the standard Euclidean norm, $\|\cdot\|_1$ denotes the ℓ_1-norm and $\|\cdot\|_\infty$ – the sup-norm of a vector.

Let $x_1, \ldots, x_n \in \mathbb{R}^D$ be a fixed collection of vectors and let Y_j be noisy linear measurements of $\lambda_0 \in \mathbb{R}^D$:

$$Y_j = \lambda_0^T x_j + \xi_j, \tag{4.11}$$

where ξ_j are independent zero-mean random variables such that $\mathrm{Var}(\xi_j) \le \sigma^2, 1 \le j \le n$. Set $\mathbb{X} := (x_1|\ldots|x_n)^T$.

We are interested in the case when $D \gg n$ and λ_0 is sparse, meaning that

$$N(\lambda_0) := \big|\mathrm{supp}(\lambda_0)\big| = \big|\{j : \lambda_{0,j} \ne 0\}\big| = s \ll D.$$

In this situation, a (version of) the famous Lasso estimator [42] of λ_0 is obtained as a solution of the following optimization problem:

$$\hat{\lambda}_\varepsilon := \arg\min_{\lambda \in \mathbb{R}^D} \left[\frac{1}{n} \sum_{j=1}^{n} (Y_j - \lambda^T x_j)^2 + \varepsilon \|\lambda\|_1 \right]. \tag{4.12}$$

The goal of this section is to extend the applicability of some well-known results for this estimator to the case of a heavy-tailed noise distribution.

Existing literature on high-dimensional linear regression suggests several ways to handle corrupted measurements, for instance, by using a different loss function (e.g., the so-called *Huber's loss* [26]), or by implementing a more flexible penalty term [35,47]. In particular, in [35] authors study the model

$$Y = X\lambda_0 + e^* + \xi, \tag{4.13}$$

where $X \in \mathbb{R}^{n \times D}, \xi \in \mathbb{R}^n$ is the additive noise and $e^* \in \mathbb{R}^n$ is a sparse error vector with unknown support and arbitrary large entries. It is shown that if the rows of X are independent Gaussian random vectors, then is possible to accurately estimate both λ_0 and e^* by adding an extra penalty term:

$$(\tilde{\lambda}_\varepsilon, \tilde{e}_\varepsilon) := \arg\min_{\lambda \in \mathbb{R}^D, e \in \mathbb{R}^n} \left[\frac{1}{n} \|Y - X\lambda - e\|^2 + \varepsilon_1 \|\lambda\|_1 + \varepsilon_2 \|e\|_1 \right].$$

However, as in the case of the usual Lasso, confidence of estimation depends on the distribution of ξ. In particular, Gaussian-type concentration holds only if the entries of ξ have sub-Gaussian tails.

The main result of this subsection (stated in Theorem 4.2) provides strong performance guarantees for the robust version of the usual Lasso estimator (4.12) and requires only standard conditions on the degree of sparsity and restricted eigenvalues of the design. Similar method can be used to improve performance guarantees for the model (4.13) in the case of heavy-tailed noise ξ.

Probabilistic bounds for the error $\|\hat{\lambda}_\varepsilon - \lambda_0\|$ crucially depend on integrability properties of the noise variable. We will recall some known bounds for the case when $\xi_j \sim N(0, \sigma^2), j = 1, \ldots, n$ (of course, similar results hold for *sub-Gaussian* noise as well). For $J \subset \{1, \ldots, D\}$ and $u \in \mathbb{R}^D$, define $u_J \in \mathbb{R}^D$ by $(u_J)_j = u_j, j \in J$ and $(u_J)_j = 0, j \in J^c$ (here, J^c denotes the complement of a set J).

Definition 4.1 (Restricted eigenvalue condition, [5]). *Let $1 \leq s \leq D$ and $c_0 > 0$. We will say that the restricted eigenvalue condition holds if*

$$\kappa(s, c_0) := \min_{\substack{J \subset \{1,\ldots,D\} \\ |J| \leq s}} \min_{\substack{u \in \mathbb{R}^D, u \neq 0 \\ \|u_{J^c}\|_1 \leq c_0 \|u_J\|_1}} \frac{\|\mathbb{X}u\|}{\sqrt{n} \|u_J\|} > 0.$$

Let $\Theta := \|\frac{1}{n} \sum_{j=1}^{n} \xi_j x_j\|_\infty$. The following result shows that the amount of regularization ε sufficient for recovery of λ_0 is closely related to the size of Θ.

Theorem 4.1. *Assume that the diagonal elements of the matrix $\frac{\mathbb{X}^T\mathbb{X}}{n}$ are bounded by 1 and*

$$\kappa\big(2N(\lambda_0), 3\big) > 0.$$

On the event $\mathcal{E} = \{\varepsilon \geq 4\Theta\}$, the following inequality holds:

$$\|\hat{\lambda}_\varepsilon - \lambda_0\|^2 \leq 64\varepsilon^2 \frac{N(\lambda_0)}{\kappa^4(2N(\lambda_0), 3)}.$$

In particular, when $\xi_j \sim N(0, \sigma^2)$ and $\varepsilon = 4\sigma t\sqrt{\frac{\log D}{n}}$,

$$\Pr(\mathcal{E}) \geq 1 - \frac{2}{D^{t^2-2}}.$$

Proof. This result is similar to the statement of Theorem 7.2 in [5], and its proof can be obtained along the same lines. See [30] for more details. □

Our next goal is to construct an estimator of λ_0 which admits high confidence error bounds without restrictive assumptions on the noise variable, such as sub-Gaussian tails. Let $t > 0$ be fixed, and set $k := \lfloor 3.5t \rfloor + 1$, $m = \lfloor \frac{n}{k} \rfloor$ (as before, we will assume that $k \leq \frac{n}{2}$). For $1 \leq l \leq k$, let $G_l := \{(l-1)m + 1, \ldots, lm\}$ and

$$\mathbb{X}_l = (x_{j_1} | \ldots | x_{j_m})^T, \qquad j_i = i + (l-1)m \in G_l$$

be the $m \times D$ design matrix corresponding to the lth group of design vectors $\{x_j, j \in G_l\}$. Moreover, let $\kappa_l(s, c_0)$ be the corresponding restricted eigenvalues.

Define

$$\hat{\lambda}_\varepsilon^l := \arg\min_{\lambda \in \mathbb{R}^D} \left[\frac{1}{|G_l|} \sum_{j \in G_l} \left(Y_j - \lambda^T x_j\right)^2 + \varepsilon\|\lambda\|_1 \right]$$

and

$$\hat{\lambda}_\varepsilon^* := \text{med}\big(\hat{\lambda}_\varepsilon^1, \ldots, \hat{\lambda}_\varepsilon^k\big), \tag{4.14}$$

where the geometric median is taken with respect to the standard Euclidean norm in $\mathbb{R}^D$. The following result holds.

Theorem 4.2. *Assume that $\|x_j\|_\infty \leq M, 1 \leq j \leq n$ and $\bar{\kappa}(2N(\lambda_0), 3) := \min_{1 \leq l \leq k} \kappa_l(2N(\lambda_0), 3) > 0$. Then for any*

$$\varepsilon \geq 95M\sigma\sqrt{\frac{t + 2/7}{n}}\log(2D),$$

with probability $\geq 1 - e^{-t}$

$$\big\|\hat{\lambda}_\varepsilon^* - \lambda_0\big\|^2 \leq 83\varepsilon^2 \frac{N(\lambda_0)}{\bar{\kappa}^4(2N(\lambda_0), 3)}.$$

Proof. We will first obtain a "weak concentration" bound from Theorem 4.1 and then apply Theorem 3.1 with $\alpha = \frac{7}{18}$ to get the result.

To this end, we need to estimate $\Theta_l := \|\frac{1}{m} \sum_{j \in G_l} \xi_j x_j\|_\infty, l = 1, \ldots, k$.

Lemma 4.1 (Nemirovski's inequality, Lemma 5.2.2 in [33] or Lemma 14.24 in [8]). *Assume that $D \geq 3$. Then for any l, $1 \leq l \leq k$,*

$$\mathbb{E}\Theta_l^2 \leq \frac{8 \log(2D)}{m} \frac{1}{m} \sum_{j \in G_l} \|x_j\|_\infty^2 \mathbb{E}\xi_j^2.$$

By our assumptions, $\|x_j\|_\infty \leq M$ and $\mathbb{E}\xi_j^2 \leq \sigma^2$ for all j, hence Chebyshev's inequality gives that for any $1 \leq l \leq k$,

$$\Pr(\Theta_l \geq t) \leq \frac{8 \log(2D) M^2 \sigma^2}{m t^2} \leq 0.1$$

whenever $t \geq 4\sigma M \sqrt{\frac{k \log(2D)}{0.1 n}}$. In particular, for $\varepsilon \geq 16\sigma M \sqrt{\frac{3.5(t+2/7) \log(2D)}{0.1 n}}$, the bound of Theorem 4.1 holds for $\hat{\lambda}_\varepsilon^l$ with probability $\geq 1 - 0.1$; note that $16\sqrt{\frac{3.5}{0.1}} \leq 95$. It remains to apply Theorem 3.1 to complete the proof. $\square$

Remark 4.3. We stated the bounds only for the Euclidean distance $\|\hat{\lambda}_\varepsilon^* - \lambda_0\|$; this formulation is close to the compressed sensing framework [12]. If, for example, the design vectors $x, x_1, \ldots, x_n$ are i.i.d. with some known distribution Π, one can use the median with respect to $\|\cdot\|_{L_2(\Pi)}$ norm in the definition of $\hat{\lambda}_\varepsilon^*$ and obtain the bounds for the *prediction error* $\|\hat{\lambda}_\varepsilon^* - \lambda_0\|_{L_2(\Pi)}^2 := \mathbb{E}((\hat{\lambda}_\varepsilon^* - \lambda_0)^T x)^2$.

4.4. Matrix regression with isotropic sub-Gaussian design

In this section, we will extend some results related to recovery of low-rank matrices from noisy linear measurements to the case of heavy-tailed noise distribution. Assume that the random couple $(X, Y) \in \mathbb{R}^{D \times D} \times \mathbb{R}$ is generated according to the following matrix regression model:

$$Y = f_*(X) + \xi, \tag{4.15}$$

where f_* is the regression function, $X \in \mathbb{R}^{D \times D}$ is a random symmetric matrix with (unknown) distribution Π and ξ is a zero-mean random variable independent of X with $\text{Var}(\xi) \leq \sigma^2$. We will be mostly interested in a situation when f_* can be well approximated by a linear functional $\langle A, \cdot \rangle$, where A a symmetric matrix of small rank and $\langle A_1, A_2 \rangle := \text{tr}(A_1^T A_2)$.

The problem of low-rank matrix estimation has attracted a lot of attention during the last several years, for example, see [11,39] and references therein. Recovery guarantees were later extended to allow the presence of noise. Results in this direction can be found in [10,24,32,40], to name a few.

In this section, we mainly follow the approach of [23] (Chapter 9) which deals with an important case of *sub-Gaussian design* (also, see [10,31] for a discussion of related problems), and use results of this work as a basis for our exposition. Everywhere below, $\| \cdot \|_F$ stands for the Frobenius norm of a matrix, $\| \cdot \|_{Op}$ – for the operator (spectral) norm, and $\| \cdot \|_1$ – for the nuclear norm of a matrix.

Given $A \in \mathbb{R}^{D \times D}$, denote $\|A\|^2_{L_2(\Pi)} := \mathbb{E}(\mathrm{tr}(A^T X))^2$. Recall that a random variable ζ is called *sub-Gaussian* with parameter γ^2 if for all $s \in \mathbb{R}$, $\mathbb{E}^{s\zeta} \leq e^{s^2 \gamma^2 / 2}$. We will be interested in the special case when X is *sub-Gaussian*, meaning that there exists $\gamma = \gamma(\Pi) > 0$ such that for all symmetric matrices A, $\langle A, X \rangle$ is a sub-Gaussian random variable with parameter $\gamma^2 \|A\|^2_{L_2(\Pi)}$ (in particular, this is the case when the entries of X are jointly Gaussian, with $\gamma = 1$). Additionally, we will assume that X is isotropic, so that $\|A\|_{L_2(\Pi)} = \|A\|_F$ for any symmetric matrix A.

In particular, these assumptions hold in the following important cases:

(a) X is symmetric and such that $\{X_{i,j}, 1 \leq i \leq j \leq D\}$ are i.i.d. centered normal random variables with $\mathbb{E}X^2_{i,j} = \frac{1}{2}$, $i < j$ and $\mathbb{E}X^2_{i,i} = 1$, $i = 1, \ldots, D$.

(b) X is symmetric and such that $X_{i,j} = \frac{1}{\sqrt{2}} \varepsilon_{i,j}$, $i < j$, $X_{i,i} = \varepsilon_{i,i}$, $1 \leq i \leq D$, where $\varepsilon_{i,j}$ are i.i.d. Rademacher random variables (i.e., random signs).

(c) In a special case when all involved matrices are diagonal, the problem becomes a version of sparse linear regression with random design. In this case, isotropic design includes a situation when X is a random diagonal matrix $X = \mathrm{diag}(x_1, \ldots, x_D)$, where x_i are i.i.d. standard normal or Rademacher random variables.

Remark 4.4. In what follows, $C_1, C_2, \ldots$ denote the constants that may depend on parameters of the underlying distribution (such as γ).

Given $\alpha \geq 1$, define $\|\zeta\|_{\psi_\alpha} := \min\{r > 0 : \mathbb{E}\exp((\frac{|\zeta|}{r})^\alpha) \leq 2\}$. We will mainly use $\| \cdot \|_{\psi_\alpha}$-norms for $\alpha = 1, 2$. The following elementary inequality holds: for any random variables ζ_1, ζ_2,

$$\|\zeta_1 \zeta_2\|_{\psi_1} \leq \|\zeta_1\|_{\psi_2} \|\zeta_2\|_{\psi_2}. \tag{4.16}$$

It is easy to see that $\|\zeta\|_{\psi_2} < \infty$ for any sub-Gaussian random variable ζ. It follows from Proposition 9.1 in [23] that there exists $C(\gamma) > 0$ such that for any sub-Gaussian isotropic matrix X,

$$\big\| \|X\|_{Op} \big\|_{\psi_2} \leq C(\gamma)\sqrt{D}. \tag{4.17}$$

We will also need the following useful inequality: for any $p \geq 1$,

$$\mathbb{E}^{1/p} \big| \langle A, X \rangle \big|^p \leq C_{p,\gamma} \|A\|_{L_2(\Pi)}. \tag{4.18}$$

The proofs of the facts mentioned above can be found in [23].

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. observations with the same distribution as (X, Y). We are mainly interested in the case when $D < n \ll D^2$. In this situation, it is impossible to estimate A_0 consistently in general, however, if A_0 is low-rank (or approximately low-rank), then the solution

of the following optimization problem provides a good approximation to A_0:

$$\hat{A}_\varepsilon := \arg\min_{A \in \mathbb{L}} \left[\frac{1}{n} \sum_{j=1}^{n} (Y_j - \langle A, X_j \rangle)^2 + \varepsilon \|A\|_1 \right]. \tag{4.19}$$

Here, $\mathbb{L}$ is a bounded, closed, convex subset of a set of all $D \times D$ symmetric matrices.

Remark 4.5. All results of this subsection extend to the case of unbounded $\mathbb{L}$ and non-isotropic sub-Gaussian design. However, our assumptions still cover important examples and yield less technical statements; see Theorem 9.3 in [23] for details on the general case. Results for the arbitrary rectangular matrices follow from the special case discussed here, see the remark on page 202 of [23].

We proceed by recalling the performance guarantees for $\hat{A}_\varepsilon$. Let $R_{\mathbb{L}} := \sup_{A \in \mathbb{L}} \|A\|_1$, and define

$$\Theta := \frac{1}{n} \sum_{j=1}^{n} \xi_j X_j.$$

Theorem 4.3 (Theorem 9.4 in [23]). *There exist constants c, C with the following property*: *let $\kappa := \log \log_2(D R_{\mathbb{L}})$, and assume that $t \geq 1$ is such that $t_{n,D} \leq cn$, where $t_{n,D} := (t + \kappa) \log n + \log(2D)$. Define the event $\mathcal{E} := \{\varepsilon \geq 2\|\Theta\|_{\mathrm{Op}}\}$. The following bound holds with probability $\geq \Pr(\mathcal{E}) - e^{-t}$:*

$$\|\hat{A}_\varepsilon - A_0\|_{\mathrm{F}}^2 \leq \inf_{A \in \mathbb{L}} \left[2\|A - A_0\|_{\mathrm{F}}^2 + C \left(\varepsilon^2 \operatorname{rank}(A) + R_{\mathbb{L}}^2 \frac{D t_{n,D}}{n} + \frac{1}{n} \right) \right]. \tag{4.20}$$

Constant 2 in front of $\|A - A_0\|_{\mathrm{F}}^2$ can be replaced by $(1 + \nu)$ for any $\nu > 0$ if C is replaced by C/ν.

Assumption 4.1. *The noise variable ξ is such that $\|\xi\|_{\psi_2} < \infty$.*

If Assumption 4.1 is satisfied, then, whenever

$$\varepsilon \geq \bar{C}(\gamma) \sqrt{\frac{D}{n}} \left(\sigma \sqrt{t + \log(2D)} \vee \|\xi\|_{\psi_2} \log \left(2 \vee \frac{\|\xi\|_{\psi_2}}{\sigma} \right) \frac{t + \log(2D)}{\sqrt{n}} \right) \tag{4.21}$$

we have that $\Pr(\mathcal{E}) \geq 1 - e^{-t}$ (hence, (4.20) holds with probability $1 - 2e^{-t}$). This follows from the following variant of the non-commutative Bernstein's inequality.

Theorem 4.4 (Theorem 2.7 in [23]). *Let $Y_1, \ldots, Y_n \in \mathbb{R}^{D \times D}$ be symmetric independent random matrices such that $\mathbb{E} Y_j = 0$ and*

$$\max_{1 \leq j \leq n} \left(\left\| \|Y_j\|_{\mathrm{Op}} \right\|_{\psi_1} \vee 2\mathbb{E}^{1/2} \|Y_j\|_{\mathrm{Op}}^2 \right) \leq U.$$

Let

$$\Psi^2 \geq \frac{1}{n}\left\|\sum_{i=1}^{n}\mathbb{E}Y_i^2\right\|_{\text{Op}}.$$

Then, for all $t > 0$, with probability $\geq 1 - e^{-t}$

$$\left\|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}Y_j\right\|_{\text{Op}} \leq \bar{C}_1 \max\left(\Psi\sqrt{t + \log(2D)}, U\log\left(\frac{U}{\Psi}\right)\frac{t + \log(2D)}{\sqrt{n}}\right),$$

where $\bar{C}_1 > 0$ is an absolute constant.

Indeed, recall that $\Theta = \frac{1}{n}\sum_{j=1}^{n}\xi_j X_j$, and apply Theorem 4.4 to $Y_j := \xi_j X_j$, noting that by (4.16), (4.17)

$$\left\|\mathbb{E}\xi^2 X^2\right\|_{\text{Op}} \leq \sigma^2 \mathbb{E}\|X\|_{\text{Op}}^2 \leq C_2\sigma^2 D,$$

$$\left\|\|\xi X\|_{\text{Op}}\right\|_{\psi_1} \leq \|\xi_1\|_{\psi_2}\left\|\|X\|_{\text{Op}}\right\|_{\psi_2} \leq C(\gamma)\|\xi\|_{\psi_2}\sqrt{D}.$$

It implies that with probability $\geq 1 - e^{-t}$,

$$\left\|\frac{1}{n}\sum_{j=1}^{n}\xi_j X_j\right\|_{\text{Op}} \leq C_3\sqrt{\frac{D}{n}}\left(\sigma\sqrt{t + \log(2D)} \vee \|\xi\|_{\psi_2}\log\left(2 \vee \frac{\|\xi\|_{\psi_2}}{\sigma}\right)\frac{t + \log(2D)}{\sqrt{n}}\right), \quad (4.22)$$

where C_2, C_3 depend only on γ, hence giving the result.

As we mentioned above, our goal is to construct the estimator of A_0 which admits bounds in flavor of Theorem 4.3 that hold with high probability under a much weaker assumption on the tail of the noise variable ξ.

To achieve this goal, we follow the same pattern as before. Let $t \geq 1$ be fixed, let $k := \lfloor t \rfloor + 1, m = \lfloor \frac{n}{k} \rfloor$, and assume that $k \leq \frac{n}{2}$. Divide the data $\{(X_j, Y_j)\}_{j=1}^{n}$ into k disjoint groups $G_1, \ldots, G_k$ of size m each, and define

$$\hat{A}_\varepsilon^l := \arg\min_{A \in \mathbb{L}}\left[\frac{1}{|G_l|}\sum_{j \in G_l}\left(Y_j - \langle A, X_j\rangle\right)^2 + \varepsilon\|A\|_1\right]$$

and

$$\hat{A}_\varepsilon^* = \hat{A}_\varepsilon(t) := \text{med}\left(\hat{A}_\varepsilon^1, \ldots, \hat{A}_\varepsilon^k\right),$$

where the geometric median is evaluated with respect to the Frobenius norm.

Assumption 4.2.

$$\|\xi\|_{2,1} := \int_0^\infty \sqrt{\Pr(|\xi| > x)}\,dx < \infty.$$

In particular, $\|\xi\|_{2,1} < \infty$ if $\mathbb{E}|\xi|^{2+\delta} < \infty$ for some $\delta > 0$, which is a mild requirement compared to Assumption 4.1. Finally, given $\alpha \in (0, 1/2)$, it will be convenient to define

$$p^* = p^*(\alpha) := \max\{p \in (0, \alpha) : \psi(\alpha; p) \geq 1\}.$$

Theorem 4.5. *Suppose that Assumption 4.2 is satisfied. For any $\alpha \in (0, 1/2)$, there exist constants c, C, B with the following properties: let $\kappa := \log \log_2(D R_{\mathbb{L}})$, $s_{n,t,D} := (\log(2/p^*(\alpha)) + \kappa) \log(n/t) + \log(2D)$, and assume that $s_{n,t,D} \leq c(n/t)$. Then for all*

$$\varepsilon \geq \frac{B}{p^*(\alpha)} \|\xi\|_{2,1} \sqrt{\frac{Dt}{n}} \log(2D),$$

with probability $\geq 1 - 2\mathrm{e}^{-t}$

$$\|\hat{A}_{\varepsilon}^* - A_0\|_{\mathrm{F}}^2 \leq C_\alpha \inf_{A \in \mathbb{L}} \left[2\|A - A_0\|_{\mathrm{F}}^2 + C\left(\varepsilon^2 \operatorname{rank}(A) + R_{\mathbb{L}}^2 s_{n,t,D} \frac{Dt}{n} + \frac{t}{n} \right) \right], \qquad (4.23)$$

where C_α is defined by (2.2).

Proof. We will start by deriving a "weak concentration" bound from Theorem 4.3. To this end, we need to estimate

$$\mathbb{E}\|\Theta_l\|_{\mathrm{Op}} := \mathbb{E}\left\| \frac{1}{|G_l|} \sum_{j \in G_l} \xi_j X_j \right\|_{\mathrm{Op}}, \qquad l = 1, \ldots, k.$$

The following result is a direct consequence of the so-called *multiplier inequality* (Lemma 2.9.1 in [43]).

Lemma 4.2. *Let $\varepsilon_1, \ldots, \varepsilon_m$ be i.i.d. Rademacher random variables independent of $X_1, \ldots, X_m$. Then*

$$\mathbb{E}\left\| \frac{1}{m} \sum_{j=1}^{m} \xi_j X_j \right\|_{\mathrm{Op}} \leq \frac{2\sqrt{2}\|\xi\|_{2,1}}{\sqrt{m}} \max_{1 \leq i \leq m} \mathbb{E}\left\| \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \varepsilon_j X_j \right\|_{\mathrm{Op}}. \qquad (4.24)$$

To estimate $\mathbb{E}\|\frac{1}{\sqrt{i}} \sum_{j=1}^{i} \xi_j X_j\|_{\mathrm{Op}}$, we use the formula $\mathbb{E}|\eta| = \int_0^\infty \Pr(|\eta| \geq t)\,dt$ and the tail bound of Theorem 4.4, which implies (in a way similar to (4.22)) that with probability $\geq 1 - \mathrm{e}^{-t}$

$$\left\| \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \varepsilon_j X_j \right\|_{\mathrm{Op}} \leq C_4 \sqrt{D}\left(\sqrt{t + \log(2D)} \vee \frac{t + \log(2D)}{\sqrt{i}} \right), \qquad (4.25)$$

hence for any $1 \leq i \leq m$

$$\mathbb{E}\left\| \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \varepsilon_j X_j \right\|_{\mathrm{Op}} \leq C_5 \sqrt{D}\left(\sqrt{\log(2D)} + \frac{\log(2D)}{\sqrt{i}} \right),$$

and (4.24) yields

$$\mathbb{E}\left\|\frac{1}{m}\sum_{j=1}^{m}\xi_j X_j\right\|_{\mathrm{Op}} \le C_6 \|\xi\|_{2,1}\sqrt{\frac{D}{m}}\log(2D).$$

Next, it follows from Chebyshev's inequality that for any $1 \le l \le k$, with probability $\ge 1 - \frac{p^*(\alpha)}{2}$

$$2\|\Theta_l\|_{\mathrm{Op}} \le \frac{4C_6}{p^*(\alpha)}\|\xi\|_{2,1}\sqrt{\frac{D}{m}}\log(2D).$$

Hence, if $\alpha \in (0, 1/2)$ and

$$\varepsilon \ge \frac{4C_6}{p^*(\alpha)}\|\xi\|_{2,1}\sqrt{\frac{D}{m}}\log(2D),$$

the inequality of Theorem 4.3 (with the confidence parameter equal to $\log(2/p^*(\alpha))$) applied to the estimator $\hat{A}_\varepsilon^l$ gives that with probability $\ge 1 - p^*(\alpha)$

$$\left\|\hat{A}_\varepsilon^* - A_0\right\|_{\mathrm{F}}^2 \le \inf_{A \in \mathbb{L}}\left[2\|A - A_0\|_{\mathrm{F}}^2 + C\left(\varepsilon^2 \operatorname{rank}(A) + R_{\mathbb{L}}^2 s_{m,D}\frac{D}{m} + \frac{1}{m}\right)\right].$$

The claim (4.23) now follows from Theorem 3.1. □

5. Numerical evaluation of the geometric median and simulation results

In this section, we briefly discuss computational aspects of our method in $\mathbb{R}^D$ equipped with the standard Euclidean norm $\|\cdot\|$, and present results of numerical simulation.

5.1. Overview of some numerical algorithms

As was mentioned in the introduction, the function $F(z) := \sum_{j=1}^{k}\|z - x_j\|$ is convex, moreover, its minimum is unique unless $\{x_1, \ldots, x_k\}$ are on the same line.

One of the computationally efficient ways to approximate $\arg\min_{z \in \mathbb{R}^D} F(z)$ is the famous *Weiszfeld's algorithm* [46]: starting from some z_0 in the affine hull of $\{x_1, \ldots, x_k\}$, iterate

$$z_{m+1} = \sum_{j=1}^{k}\alpha_{m+1}^{(j)}x_j, \tag{5.1}$$

where $\alpha_{m+1}^{(j)} = \frac{\|x_j - z_m\|^{-1}}{\sum_{j=1}^{k}\|x_j - z_m\|^{-1}}$. Kuhn proved [25] that Weiszfeld's algorithm converges to the geometric median for all but countably many initial points (additionally, his result states that z_m

converges to the geometric median if none of z_m belong to $\{x_1, \ldots, x_k\}$). It is straightforward to check that (5.1) is actually a gradient descent scheme: indeed, it is equivalent to

$$z_{m+1} = z_m - \beta_{m+1} g_{m+1},$$

where $\beta_{m+1} = \frac{1}{\sum_{j=1}^{k} \|x_j - z_m\|^{-1}}$ and $g_{m+1} = \sum_{j=1}^{k} \frac{z_m - x_j}{\|z_m - x_j\|}$ is the gradient of F (we assume that $z_m \notin \{x_1, \ldots, x_k\}$).

Ostresh [36] proposed a method which avoids the possibility of hitting one of the vertices $\{x_1, \ldots, x_k\}$ by considering the following descent scheme: starting with some z_0 in the affine hull of $\{x_1, \ldots, x_k\}$, let

$$z_{m+1} = z_m - \zeta \tilde{\beta}_{m+1} \tilde{g}_{m+1},$$

where $\zeta \in [1, 2]$, $\tilde{g}_{m+1}$ is the properly defined "generalized" gradient (see [36] for details), and $\tilde{\beta}_{m+1} = \frac{1}{\sum_{j: x_j \neq z_m} \|x_j - z_m\|^{-1}}$. It is shown that z_m converges to the geometric median whenever it is unique. Further improved modifications of original Weiszfeld's method can be found in [44].

For other approaches to fast numerical evaluation of the geometric median, see [6,13,15,37] and references therein.

5.2. Simulation results

### 5.2.1. *Principal component analysis*

Data points $X_1, \ldots, X_{156}$ were sampled from the distribution on $\mathbb{R}^{120}$ such that $X_1 \overset{d}{=} AY$, where the coordinates of Y are independent random variables with density $p(y) = \frac{3y^2}{2(1+|y|^3)^2}$ and A is a full-rank diagonal matrix with 5 "large" eigenvalues $\{5^{1/2}, 6^{1/2}, 7^{1/2}, 8^{1/2}, 9^{1/2}\}$ while the remaining diagonal elements are equal to $\frac{1}{\sqrt{120}}$. Additionally, the data set contained 4 "outliers" $Z_1, \ldots, Z_4$ generated from the uniform distribution on $[-20, 20]^{120}$ and independent of X_i's.

In this case, the usual sample covariance matrix does not provide any useful information about the principal components. However, in most cases our method gave reasonable approximation to the truth. We used the estimator described in Section 4.2 with the number of groups $k = 10$ containing 16 observations each. The error was measured by the spectral norm $\|\widehat{\text{Proj}}_5 - \text{Proj}_5\|_{\text{Op}}$, where $\widehat{\text{Proj}}_5$ is a projector on the eigenvectors corresponding to 5 largest eigenvalues of the estimator. Figures 2, 3 show the histograms of the errors evaluated over 100 runs of the simulation. Figure 4 shows performance of a "thresholded geometric median" estimator which is defined in Section 6 below.

5.2.2. *High-dimensional sparse linear regression*

The following model was used for simulation:

$$Y_j = \lambda_0^T x_j + \xi_j, \qquad j = 1, \ldots, 300,$$
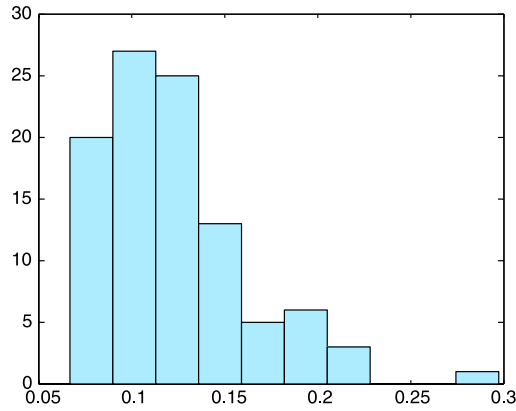
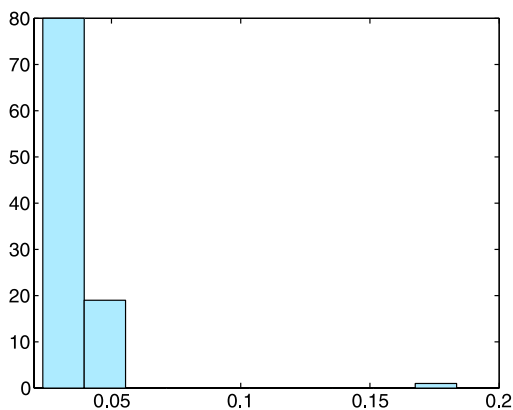**Figure 2.** Error of the "geometric median" estimator (4.7).

where $\lambda_0 \in \mathbb{R}^{1000}$ is a vector with 10 non-zero entries sampled from the uniform distribution on $[-15, 15]$, and $x_j \in \mathbb{R}^{1000}$, $j = 1, \ldots, 300$, are generated according to the normal distribution $N(0, I_{1000})$. Noise ξ_j was sampled from the mixture

$$\xi_j = \begin{cases} \xi_{1,j} & \text{with probability } 1 - 1/500, \\ \xi_{2,j} & \text{with probability } 1/500, \end{cases}$$

where $\xi_{1,j} \sim N(0, 1/8)$ and $\xi_{2,j}$ takes values $\pm\frac{250}{\sqrt{2}}$ with probability $1/2$ each. All parameters $\lambda_0, x_j, \xi_j, j = 1, \ldots, 300$, were sampled independently. Error of the estimator $\hat{\lambda}$ was measured by the ratio $\frac{\|\hat{\lambda} - \lambda_0\|}{\|\lambda_0\|}$. Size of the regularization parameter ε was chosen based on 4-fold cross
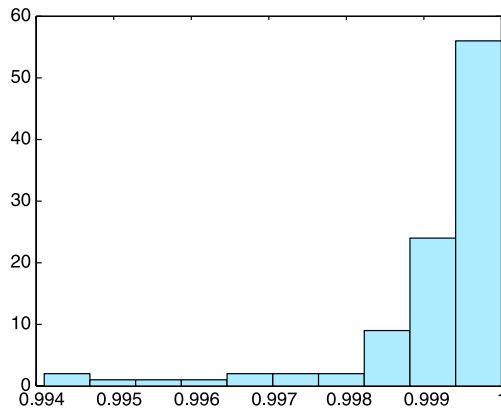

Figure 3. Error of the sample covariance estimator.

Figure 4. Error of the "thresholded geometric median" estimator (6.1), $\nu = 0.5$.

validation. On each stage of the simulation, we evaluated the usual Lasso estimator (4.12) and the "median Lasso" estimator (4.14) based on partitioning the observations into 4 groups of size 75 each. Figures 5 and 6 show the histograms of the errors over 50 runs of the simulation. Note that the maximal error of the "median Lasso" is 0.055 while the error of the usual Lasso exceeded 0.15 in 18 out of 50 cases.

6. Final remarks

Let $\hat{\alpha}_1, \ldots, \hat{\alpha}_k \geq 0$, $\sum_{j=1}^{k} \alpha_j = 1$ be the coefficients such that $\hat{\mu} = \sum_{j=1}^{k} \hat{\alpha}_k \mu_k$ is the geometric median of a collection of estimators $\{\mu_1, \ldots, \mu_k\}$. Our numerical experiments reveal that perfor-
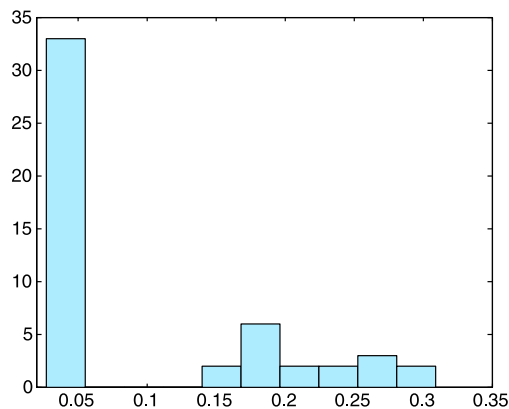
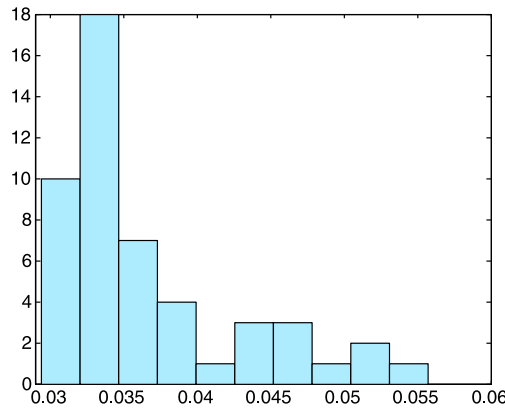**Figure 5.** Error of the standard Lasso estimator (4.12).

Figure 6. Error of the "median Lasso" estimator (4.14).

mance of $\hat{\mu}$ can be significantly improved by setting the coefficients below a certain threshold level ν to 0, that is,

$$
\tilde{\alpha}_j := \frac{\hat{\alpha}_j I\{\hat{\alpha}_j \geq \nu/k\}}{\sum_{i=1}^{k} \hat{\alpha}_j I\{\hat{\alpha}_j \geq \nu/k\}},
$$

$$
\tilde{\mu} := \sum_{j=1}^{k} \tilde{\alpha}_j \mu_j.
$$

(6.1)

An interesting problem that we plan to address in subsequent work is the possibility of adaptive choice of the threshold parameter.

Examples presented above cover only a small area on the map of possible applications. For instance, it would be interesting to obtain an estimator in the low-rank matrix completion framework [11,24] that admits strong performance guarantees for the heavy-tailed noise model. Results obtained in Section 4.4 for the matrix regression problem do not seem to yield a straightforward solution in this case. Another promising direction is related to design of robust techniques for Bayesian inference and evaluation of the geometric median in the space of probability measures. We plan to address these questions in the future work.

Appendix: Proof of Lemma 2.1, part (b)

Once again, assume that the claim does not hold and $\|x_i - z\| \leq r$, $i = 1, \ldots, \lfloor (1-\alpha)k \rfloor + 1$.

We will need the following general description of the subdifferential of a norm $\|\cdot\|$ in a Banach space $\mathbb{X}$ (see, e.g., [21]):

$$
\partial \|x\| = \begin{cases} \{x^* \in \mathfrak{X}^* : \|x^*\|_* = 1, x^*(x) = \|x\|\}, & x \neq 0, \\ \{x^* \in \mathfrak{X}^* : \|x^*\|_* \leq 1\}, & x = 0, \end{cases}
$$

where $\mathbb{X}^*$ is the dual space with norm $\|\cdot\|_*$.

For $x, u \in \mathbb{X}$, let

$$D\big(\|x\|; u\big) = \lim_{t \searrow 0} \frac{\|x + tu\| - \|x\|}{t}$$

be the directional derivative of $\| \cdot \|$ at the point x in direction u. We need the following useful fact from convex analysis:

Lemma A.1. *There exists $g^* := g^*_{x,u} \in \partial \|x\|$ such that $D(\|x\|; u) = \langle g^*, u \rangle$, where $\langle g^*, u \rangle :=$ $g^*(u)$.*

Proof. In follows from the results of Chapter 4 in [21] that $D(\|x\|; u)$ is a continuous convex function of u (for fixed x), hence its subdifferential is non-empty. Let $\tilde{g} \in \partial D(\|x\|; u)$. Then for all $s > 0, v \in \mathbb{X}$

$$s D\big(\|x\|; v\big) = D\big(\|x\|; sv\big) \geq D\big(\|x\|; u\big) + \langle \tilde{g}, sv - u \rangle.$$

Letting $s \to \infty$, we get $D(\|x\|; v) \geq \langle \tilde{g}, v \rangle$, hence $\tilde{g} \in \partial \|x\|$. Taking $s = 0$, we get $D(\|x\|; u) \leq \langle \tilde{g}, u \rangle$, hence $g^* := \tilde{g}$ satisfies the requirement. $\square$

Using Lemma A.1, it is easy to see that there exist $g^*_j \in \partial \|x_j - x_*\|$, $j = 1, \ldots, k$ such that

$$DF\left(x_*; \frac{z - x_*}{\|z - x_*\|}\right) = -\sum_{j: x_j \neq x_*} \frac{\langle g^*_j, z - x_* \rangle}{\|z - x_*\|} + \sum_{j=1}^{k} I\{x_j = x_*\}.$$

Moreover, for any u, $DF(x_*; z - x_*) \geq 0$ by the definition of x_*. Note that for any j

$$\frac{\langle g^*_j, z - x_* \rangle}{\|z - x_*\|} = \frac{\langle g^*_j, x_j - x_* \rangle + \langle g^*_j, z - x_j \rangle}{\|z - x_*\|}. \tag{A.1}$$

By the definition of g^*_j and triangle inequality,

$$\langle g^*_j, x_j - x_* \rangle = \|x_j - x_*\| \geq \|z - x_*\| - \|z - x_j\|$$

and, since $\|g^*_j\|_* \leq 1$,

$$\langle g^*_j, z - x_j \rangle \geq -\|z - x_j\|.$$

Substituting this in (A.1), we get

$$\frac{\langle g^*_j, z - x_* \rangle}{\|z - x_*\|} \geq 1 - 2\frac{\|z - x_j\|}{\|z - x_*\|} > 1 - \frac{2}{C_\alpha},$$

hence

$$DF\left(x_*; \frac{z - x_*}{\|z - x_*\|}\right) < -(1 - \alpha)k\left(1 - \frac{2}{C_\alpha}\right) + \alpha k \leq 0$$

whenever $C_\alpha \geq \frac{2(1-\alpha)}{1-2\alpha}$.

Acknowledgements

References

[1] Alon, N., Matias, Y. and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing* (*Philadelphia*, *PA*, 1996) 20–29. New York: ACM. MR1427494

[2] Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann*. *Statist*. **39** 2766–2794. MR2906886

[3] Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann*. *Statist*. **36** 199–227. MR2387969

[4] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann*. *Statist*. **36** 2577–2604. MR2485008

[5] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann*. *Statist*. **37** 1705–1732. MR2533469

[6] Bose, P., Maheshwari, A. and Morin, P. (2003). Fast approximations for sums of distances, clustering and the Fermat–Weber problem. *Comput*. *Geom*. **24** 135–146. MR1947896

[7] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Trans*. *Inform*. *Theory* **59** 7711–7717. MR3124669

[8] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761

[9] Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *J*. *ACM* **58** Art. 11, 37. MR2811000

[10] Candès, E.J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans*. *Inform*. *Theory* **57** 2342–2359. MR2809094

[11] Candès, E.J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found*. *Comput*. *Math*. **9** 717–772. MR2565240

[12] Candès, E.J., Romberg, J.K. and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm*. *Pure Appl*. *Math*. **59** 1207–1223. MR2230846

[13] Cardot, H., Cénac, P. and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli* **19** 18–43. MR3019484

[14] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann*. *Inst*. *Henri Poincaré Probab*. *Stat*. **48** 1148–1185. MR3052407

[15] Chandrasekaran, R. and Tamir, A. (1990). Algebraic optimization: The Fermat–Weber location problem. *Math*. *Program*. **46** 219–224. MR1047376

[16] Davis, C. and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J*. *Numer*. *Anal*. **7** 1–46. MR0264450

[17] Haldane, J.B.S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35** 414–417.

[18] Hsu, D. and Sabato, S. (2013). Loss minimization and parameter estimation with heavy tails. Preprint. Available at arXiv:1307.1827.

[19] Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2488795

[20] Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statist*. *Sci*. **23** 92–119. MR2431867

[21] Ioffe, A.D. and Tikhomirov, V.M. (1974). *Theory of Extremal Problems*. Moscow: Nauka.

[22] Kemperman, J.H.B. (1987). The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (*Neuchâtel*, 1987) 217–230. Amsterdam: North-Holland. MR0949228

[23] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. *Lectures from the 38th Probability Summer School held in Saint-Flour 2008. École d'Été de Probabilités de Saint-Flour*. Heidelberg: Springer. MR2829871

[24] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann*. *Statist*. **39** 2302–2329. MR2906869

[25] Kuhn, H.W. (1973). A note on Fermat's problem. *Math*. *Program*. **4** 98–107. MR0316102

[26] Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron*. *J*. *Stat*. **5** 1015–1053. MR2836768

[27] Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann*. *Statist*. **40** 1024–1060. MR2985942

[28] Lerasle, M. and Oliveira, R.I. (2011). Robust empirical mean estimators. Preprint. Available at arXiv:1112.3914.

[29] Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437

[30] Minsker, S. (2013). Geometric median and robust estimation in Banach spaces. Preprint. Available at http://sminsker.wordpress.com/publications/.

[31] Negahban, S. and Wainwright, M.J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann*. *Statist*. **39** 1069–1097. MR2816348

[32] Negahban, S. and Wainwright, M.J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J*. *Mach*. *Learn*. *Res*. **13** 1665–1697. MR2930649

[33] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Berlin: Springer. MR1775640

[34] Nemirovski, A. and Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley.

[35] Nguyen, N.H. and Tran, T.D. (2013). Robust Lasso with missing and grossly corrupted observations. *IEEE Trans*. *Inform*. *Theory* **59** 2036–2058. MR3043781

[36] Ostresh, L.M. Jr. (1978). On the convergence of a class of iterative methods for solving the Weber location problem. *Oper*. *Res*. **26** 597–609. MR0496728

[37] Overton, M.L. (1983). A quadratically convergent method for minimizing a sum of Euclidean norms. *Math*. *Program*. **27** 34–63. MR0712109

[38] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Lond*. *Edinb*. *Dubl*. *Phil*. *Mag*. *J*. *Sci*. **2** 559–572.

[39] Recht, B., Fazel, M. and Parrilo, P.A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev*. **52** 471–501. MR2680543

[40] Rohde, A. and Tsybakov, A.B. (2011). Estimation of high-dimensional low-rank matrices. *Ann*. *Statist*. **39** 887–930. MR2816342

[41] Small, C. (1990). A survey of multidimensional medians. *Internat*. *Statist*. *Rev*. **58** 263–277.

[42] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **58** 267–288. MR1379242

[43] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer. MR1385671

[44] Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97** 1423–1426 (electronic). MR1740461

[45] Weber, A. (1929). *Uber Den Standort der Industrien* (*Alfred Weber's Theory of the Location of Industries*). Chicago, IL: Univ. Chicago Press.

[46] Weiszfeld, E. (1937). Sur un problème de minimum dans l'espace. *Tohoku Math. J.* (2) **42** 274–280.

[47] Wright, J. and Ma, Y. (2010). Dense error correction via ℓ_1-minimization. *IEEE Trans. Inform. Theory* **56** 3540–3560. MR2799012

[48] Zhang, T. and Lerman, G. (2014). A novel M-estimator for robust PCA. *J. Mach. Learn. Res.* **15** 749–808.

[49] Zwald, L. and Blanchard, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems* **18** 1649–1656. Cambridge, MA: MIT Press.