Consistent Estimation for PCA and Sparse Regression with Oblivious Outliers

Anonymous Author(s)
Affiliation
Address
email

Abstract

We develop machinery to design efficiently computable, *consistent* estimators, achieving estimation error approaching zero as the number of observation grows, when facing an oblivious adversary that may corrupt all but an α fraction of the samples. As concrete examples, we investigate two problems: sparse regression and principal component analysis (PCA). For sparse regression, we achieve consistency for optimal sample size $n \gtrsim (k \log d)/\alpha^2$ and optimal error rate $O(\sqrt{(k \log d)/(n \cdot \alpha^2)})$ where n is the number of observations, d is the number of dimensions and k is the sparsity of the parameter vector, allowing the fraction of inliers to be inverse-polynomial in the number of samples. Prior to this work, no estimator was known to be consistent when the fraction of inliers α is $o(1/\log \log n)$, even for (non-spherical) Gaussian design matrices. Results holding in the presence of such general noise has only been shown in dense setting (i.e., general linear regression) very recently by d'Orsi et al. [dNS21]. In the context of PCA, we attain optimal error guarantees under broad spikiness assumptions on the parameter matrix (usually used in matrix completion). Previous works could obtain non-trivial guarantees only under the assumptions that the measurement noise corresponding to the inliers is polynomially small in n (e.g., Gaussian with variance $1/n^2$).

To devise our estimators, we equip the Huber loss with non-smooth regularizers such as the ℓ_1 norm or the nuclear norm, and extend d'Orsi et al.'s approach [dNS21] in a novel way to analyze the loss function. Our machinery appears to be easily applicable to a wide range of estimation problems. We complement these algorithmic results with statistical lower bounds showing that the fraction of inliers that our PCA estimator can deal with is optimal up to a constant factor.

1 Introduction

Estimating information from structured data is a central theme in statistics that by now has found applications in a wide array of disciplines. On a high level, a typical assumption in an estimation problem is the existence of a –known *a priori*– family of probability distributions $\mathcal{P} := \left\{ \mathbb{P}_\beta \;\middle|\; \beta \in \Omega \right\}$ over some space $\mathcal{Z}$ that are each indexed by some parameter $\beta \in \Omega$. We then observe a collection of n independent observations $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ drawn from an unknown probability distribution $\mathbb{P}_{\beta^*} \in \mathcal{P}$. The goal is to (approximately) recover the *hidden* parameter β^*.[1]

Oftentimes, real-world data may contain skewed, imprecise or corrupted measurements. Hence, a desirable property for an estimator is to be robust to significant, possibly malicious, noise perturbations

[1] In this paper, we assume $\Omega \subseteq \mathbb{R}^d$ and $\mathcal{Z} \subseteq \mathbb{R}^D$ for some d, D, denote random variables in bold face, and hide absolute constant factors with the notation $O(\cdot), \Omega(\cdot), \gtrsim, \lesssim$ and logarithmic factors with $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

on the given observations. Indeed, in the last two decades, a large body of work has been developed on designing robust algorithms (e.g see [BGN09, BBC11, MMYSB19]). However, proving strong guarantees often either demands strong assumptions on the noise model or requires that the fraction of perturbed observations is small. More concretely, when we allow the noise to be chosen adaptively, i.e., chosen dependently on the observations and hidden parameters, a common theme is that *consistent* estimators – estimators whose error tends to zero as the number of observations grows – can be attained only when the fraction of outliers is small.

In order to make vanishing error possible in the presence of large fractions of outliers, it is necessary to consider weaker adversary models that are *oblivious* to the underlying structured data. In recent years, a flurry of works have investigated oblivious noise models [CLMW11, ZLW+10, TJSO14, BJKK17, SBRJ19, dNS21]. These results however are tailor-made to the specific models and problems. To overcome this limitation, in this paper we aim to provide a simple blueprint to design provably robust estimators under minimalistic noise assumptions for a large class of estimation problems. As a testbed for our blueprint, we investigate two well-studied problems:

Principal component analysis (PCA): Given a matrix $\boldsymbol{Y} := L^* + \boldsymbol{N}$ where $L^* \in \mathbb{R}^{n\times n}$ is an unknown parameter matrix and $\boldsymbol{N}$ is an n-by-n *random* noise matrix, the goal is to find an estimator $\hat{\boldsymbol{L}}$ for L^* which is as close as possible to L^* in Frobenius norm.

Sparse regression: Given observations $(X_1, \boldsymbol{y}_1), \dots, (X_n, \boldsymbol{y}_n)$ following the linear model $\boldsymbol{y}_i = \langle X_i, \beta^* \rangle + \boldsymbol{\eta}_i$ where $X_i \in \mathbb{R}^d$, $\beta^* \in \mathbb{R}^d$ is the k-sparse parameter vector of interest (by k-sparse we mean that it has at most k nonzero entries) and $\boldsymbol{\eta}_1, \dots \boldsymbol{\eta}_n$ is noise, the goal is to find an estimator $\hat{\boldsymbol{\beta}}$ for β^* achieving small *prediction error* $\frac{1}{\sqrt{n}}\left\|X(\hat{\boldsymbol{\beta}} - \beta^*)\right\|$, where X is the matrix whose rows are $X_1, \dots, X_n$.[2]

Principal component analysis. A natural way to describe principal component analysis under oblivious preturbations is that of assuming the noise matrix $\boldsymbol{N}$ to be an n-by-n matrix with a uniformly random set of $\alpha \cdot n^2$ entries bounded by some small $\zeta \geqslant 0$ in absolute value. In these settings, we may think of the α fraction of entries with small noise as the set of uncorrupted observations. Moreover, for $\zeta > 0$, the fact that even for uncorrupted observations the noise is non-zero allows us to capture both gross sparse errors and small entry-wise noise in the measurements at the same time (for example if $\zeta = 1$ then the model captures settings with additional standard Gaussian noise). Remarkably, for $\zeta = 0$, Candès et al.'s seminal work [CLMW11] provided an algorithm that exactly recovers L^* even for a vanishing fraction of inliers, under *incoherence* conditions on the signal L^*. The result was slightly extended in [ZLW+10] where the authors provided an algorithm recovering L^* up to squared error $O(\zeta^2 \cdot n^4)$ thus allowing polynomially small measurement noise ζ, but still failing to capture settings where standard Gaussian measurement noise is added to the sparse noise. Even for simple signal matrices L^*, prior to this work, it remained an open question whether consistent estimators could be designed in presence of both oblivious noise and more reasonable measurement noise (e.g., standard Gaussian).

Linear regression. Similarly to the context of principal component analysis, a convenient model for oblivious adversarial corruptions is that of assuming the noise vector $\boldsymbol{\eta}$ to have a random set of $\alpha \cdot n$ entries bounded by ζ in absolute value (here all results of interest can be extended to any $\zeta > 0$ just by scaling, so we consider only the case $\zeta = 1$). This model trivially captures the classical settings with Gaussian noise (a Gaussian vector with variance σ^2 will have an $\alpha = \Theta(\frac{1}{\sigma})$ fraction of inliers) and, again, allows us to think of the α-fraction of entries with small noise as the set of uncorrupted observations. Early works on consistent regression focused on the regime with Gaussian design $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \sim N(0, \Sigma)$ and deterministic noise η.[3] [BJKK17] presented an estimator achieving error $\tilde{O}(d/(\alpha^2 \cdot n))$ for any α larger than a fixed constant. [SBRJ19] extended this result by achieving comparable error rates even for a vanishing fraction of inliers $\alpha \gtrsim 1/\log\log n$. Assuming $n \gtrsim d^2/\alpha^2$ [TJSO14] proved that the Huber-loss estimator [Hub64] achieves optimal error rate $O(d/\alpha^2 \cdot n)$ even

[2] Our analysis also works for the *parameter error* $\|\beta^* - \hat{\boldsymbol{\beta}}\|$.

[3] For Gaussian design $\boldsymbol{X}$ one may also consider a deterministic noise model. This noise model is subsumed by the random noise model discussed here. As shown in [dNS21], roughly speaking the underlying reason is that the Gaussianity of $\boldsymbol{X}$ allows one to obtain several other desirable properties "for free". For example, one could ensure that the noise vector is symmetric by randomly flipping the sign of each observation $(\boldsymbol{y}_i, \boldsymbol{X}_i)$, as the design matrix will still be Gaussian. See Section 2 for a more in-depth discussion.

for polynomially small fraction of inliers $\alpha \gtrsim \sqrt{d/n}$. This line of work culminated in [dNS21] which extended the result of [TJSO14] achieving the same guarantees with optimal sample complexity $n \gtrsim d/\alpha^2$. Furthermore, the authors of [dNS21] extended these guarantees to deterministic design matrices satisfying only a spreadness condition (trivially satisfied by sub-Gaussian design matrices).

In the context of sparse regression, less is known. When the fraction of observations is constant $\alpha \geqslant \Omega(1)$, [NTN11] provided a first consistent estimator albeit with sub-optimal error rate. For Gaussian designs, the algorithm of [SBRJ19] achieves nearly optimal error convergence $\tilde{O}\big((k \log^2 d)/\alpha^2 n\big)$, but requires $\alpha \gtrsim 1/\log\log n$. More recently, [dNS21] presented an algorithm for standard Gaussian design $\boldsymbol{X} \sim N(0,1)^{n\times d}$, achieving the nearly optimal error convergence $\tilde{O}(k/(\alpha^2 \cdot n))$ for nearly optimal inliers fraction $\alpha \gtrsim \sqrt{(k \cdot \log^2 d)/n}$. The first two works use an iterative process, and require a bound $\|\beta^*\| \leqslant d^{O(1)}$ (for larger $\|\beta^*\|$ these algorithms also work, but the fraction of inliers or the error convergence is worse). The algorithm from [dNS21] , however, heavily relies on the assumption that $\boldsymbol{X} \sim N(0,1)^{n\times d}$ and appears unlikely to be generalizable to more general families of matrices (including non-spherical Gaussian designs).

Our contribution. We propose new machinery to design efficiently computable *consistent estimators* achieving optimal error rates and sample complexity against oblivious outliers. In particular, we extend the approach of [dNS21] to structured estimation problems, by finding a way to exploit adequately the structure therein. While consistent estimators have already been designed under more benign noise assumptions (e.g. the LASSO estimator for sparse linear regression under Gaussian noise), it was previously unclear how to exploit this structure in the setting of oblivious noise. One key consequence of our work is hence to demonstrate what minimal assumptions on the noise are sufficient to make effective recovery (in the sense above) possible. Concretely, we show

Oblivious PCA: Under mild assumptions on the noise matrix $\boldsymbol{N}$ and common assumption on the parameter matrix L^* –traditionally applied in the context of matrix completion [NW12]– we provide an algorithm that achieves optimal error guarantees.

Sparse regression: Under mild assumptions on the design matrix and the noise vector –similar to the ones used in [dNS21] for dense parameter vectors β^* – we provide an algorithm that achieves optimal error guarantees and sample complexity.

For both problems, our analysis improves over the state-of-the-art and recovers the classical optimal guarantees, not only for Gaussian noise, but also under much less restrictive noise assumptions. At a high-level, we achieve the above results by equipping the Huber loss estimator with appropriate regularizers. Our techniques closely follow standard analyses for M-estimators, but crucially depart from them when dealing with the observations with large perturbations. Furthermore, our analysis appears to be mechanical and thus easily applicable to many different estimation problems.

2 Results

Our estimators are based on regularized versions of the *Huber loss*. The regularizer we choose depends on the underlying structure of the estimation problem: We use ℓ_1 regularization to enforce sparsity in linear regression and nuclear norm regularization to enforce a low-rank structure in the context of PCA. More formally, the *Huber penalty* is defined as the function $f_h : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ such that

$$f_h(t) := \begin{cases} \frac{1}{2}t^2 & \text{for } |t| \leqslant h\,, \\ h(|t| - \frac{h}{2}) & \text{otherwise.} \end{cases} \tag{2.1}$$

where $h > 0$ is a penalty parameter. For $X \in \mathbb{R}^D$, the *Huber loss* is the piecewise function $F_h(X) := \sum_{i\in[D]} f_h(X_i)$. We will define the regularized versions in the following sections. For a matrix A, we use $\|A\|$, $\|A\|_{\text{nuc}}$, $\|A\|_{\text{F}}$, $\|A\|_{\max}$ to denote its spectral, nuclear, Frobenius, maximum norms, respectively. For a vector v, we use $\|v\|$ and $\|v\|_1$ to denote its ℓ_2 and ℓ_1 norms.

2.1 Oblivious principal component analysis

For oblivious PCA, we provide guarantees for the following estimator (ζ will be defined shortly):

$$\hat{\boldsymbol{L}} := \operatorname*{argmin}_{L\in\mathbb{R}^{n\times n},\, \|L\|_{\max}\leqslant \rho/n} \Big(F_h(\boldsymbol{Y} - L) + 100\sqrt{n}\big(\zeta + \rho/n\big)\|L\|_{\text{nuc}} \Big). \tag{2.2}$$

Theorem 2.1. *Let $L^* \in \mathbb{R}^{n\times n}$ be an unknown deterministic matrix and let $\boldsymbol{N}$ be an n-by-n random matrix with independent, symmetrically distributed (about zero) entries and $\alpha := \min_{i,j\in[n]} \mathbb{P}\{|\boldsymbol{N}_{ij}| \leqslant \zeta\}$ for some $\zeta \geqslant 0$. Suppose that $\operatorname{rank}(L^*) = r$ and $\|L^*\|_{\max} \leqslant \rho/n$. Then, with probability at least $1 - 2^{-n}$ over $\boldsymbol{N}$, given $\boldsymbol{Y} = L^* + \boldsymbol{N}$, ζ and ρ, the estimator Eq. (2.2) with Huber parameter $h = \zeta + \rho/n$ satisfies*

$$\left\|\hat{\boldsymbol{L}} - L^*\right\|_{\mathrm{F}} \leqslant O\left(\frac{\sqrt{rn}}{\alpha}\right) \cdot (\zeta + \rho/n)\,.$$

We first compare the guarantees of Theorem 2.1 with the previous results on robust PCA [CLMW11, ZLW+10].[4] The first difference is that they require L^* to satisfy certain incoherent conditions.[5] Concretely, they provide theoretical guarantees for $r \leqslant O\big(\mu^{-1} n (\log n)^{-2}\big)$, where μ is the incoherence parameter. In certain regimes, such a constraint strongly binds with the eigenvectors of L^*, restricting the set of admissible signal matrices. Using the different assumption $\|L^*\|_{\max} \leqslant \rho/n$ (commonly used for matrix completion, see [NW12]), we can obtain nontrivial guarantees (i.e. $\left\|\hat{\boldsymbol{L}} - L^*\right\|_{\mathrm{F}} / \|L^*\|_{\mathrm{F}} \to 0$ as $n \to \infty$) even when the μ-incoherence conditions are not satisfied for any $\mu \leqslant n/\log^2 n$, and hence the results [CLMW11, ZLW+10] cannot be applied. We remark that, without assuming incoherence, the dependence of the error on the maximal entry of L^* is inherent (see Remark 3.2).

The second difference is that Theorem 2.1 provides a significantly better dependence on the magnitude ζ of the entry-wise measurement error. Specifically, in the settings of Theorem 2.1, [ZLW+10] showed that if L^* satisfies the incoherence conditions, the error of their estimator is $O\big(n^2\zeta\big)$. If the entries of $\boldsymbol{N}$ are standard Gaussian with probability α (and hence $\zeta \leqslant O(1)$), and the entries of L^* are bounded by $O(1)$, then the error of our estimator is $O(\sqrt{rn}/\alpha)$, which is considerably better than $O(n^2)$ as in [ZLW+10]. On the other hand, their error does not depend on the magnitude ρ/n of the signal entries, so in the extreme regimes when the singular vectors of L^* satisfy the incoherence conditions but L^* has very large singular values (so that the magnitude of the entries of L is significantly larger than n), their analysis provides better guarantees than Theorem 2.1.

As another observation to understand Theorem 2.1, notice that our robust PCA model also captures the classical matrix completion settings. In fact, any instance of matrix completion can be easily transformed into an instance of our PCA model: for the entries (i, j) that we do not observe, we can set $N_{i,j}$ to some arbitrarily large value $\pm C(\rho, n) \gg \rho/n$, making the signal-to-noise ratio of the entry arbitrarily small. The observed (i.e. uncorrupted) $\Theta(\alpha \cdot n^2)$ entries may additionally be perturbed by Gaussian noise with variance $\Theta(\zeta^2)$. The error guarantees of the estimator in Theorem 2.1 is $O\big((\rho/n + \zeta)\sqrt{rn}/\alpha\big)$. Thus, the dependency on the parameters ρ, n, ζ, and r is the *same* as in matrix completion and the error is within a factor of $\Theta(\sqrt{1/\alpha})$ from the optimum for matrix completion. However, this worse dependency on α is intrinsic to the more general model considered and it turns out to be optimal (see Theorem 2.4). On a high level, the additional factor of $\Theta(\sqrt{1/\alpha})$ comes from the fact that in our PCA model we don't know which entries are corrupted. The main consequence of this phenomenon is that a condition of the form $\alpha \gtrsim \sqrt{r/n}$ appears *inherent* to achieve consistency. To get some intuition on why this condition is necessary, consider the Wigner model where we're given a matrix $\boldsymbol{Y} = xx^\top + \sigma \boldsymbol{W}$ for a flat vector $x \in \{\pm 1\}^n$ and a standard Gaussian matrix $\boldsymbol{W}$. Note, that the entries of $\boldsymbol{W}$ fit our noise model for $\zeta = 1\,, \rho/n = 1\,, r = 1$ and $\alpha = \Theta(1/\sigma)$. The spectral norm of $\sigma\boldsymbol{W}$ concentrates around $2\sigma\sqrt{n}$ and thus it is information-theoretically not possible to approximately recover the vector x for $\sigma = 1/\alpha \gtrsim \sqrt{n}$ (see [PWBM16]).

2.2 Sparse regression

Our regression model considers a *fixed* design matrix $X \in \mathbb{R}^{n\times d}$ and observations $\boldsymbol{y} := X\beta^* + \boldsymbol{\eta} \in \mathbb{R}^n$ where β^* is an *unknown* k-sparse parameter vector and $\boldsymbol{\eta}$ is *random* noise with $\mathbb{P}(|\eta_i| \leqslant 1) \geqslant \alpha$ for all $i \in [n]$. Earlier works [BJKK17], [SBRJ19] focused on the setting that the design matrix

[4] We remark that in [CLMW11] the authors showed that they can consider non-symmetric noise when the fraction of inliers is large $\alpha \geqslant 1/2$. However for smaller fraction of inliers their analysis requires the entries of the noise to be symmetric and independent, so for $\alpha < 1/2$, their assumptions are captured by Theorem 2.1.

[5] A rank-r $n \times n$ dimensional matrix M is *μ-incoherent* if its *singular vector decomposition* $M := U\Sigma V^\top$ satisfies $\max_{i\in[n]} \|U^\top e_i\|^2 \leqslant \frac{\mu r}{n}$, $\max_{i\in[n]} \|V^\top e_i\|^2 \leqslant \frac{\mu r}{n}$ and $\|UV^\top\|_\infty \leqslant \frac{\sqrt{\mu r}}{n}$.

consists of i.i.d. rows with Gaussian distribution $\mathcal{N}(0, \Sigma)$ and the noise is $\eta = \zeta + \boldsymbol{w}$ where ζ is deterministic $(\alpha \cdot n)$-sparse vector and $\boldsymbol{w}$ is subgaussian. As in [dNS21], our results for a fixed design and random noise can, in fact, extend to yield the same guarantees for this early setting (see Theorem 2.3). Hence, a key advantage of our results is that the design X does not have to consist of Gaussian entries. Remarkably, we can handle arbitrary deterministic designs as long as they satisfy some mild conditions. Concretely, we make the following three assumptions, the first two of which are standard in the literature of sparse regression (e.g., see [Wai19]):

1. For every column X^i of X, $\left\|X^i\right\| \leqslant \sqrt{n}$.
2. *Restricted eigenvalue property (RE-property)*: For every vector $u \in \mathbb{R}^d$ such that[6] $\left\|u_{\mathrm{supp}(\beta^*)}\right\| \geqslant 0.1 \cdot \|u\|$, $\frac{1}{n}\|Xu\|^2 \geqslant \lambda \cdot \|u\|^2$ for some parameter $\lambda > 0$.
3. *Well-spreadness property*: For some $m \in [n]$ and for every vector $u \in \mathbb{R}^d$ such that $\left\|u_{\mathrm{supp}(\beta^*)}\right\| \geqslant 0.1 \cdot \|u\|$ and for every subset $S \subseteq [n]$ with $|S| \geqslant n - m$, it holds that $\|(Xu)_S\| \geqslant \frac{1}{2}\|Xu\|$.

We devise our estimator for sparse regression and state its statistical guarantees below:

$$\hat{\boldsymbol{\beta}} := \arg\min_{\beta \in \mathbb{R}^d} \left(\boldsymbol{F}_2(\beta) + 100\sqrt{n \log d} \cdot \left\|\beta\right\|_1 \right). \tag{2.3}$$

Theorem 2.2. *Let $\beta^* \in \mathbb{R}^d$ be an unknown k-sparse vector and let $X \in \mathbb{R}^{n \times d}$ be a deterministic matrix such that for each column X^i of X, $\|X^i\| \leqslant \sqrt{n}$, satisfying the RE-property with $\lambda > 0$ and well-spreadness property with $m \gtrsim \frac{k \log d}{\lambda \cdot \alpha^2}$ (recall that $n \geqslant m$). Further, let $\boldsymbol{\eta}$ be an n-dimensional random vector with independent, symmetrically distributed (about zero) entries and $\alpha = \min_{i \in [n]} \mathbb{P}\left\{\left|\boldsymbol{\eta}_i\right| \leqslant 1\right\}$. Then with probability at least $1 - d^{-10}$ over $\boldsymbol{\eta}$, given X and $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$, the estimator Eq. (2.3) satisfies*

$$\frac{1}{n}\left\|X\left(\hat{\boldsymbol{\beta}} - \beta^*\right)\right\|^2 \leqslant O\left(\frac{1}{\lambda} \cdot \frac{k \log d}{\alpha^2 \cdot n}\right) \qquad \textit{and} \qquad \left\|\hat{\boldsymbol{\beta}} - \beta^*\right\|^2 \leqslant O\left(\frac{1}{\lambda^2} \cdot \frac{k \log d}{\alpha^2 \cdot n}\right).$$

There are important considerations when interpreting this theorem. The first is the special case $\eta \sim N(0, \sigma^2)^n$, which satisfies our model for $\alpha = \Theta(1/\sigma)$. For this case, it is well known (e.g., in [Wai19]) that under the same RE assumption, the LASSO estimator achieves a prediction error rate of $O(\frac{\sigma^2}{\lambda} \cdot \frac{k \log d}{n}) = O(\frac{k \log d}{\lambda \cdot \alpha^2 \cdot n})$, matching our result. Moreover, this error rate is essentially optimal. Under a standard assumption in complexity theory ($\mathbf{NP} \not\subset \mathbf{P/poly}$), the RE assumption is necessary when considering polynomial-time estimators [ZWJ14]. Further, this also shows that the dependence on the RE constant seems unavoidable. Under mild conditions on the design matrix, trivially satisfied if the rows are i.i.d. Gaussian with covariance Σ whose condition number is constant, our guarantees are optimal up to constant factors for *all* estimators if $k \leqslant d^{1-\Omega(1)}$ (e.g. $k \leqslant d^{0.99}$), see [RWY11]. This optimality also shows that our bound on the number of samples is best possible since otherwise we would not be able to achieve vanishing error. The (non-sparse version of) well-spreadness property was first used in the context of regression in [dNS21]. In the same work the authors also showed that, under oblivious noise assumptions, some weak form of spreadness property is indeed necessary.

The second consideration is the *optimal* dependence on α: Theorem 2.2 achieves consistency as long as the fraction of inliers satisfies $\alpha = \omega(k \log d/n)$. To get an intuition, observe that lower bounds for standard sparse regression show that, already for $\boldsymbol{\eta} \sim N(0, \sigma \cdot \mathrm{Id}_n)$, it is possible to achieve consistency only for $n = \omega(\sigma^2 k \log d)$ (if $k \leqslant d^{1-\Omega(1)}$). As for this $\boldsymbol{\eta}$, the number of entries of magnitude at most 1 is $O(n/\sigma)$ with high probability, it follows that for $\alpha = 1/\sigma \leqslant \sqrt{(k \log d)/n}$, no estimator is consistent.

To the best of our knowledge, Theorem 2.2 is the first result to achieve consistency under such minimalistic noise settings and deterministic designs. Previous results [BJKK17, SBRJ19, dNS21] focused on simpler settings of Gaussian design $\boldsymbol{X}$ and deterministic noise, and provide no guarantees for more general models. The techniques for Theorem 2.2 also extend to this case.

Theorem 2.3. *Let $\beta^* \in \mathbb{R}^d$ be an unknown k-sparse vector and let $\boldsymbol{X}$ be a n-by-d random matrix with i.i.d. rows $\boldsymbol{X}_1, \ldots \boldsymbol{X}_n \sim N(0, \Sigma)$ for a positive definite matrix Σ. Further, let $\eta \in \mathbb{R}^n$ be a*

[6] For a vector $v \in \mathbb{R}^d$ and a set $S \subseteq [d]$, we denote by v_S the restriction of v to the coordinates in S.

deterministic vector with $\alpha \cdot n$ coordinates bounded by 1 in absolute value. Suppose that $n \gtrsim \frac{\nu(\Sigma) \cdot k \log d}{\sigma_{\min}(\Sigma) \cdot \alpha^2}$, where $\nu(\Sigma)$ is the maximum diagonal entry of Σ and $\sigma_{\min}(\Sigma)$ is its smallest eigenvalue. Then, with probability at least $1 - d^{-10}$ over $\boldsymbol{X}$, given $\boldsymbol{X}$ and $\boldsymbol{y} = \boldsymbol{X}\beta^* + \eta$, the estimator Eq. (2.3) satisfies

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \beta^*\right)\right\|^2 \leqslant O\left(\frac{\nu(\Sigma) \cdot k \log d}{\sigma_{\min}(\Sigma) \cdot \alpha^2 \cdot n}\right) \qquad \textit{and} \qquad \left\|\hat{\boldsymbol{\beta}} - \beta^*\right\|^2 \leqslant O\left(\frac{\nu(\Sigma) \cdot k \log d}{\sigma_{\min}^2(\Sigma) \cdot \alpha^2 \cdot n}\right).$$

Even for standard Gaussian design $X \sim N(0,1)^{n \times d}$, the above theorem improves over previous results, which required sub-optimal sample complexity $n \gtrsim \left(k/\alpha^2\right) \cdot \log d \cdot \log\left\|\beta^*\right\|$. For non-spherical Gaussian designs, the improvement over state of the art [SBRJ19] is more serious: their algorithm requires $\alpha \geqslant \Omega(1/\log\log n)$, while our Theorem 2.3 doesn't have such restrictions and works for all $\alpha \gtrsim \sqrt{\frac{\nu(\Sigma)}{\sigma_{\min}(\Sigma)} \cdot \frac{k \log d}{n}}$; in many interesting regimes α is allowed to be smaller than $n^{-\Omega(1)}$.

Note that while we can deal with general covariance matrices, to compare Theorem 2.3 with Theorem 2.2 it is easier to consider Σ in a normalized form, when $\nu(\Sigma) \leqslant 1$. This can be easily achieved by scaling $\boldsymbol{X}$. Also note that Theorem 2.2 can be generalized to the case $\|X^i\| \leqslant \nu\sqrt{n}$ for arbitrary $\nu > 0$, and then the error bounds and the bound on m should be multiplied by ν.

The RE-property of Theorem 2.2 is a standard assumption in sparse regression and is satisfied by a large family of matrices. For example, with high probability a random matrix $\boldsymbol{X}$ with i.i.d. rows sampled from $\mathcal{N}(0, \Sigma)$, with positive definite $\Sigma \in \mathbb{R}^{d \times d}$ whose diagonal entries are bounded by 1, satisfies the RE-property with parameter $\Omega(\sigma_{\min}(\Sigma))$ for all subsets of $[d]$ of size k (so for every possible support of β^*) as long as long as $n \gtrsim \frac{1}{\sigma_{\min}(\Sigma)} \cdot k \log d$ (see [Wai19]). The well-spread assumption is satisfied for such $\boldsymbol{X}$ with for all sets $S \subset [n]$ of size $m \leqslant cn$ (for sufficiently small c) and for all subsets of $[d]$ of size k as long as $n \gtrsim \frac{1}{\sigma_{\min}(\Sigma)} \cdot k \log d$.

2.3 Optimal fraction of inliers for principal component analysis under oblivious noise

We show here that the dependence on α we obtain in Theorem 2.1 is information theroetically optimal up to constant factors. Concretely, let $L^*, \boldsymbol{N}, \boldsymbol{Y}, \alpha, \rho$ and ζ be as in Theorem 2.1, and let $0 < \varepsilon < 1$ and $0 < \delta < 1$. A successful (ε, δ)-weak recovery algorithm for PCA is an algorithm that takes $\boldsymbol{Y}$ as input and returns a matrix $\hat{\boldsymbol{L}}$ such that $\left\|\hat{\boldsymbol{L}} - L^*\right\|_{\mathrm{F}} \leqslant \varepsilon \cdot \rho$ with probability at least $1 - \delta$.

It can be easily seen that the Huber-loss estimator of Theorem 2.1 fails to be a successful weak-recovery algorithm if $\alpha = o(\sqrt{r/n})$ (for both cases $\zeta \leqslant \rho/n$ and $\rho/n \leqslant \zeta$, we need $\alpha = \Omega(\sqrt{r/n})$.) A natural question to ask is whether the condition $\alpha = \Omega(\sqrt{r/n})$ is necessary in general. The following theorem shows that if $\alpha = o(\sqrt{r/n})$, then weak-recovery is information-theoretically impossible. This means that the (polynomially small) fraction of inliers that the Huber-loss estimator of Theorem 2.1 can deal with is optimal up to a constant factor.

Theorem 2.4. *There exists a universal constant $C_0 > 0$ such that for every $0 < \varepsilon < 1$ and $0 < \delta < 1$, if $\alpha := \min_{i,j \in [n]} \mathbb{P}[|\boldsymbol{N}_{i,j}| \leqslant \zeta]$ satisfies $\alpha < C_0 \cdot (1 - \varepsilon^2)^2 \cdot (1 - \delta) \cdot \sqrt{r/n}$, and n is large enough, then it is information-theoretically impossible to have a successful (ε, δ)-weak recovery algorithm. The problem remains information-theoretically impossible (for the same regime of parameters) even if we assume that L^* is incoherent; more precisely, even if we know that L^* has incoherence parameters that are as good as those of a random flat matrix of rank r, the theorem still holds.*

3 Techniques

To illustrate our techniques in proving statistical guarantees for the Huber-loss estimator, we first use sparse linear regression as a running example. Then, we discuss how the same ideas apply to principal component analysis. Finally, we also remark our techniques for lower bounds.

3.1 Sparse linear regression under oblivious noise

We consider the model of Theorem 2.2. Our starting point to attain the guarantees for our estimator Eq. (2.3), i.e., $\hat{\boldsymbol{\beta}} := \arg\min_{\beta \in \mathbb{R}^d} F_2(\beta) + 100\sqrt{n \log d}\left\|\beta\right\|_1$, is a classical approach for M-estimators

(see e.g. [Wai19]). For simplicity, we will refer to $F_2(\beta)$ as the loss function and to $\|\beta\|_1$ as the regularizer. At a high level, it consists of the following two ingredients:

(I) an upper bound on some norm of the gradient of the loss function at the parameter β^*,

(II) a lower bound on the curvature of the loss function (in form of a local strong convexity bound) within a structured neighborhood of β^*. The structure of this neighborhood can roughly be controlled by choosing the appropriate regularizer.

The key aspect of this strategy is that the strength of the statistical guarantees of the estimator crucially depends on the *directions* and the *radius* in which we can establish lower bounds on the curvature of the function. Since these features are inherently dependent on the landscape of the loss function and the regularizer, they may differ significantly from problem to problem. This strategy has been applied successfully for many related problems such as compressed sensing or matrix completion albeit with standard noise assumptions.[7] Under oblivious noise, [dNS21] used a particular instantiation of this framework to prove optimal convergence of the Huber-loss – without any regularizer – for standard linear regression. Such estimator, however, doesn't impose any structure on the neighborhood of β^* considered in (II) and thus, can only be used to obtain sub-optimal guarantees for sparse regression.

In the context of sparse regression, the above two conditions translate to: (I) an upper bound on the largest entry in absolute value of the gradient of the loss function at β^*, and (II) a lower bound on the curvature of $\boldsymbol{F}_2$ within the set of *approximately* k-sparse vectors[8] close to β^*. We use this recipe to show that *all* approximate minimizers of $\boldsymbol{F}$ are close to β^*. While the idea of restricting to only approximately sparse directions has also been applied for the LASSO estimator in sparse regression under standard (sub)-Gaussian noise, in the presence of oblivious noise, our analysis of the Huber-loss function requires a more careful approach.

More precisely, under the assumptions of Theorem 2.2, the error bound can be computed as

$$O\left(\frac{s\|G\|^*_{\mathrm{reg}}}{\kappa}\right), \tag{3.1}$$

where G is a gradient of Huber loss at β^*, $\|\cdot\|^*_{\mathrm{reg}}$ is a norm dual to the regularization norm (which is equal to $\|\cdot\|_{\max}$ for ℓ_1 regularizer), s is a *structure* parameter, which is equal to $\sqrt{k/\lambda}$, and κ is a restricted strong convexity parameter. Note that by error here we mean $\frac{1}{\sqrt{n}}\left\|X\left(\hat{\boldsymbol{\beta}} - \beta^*\right)\right\|$.

Similarly, under under assumptions of Theorem 2.1, we get the error bound Eq. (3.1), where G is a gradient of Huber loss at L^*, $\|\cdot\|^*_{\mathrm{reg}}$ is a norm dual to the nuclear norm (i.e. the spectral norm), structure parameter s is $\sqrt{r}$, and κ is a restricted strong convexity parameter. For more details on the conditions of the error bound, see the supplementary material. Below, we explore the bounds on the norm of the gradient and on the restricted strong convexity parameter.

Bounding the gradient of the Huber loss. The gradient of the Huber-loss $\boldsymbol{F}_2(\cdot)$ at β^* has the form $\nabla \boldsymbol{F}_2(\beta^*) = \sum_{i=1}^n f_2'\left[\boldsymbol{\eta}_i\right] \cdot X_i$, where X_i is the i-th row of X. The random variables $f_2'\left[\boldsymbol{\eta}_i\right]$, $i \in [n]$, are independent, centered, symmetric and bounded by 2. Since we assume that each column of X has norm at most $\sqrt{n}$, the entries of the row X_i are easily bounded by $\sqrt{n}$. Thus, $\nabla \boldsymbol{F}_2(\beta^*)$ is a vector with independent, symmetric entries with bounded variance, so its behavior can be easily studied through standard concentration bounds. In particular, by a simple application of Hoeffding's inequality, we obtain, with high probability,

$$\left\|\nabla \boldsymbol{F}_2(\beta^*)\right\|_{\max} = \max_{j\in[d]}\left|\sum_{i\in[n]} f_2'\left[\boldsymbol{\eta}_i\right] \cdot X_{ij}\right| \leqslant O\left(\sqrt{n \log d}\right). \tag{3.2}$$

Local strong convexity of the Huber loss. Proving local strong convexity presents additional challenges. Without the sparsity constraint, [dNS21] showed that under a slightly stronger spreadness assumptions than Theorem 2.2, the Huber loss is locally strongly convex within a constant radius

[7] The term "standard noise assumptions" is deliberately vague; as a concrete example, we will refer to (sub)-Gaussian noise distributions. See again [Wai19] for a survey.

[8] We will clarify this notion in the subsequent paragraphs.

R ball centered at β^* whenever $n \gtrsim d/\alpha^2$. (This function is not globally strongly convex due to its linear parts.) Using their result as a black-box, one can obtain the error guarantees of Theorem 2.2, but with suboptimal sample complexity. The issue is that with the substantially smaller sample size of $n \geqslant \tilde{O}(k/\alpha^2)$ that resembles the usual considerations in the context of sparse regression, the Huber loss is *not* locally strongly convex around β^* uniformly across *all directions*, so we cannot hope to prove convergence with optimal sample complexity using this argument. To overcome this obstacle, we make use of the framework of M-estimators: Since we consider a regularized version of the Huber loss, it will be enough to show local strong convexity in a radius R uniformly across all directions which are *approximately* k-sparse. For this substantially weaker condition, $\tilde{O}(k/\alpha^2)$ will be enough.

More in details, for observations $\boldsymbol{y} = X\beta^* + \boldsymbol{\eta}$ and an arbitrary $u \in \mathbb{R}^d$ of norm $\|u\| \leqslant R$, it is possible to lower bound the Hessian[9] of the Huber loss at $\beta^* + u$ by:[10]

$$\begin{aligned} H\boldsymbol{F}_2(\beta^* + u) &= \sum_{i=1}^{n} f_2''\left[(Xu)_i - \boldsymbol{\eta}_i\right] \cdot X_i X_i^\top = \sum_{i=1}^{n} \mathbf{1}_{\left[|(Xu)_i - \boldsymbol{\eta}_i| \leqslant 2\right]} \cdot X_i X_i^\top \\ &\succeq M(u) := \sum_{i=1}^{n} \mathbf{1}_{[|\langle X_i, u\rangle| \leqslant 1]} \cdot \mathbf{1}_{\left[|\boldsymbol{\eta}_i| \leqslant 1\right]} \cdot X_i X_i^\top . \end{aligned}$$

As can be observed, we do not attempt to exploit cancellations between Xu and $\boldsymbol{\eta}$. Let $\boldsymbol{Q} := \left\{ i \in [n] \,\middle|\, |\boldsymbol{\eta}_i| \leqslant 1 \right\}$ be the set of uncorrupted entries of $\boldsymbol{\eta}$. Given that with high probability $\boldsymbol{Q}$ has size $\Omega(\alpha \cdot n)$, the best outcome we can hope for is to provide a lower bound of the form $\langle u, M(u)u\rangle \geqslant \alpha n$ in the direction $\hat{\boldsymbol{\beta}} - \beta^*$. In [dNS21], it was shown that if the span of the measurement matrix X is well spread, then $\langle u, M(u)u\rangle \geqslant \Omega(\alpha \cdot n)$.

If the direction $\hat{\boldsymbol{\beta}} - \beta^*$ was fixed, it would suffice to show the curvature in that single direction through the above reasoning. However, $\hat{\boldsymbol{\beta}}$ depends on the unknown random noise vector $\boldsymbol{\eta}$. Without the regularizer in Eq. (2.3), this dependence indicates that the vector $\hat{\boldsymbol{\beta}}$ may take any possible direction, so one needs to ensure local strong convexity to hold in a constant-radius ball centered at β^*. That is, $\min_{\|u\| \leqslant R} \lambda_{\min}(M(u)) \geqslant \Omega(\alpha \cdot n)$. It can be shown through a covering argument (of the ball) that this bound holds true for $n \geqslant d/\alpha^2$. This is the approach of [dNS21].

The minimizer of the Huber loss follows a sparse direction. The main issue with the above approach is that no information concerning the direction $\hat{\boldsymbol{\beta}} - \beta^*$ is used. In the settings of sparse regression, however, our estimator contains the regularizer $\|\beta\|_1$. The main consequence of the regularizer is that the *direction* $\hat{\boldsymbol{\beta}} - \beta^*$ is approximately flat in the sense $\|\hat{\boldsymbol{\beta}} - \beta^*\|_1 \leqslant O\left(\sqrt{k}\|\hat{\boldsymbol{\beta}} - \beta^*\|\right)$. The reason[11] is that due to the structure of the objective function in Eq. (2.3) and concentration of the gradient Eq. (3.2), the penalty for dense vectors is larger than the absolute value of the inner product $\langle \nabla \boldsymbol{F}(\beta^*), \hat{\boldsymbol{\beta}} - \beta^*\rangle$ (which, as previously argued, concentrates around its (zero) expectation).

This specific structure of the minimizer implies that it suffices to prove local strong convexity *only* in approximately sparse directions. For these set of directions, we carefully construct a sufficiently small covering set so that $n \geqslant \tilde{O}(k/\alpha^2)$ samples suffice to ensure local strong convexity over it.

Remark 3.1 (Comparison with LASSO). it is important to remark that while this approach of only considering approximately sparse directions has also been used in the context of sparse regression under Gaussian noise (e.g. the LASSO estimator), obtaining the desired lower bound is considerably easier in these settings as it directly follows from the restricted eigenvalue property of the design matrix. In our case, we require an additional careful probabilistic analysis which uses a covering argument for the set of approximately sparse vectors. As we see however, it turns out that we do not need any additional assumptions on the design matrix when compared with the LASSO estimator except for the well-spreadness property (recall that some weak version of well-spreadness is indeed necessary in robust settings, see [dNS21]).

[9] The Hessian does not exist everywhere. Nevertheless, the second derivative of the penalty function f_2 exists as an L_1 function in the sense that $f_2'(b) - f_2'(a) = 2\int_a^b \mathbf{1}_{[|t| \leqslant 2]} dt$. This property is enough for our purposes.

[10] A more extensive explanation of the first part of this analysis can be found in [dNS21].

[11] This phenomenon is a consequence of the *decomposability* of the L_1 norm, see the supplementary material.

3.2 Principal component analysis under oblivious noise

A convenient feature of the approach in Section 3.1 for sparse regression, is that it can be easily applied to additional problems. We briefly explain here how to apply it for principal component analysis. We consider the model defined in Theorem 2.1. We use an estimator based on the Huber loss equipped with the nuclear norm as a regularizer to enforce the low-rank structure in our estimator

$$\hat{\boldsymbol{L}} := \underset{L\in\mathbb{R}^{n\times n},\, \|L\|_{\max}\leqslant \rho/n}{\operatorname{argmin}} \left(F_{\zeta+\rho/n}(\boldsymbol{Y}-L) + 100\sqrt{n}\big(\zeta+\rho/n\big)\|L\|_{\mathrm{nuc}} \right). \tag{3.3}$$

In this setting, the gradient $\nabla F_{\zeta+\rho/n}(\boldsymbol{Y}-L^*)$ is a matrix with independent, symmetric entries which are bounded (by $\zeta+\rho/n$) and hence its spectral norm is $O\big((\zeta+\rho/n)\sqrt{n}\big)$ with high probability. Local strong convexity can be obtained in a similar fashion as shown in Section 3.1: due to the choice of the Huber transition point all entries with small noise are in the quadratic part of F. Moreover, the nuclear norm regularizer ensures that the minimizer is an approximately low-rank matrix in the sense that $\|M\|_{\mathrm{nuc}} \leqslant O\big(\sqrt{r}\|M\|_{\mathrm{F}}\big)$. So again, it suffices to provide curvature of the loss function only on these subset of structured directions.

Remark 3.2 (Incoherence vs. spikiness). Recall the discussion on incoherence in Section 2. If for every $\mu \leqslant n/\log^2 n$ matrix L^* doesn't satisfy μ-incoherence conditions, the results in [CLMW11, ZLW+10] cannot be applied. However, our estimator achieves error $\|\hat{\boldsymbol{L}}-L^*\|_{\mathrm{F}}/\|L^*\|_{\mathrm{F}} \to 0$ as $n\to\infty$. Indeed,let $\omega(1) \leqslant f(n) \leqslant o(\log^2 n)$ and assume $\zeta=0$. Let $u\in\mathbb{R}^n$ be an $f(n)$-sparse unit vector whose nonzero entries are equal to $1/\sqrt{f(n)}$. Let $v\in\mathbb{R}^n$ be a vector with all entries equal to $1/\sqrt{n}$. Then, uv^{T} does not satisfy incoherence with any $\mu < n/f(n)$. We have $\left\|uv^{\mathsf{T}}\right\|_{\mathrm{F}}=1$, and the error of our estimator is $O(1/(\alpha\sqrt{f(n)}))$, so it tends to zero for constant (or even some subconstant) α.

Furthermore notice that the dependence of the error of Theorem 2.1 on the maximal entry of L^* is inherent if we do not require incoherence. Indeed, consider $L_1 = b\cdot e_1 e_1^{\mathsf{T}}$ for large enough $b>0$ and $L_2 = e_2 e_2^{\mathsf{T}}$. For constant α, let $|N_{ij}|$ be 1 with probability $\alpha/2$, 0 with probability $\alpha/2$ and b with probability $1-\alpha$. Then given Y we cannot even distinguish between cases $L^*=L_1$ or $L^*=L_2$, and since $\|L_1-L_2\|_{\mathrm{F}} \geqslant b$, the error should also depend on b.

Remark 3.3 (α vs α^2: what if one knows which entries are corrupted?). As was observed in Section 2, the error bound of our estimator is worse than the error for matrix completion by a factor $1/\sqrt{\alpha}$. We observe similar effect in linear regression: if, as in matrix completion, we are given a randomly chosen α fraction of observations $\left\{\big(X_i, y_i = \langle X_i, \beta^*\rangle + \boldsymbol{\eta}_i\big)\right\}_{i=1}^n$ where $\boldsymbol{\eta}\sim N(0,1)^n$, and since for the remaining samples we may not assume any bound on the signal-to-noise ratio, then this problem is essentialy the same as linear regression with αn observations. Thus the optimal prediction error rate is $\Theta(\sqrt{d/(\alpha n)})$. Now, if we have $y = X\beta^* + \boldsymbol{\eta}$, where $\boldsymbol{\eta}\sim N(0,1/\alpha^2)$, then with probability $\Theta(\alpha)$, $|\eta_i|\leqslant 1$, but the optimal prediction error rate in this case is $\Theta(\sqrt{d/(\alpha^2 n)})$. So in both linear regression and robust PCA, *prior knowledge* of the set of corrupted entries makes the problem easier.

3.3 Optimal fraction of inliers for principal component analysis under oblivious noise

In order to prove Theorem 2.4, we will adopt a generative model for the hidden matrix $\boldsymbol{L}^*$: We will generate $\boldsymbol{L}^*$ randomly but assume that the distribution is known to the algorithm. This makes the problem easier. Therefore, any impossibility result for this generative model would imply impossibility for the more restrictive model for which L^* is deterministic but unknown.

We generate a random flat matrix $\boldsymbol{L}^*$ using $n\cdot r$ independent and uniform random bits in such a way that $\boldsymbol{L}^*$ is of rank r and incoherent with high probability. Then, for every constant $0<\xi<1$, we find a distribution for the random noise $\boldsymbol{N}$ in such a way that the fraction of inliers satisfies $\alpha := \mathbb{P}[|\boldsymbol{N}_{ij}|\leqslant\zeta] = \Theta\Big(\xi\sqrt{r/n}\Big)$, and such that the mutual information between $\boldsymbol{L}^*$ and $\boldsymbol{Y}=\boldsymbol{L}^*+\boldsymbol{N}$ can be upper bounded as $I(\boldsymbol{L}^*;\boldsymbol{Y}) \leqslant O(\xi\cdot n\cdot r)$. Roughly speaking the smaller ξ gets, the more independent $\boldsymbol{L}^*$ and $\boldsymbol{Y}$ will be. Now using an inequality that is similar to the standard Fano-inequality but adapted to weak-recovery, we show that if there is a successful (ε,δ)-weak recovery algorithm for $\boldsymbol{L}^*$ and $\boldsymbol{N}$, then $I(\boldsymbol{L}^*;\boldsymbol{Y}) \geqslant \Omega\big((1-\varepsilon^2)^2\cdot(1-\delta)\cdot n\cdot r\big)$. By combining all these observations together, we can deduce that if ξ is small enough, it is impossible to have a successful (ε,δ)-weak recovery algorithm for $\boldsymbol{L}^*$ and $\boldsymbol{N}$.

References

[BBC11] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis, *Theory and applications of robust optimization*, SIAM Review **53** (2011), no. 3, 464–501. 2

[BGN09] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski, *Robust optimization*, Princeton Series in Applied Mathematics, vol. 28, Princeton University Press, 2009. 2

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar, *Consistent robust regression*, NIPS, 2017, pp. 2107–2116. 2, 4, 5

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, *Robust principal component analysis?*, J. ACM **58** (2011), no. 3, 11:1–11:37. 2, 4, 9

[dNS21] Tommaso d'Orsi, Gleb Novikov, and David Steurer, *Consistent regression when oblivious outliers overwhelm*, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, PMLR, 2021, The current full version is in https://arxiv.org/abs/2009.14774v2. 1, 2, 3, 5, 7, 8

[Hub64] Peter J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. **35** (1964), no. 1, 73–101. 2

[MMYSB19] Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai, and MatÃ■as SalibiÃąn-Barrera, *Robust statistics: Theory and methods*, 2 ed., Wiley Series in Probability and Statistics, John Wiley & Sons, 2019. 2

[NTN11] Nasser Nasrabadi, Trac Tran, and Nam Nguyen, *Robust lasso with missing and grossly corrupted observations*, Advances in Neural Information Processing Systems (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, eds.), vol. 24, Curran Associates, Inc., 2011. 3

[NW12] Sahand Negahban and Martin J. Wainwright, *Restricted strong convexity and weighted matrix completion: Optimal bounds with noise*, Journal of Machine Learning Research **13** (2012), 1665–1697. 3, 4

[PWBM16] Amelia Perry, Alexander S. Wein, Afonso S. Bandeira, and Ankur Moitra, *Optimality and sub-optimality of PCA for spiked random matrices and synchronization*, CoRR **abs/1609.05573** (2016). 4

[RWY11] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls*, IEEE Transactions on Information Theory **57** (2011), no. 10, 6976–6994. 5

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain, *Adaptive hard thresholding for near-optimal consistent robust regression*, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA, 2019, pp. 2892–2897. 2, 3, 4, 5, 6

[SZF19] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan, *Adaptive huber regression*, Journal of the American Statistical Association (2019), 1–24.

[TJSO14] Efthymios Tsakonas, Joakim Jaldén, Nicholas D Sidiropoulos, and Björn Ottersten, *Convergence of the huber regression m-estimate in the presence of dense outliers*, IEEE Signal Processing Letters **21** (2014), no. 10, 1211–1214. 2, 3

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.

[Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019. 5, 6, 7

[ZLW+10] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel J. Candès, and Yi Ma, *Stable principal component pursuit*, ISIT, IEEE, 2010, pp. 1518–1522. 2, 4, 9

[ZWJ14] Yuchen Zhang, M. Wainwright, and Michael I. Jordan, *Lower bounds on the performance of polynomial-time algorithms for sparse linear regression*, COLT, 2014. 5

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 2.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work does not present any foreseeable societal consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary Material .
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]