Outlier-Robust Sparse Estimation via Non-Convex Optimization

Yu Cheng
Brown University
Providence, RI 02912
yu_cheng@brown.edu

Ilias Diakonikolas
University of Wisconsin-Madison
Madison, WI 53706
ilias@cs.wisc.edu

Rong Ge
Duke University
Durham, NC 27708
rongge@cs.duke.edu

Shivam Gupta
University of Texas at Austin
Austin, TX 78712
shivamgupta@utexas.edu

Daniel M. Kane
University of California, San Diego
La Jolla, CA 92093
dakane@cs.ucsd.edu

Mahdi Soltanolkotabi
University of Southern California
Los Angeles, CA 90089
soltanol@usc.edu

Abstract

We explore the connection between outlier-robust high-dimensional statistics and non-convex optimization in the presence of sparsity constraints, with a focus on the fundamental tasks of robust sparse mean estimation and robust sparse PCA. We develop novel and simple optimization formulations for these problems such that *any* approximate stationary point of the associated optimization problem yields a near-optimal solution for the underlying robust estimation task. As a corollary, we obtain that any first-order method that efficiently converges to stationarity yields an efficient algorithm for these tasks.[1] The obtained algorithms are simple, practical, and succeed under broader distributional assumptions compared to prior work.

1 Introduction

In several modern machine learning (ML) applications, such as ML security [BNJT10, BNL12, SKL17, DKK+19a] and exploratory analysis of real datasets, e.g., in population genetics [RPW+02, PLJD10, LAT+08, DKK+17], typical datasets contain a non-trivial fraction of arbitrary (or even adversarial) outliers. Robust statistics [HRRS86, HR09] is the subfield of statistics aiming to design estimators that are tolerant to a *constant fraction* of outliers, independent of the dimensionality of the data. Early work in this field, see, e.g., [Tuk60, Hub64, Tuk75] developed sample-efficient robust estimators for various basic tasks, alas with runtime exponential in the dimension.

During the past five years, a line of work in computer science, starting with [DKK+16, LRV16], has developed the first *computationally efficient* robust high-dimensional estimators for a range of tasks. This progress has led to a revival of robust statistics from an algorithmic perspective (see, e.g., [DK19, DKK+21] for surveys on the topic). In this work, we focus on high-dimensional estimation tasks in the presence of sparsity constraints. To rigorously study these problems, we need to formally define the model of data corruption. Throughout this work, we work with the following standard contamination model.

[1] An implementation of our algorithms is available at https://github.com/guptashvm/Sparse-GD.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Definition 1.1 (Strong Contamination Model, see [DKK+16]). *Given a parameter $0 < \epsilon < 1/2$ and a distribution family $\mathcal{D}$ on $\mathbb{R}^d$, the* adversary *operates as follows: The algorithm specifies a number of samples n, and n samples are drawn from some unknown $D \in \mathcal{D}$. The adversary is allowed to inspect the samples, remove up to ϵn of them and replace them with arbitrary points. This modified set of n points is then given as input to the algorithm. We say that a set of samples is* ϵ-corrupted *if it is generated by the above process.*

High-dimensional robust statistics is algorithmically challenging because the natural optimization formulations of such tasks are typically non-convex. The recent line of work on algorithmic robust statistics has led to a range of sophisticated algorithms. In some cases, such algorithms require solving large convex relaxations, rendering them computationally prohibitive for large-scale problems. In other cases, they involve a number of hyper-parameters that may require careful tuning. Motivated by these shortcomings of known algorithms, recent work [CDGS20, ZJS20] established an intriguing connection between high-dimensional robust estimation and non-convex optimization. The high-level idea is quite simple: Even though typical robust statistics tasks lead to non-convex formulations, it may still be possible to leverage the underlying structure to show that standard first-order methods provably and efficiently reach near-optimal solutions. Indeed, [CDGS20, ZJS20] were able to prove such statements for robust mean estimation under natural distributional assumptions. Specifically, these works established that any (approximate) stationary point of a well-studied non-convex formulation for robust mean estimation yields a near-optimal solution for the underlying robust estimation task.

In this work, we continue this line of work with a focus on *sparse* estimation tasks. Leveraging sparsity in high-dimensional datasets is a fundamental problem of significant practical importance. Various formalizations of this problem have been investigated in statistics and machine learning for at least the past two decades (see, e.g., [HTW15] for a textbook on the topic). We focus on *robust sparse mean estimation* and *robust sparse PCA*. Sparse mean estimation is arguably one of the most fundamental sparse estimation tasks and is closely related to the Gaussian sequence model [Tsy08, Joh17]. The task of sparse PCA in the spiked covariance model, initiated in [Joh01], has been extensively investigated (see Chapter 8 of [HTW15] and references therein).

In the context of robust sparse mean estimation, we are given an ϵ-corrupted set of samples from a distribution with unknown mean $\mu \in \mathbb{R}^d$ where μ is k-sparse, and we want to compute a vector $\widehat{\mu}$ close to μ. In the context of robust sparse PCA (in the spiked covariance model), we are given an ϵ-corrupted set of samples from a distribution with covariance matrix $I + \rho vv^T$, where $v \in \mathbb{R}^d$ is k-sparse and the goal is to approximate v. It is worth noting that for both problems, we have access to much fewer samples compared to the non-sparse case (roughly $O(k^2 \log d)$ instead of $\Omega(d)$). Consequently, the design and analysis of optimization formulations for robust sparse estimation requires new ideas and techniques that significantly deviate from the standard (non-sparse) case.

1.1 Our Results and Contributions

We show that standard first-order methods lead to robust and efficient algorithms for sparse mean estimation and sparse PCA. Our main contribution is to propose novel (non-convex) formulations for these robust estimation tasks, and to show that *approximate stationarity suffices for near-optimality*. We establish landscape results showing that *any* approximate stationary point of our objective function yields a near-optimal solution for the underlying robust estimation task. Consequently, gradient descent (or any other methods converging to stationarity) can solve these problems.

Our results provide new insights and techniques in designing and analyzing (non-convex) optimization formulations of robust estimation tasks. Our formulations and structural results immediately lead to simple and practical algorithms for robust sparse estimation. Importantly, the gradient of our objectives can be computed efficiently via a small number of basic matrix operations. In addition to their simplicity and practicality, our methods provably succeed under more general distributional assumptions compared to prior work.

For robust sparse mean estimation and robust sparse PCA, our landscape results require deterministic conditions on the original set of good samples. We refer to these conditions as *stability conditions* (Definitions 2.1 and 2.2, formally defined in Section 2). At a high level, they state that the first and second moments of a set of samples are stable when *any* ϵ-fraction of the samples are removed. These stability conditions hold with high probability for a set of clean samples drawn from natural families of distributions (e.g., subgaussian).

For robust sparse mean estimation, we establish the following result.

Theorem 1.2 (Robust Sparse Mean Estimation)**.** *Let $0 < \epsilon < \epsilon_0$ for some universal constant ϵ_0 and let $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (per Definition 2.1) w.r.t. a distribution with unknown k-sparse mean $\mu \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$.* [2] *There is an algorithm that on inputs S, k, ϵ, and δ, runs in polynomial time and returns a k-sparse vector $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2 \leq O(\delta)$.*

We emphasize that a key novelty of Theorem 1.2 is that the underlying algorithm is a *first-order method* applied to a *novel non-convex formulation* of the problem. The major advantage of our algorithm over prior work [BDLS17, DKK+19b] is its simplicity, practicality, and the fact that it seamlessly applies to a wider class of distributions on the clean data.

As we will discuss in Section 3, when the ground-truth distribution D is subgaussian with unknown k-sparse mean $\mu \in \mathbb{R}^d$ and identity covariance, a set of $n = \widetilde{\Omega}(k^2 \log d/\epsilon^2)$ samples drawn from D is (k, ϵ, δ)-stable (Definition 2.1) with high probability for $\delta = O(\epsilon\sqrt{\log(1/\epsilon)})$. It follows as an immediate corollary of Theorem 1.2 that, given an ϵ-corrupted set of samples, we can compute a vector $\widehat{\mu}$ that is $O(\delta) = O(\epsilon\sqrt{\log(1/\epsilon)})$ close to the true mean μ. This sample complexity matches the known computational-statistical lower bounds [DKS17, BB20]. More generally, one can relax the concentration assumption on the clean data and obtain qualitatively similar error guarantees.

Next we state our main result for robust sparse PCA.

Theorem 1.3 (Robust Sparse PCA)**.** *Let $0 < \rho \leq 1$ and $0 < \epsilon < \epsilon_0$ for some universal constant ϵ_0. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (as in Definition 2.2) w.r.t. a centered distribution with covariance $\Sigma = I + \rho vv^\top$, for an unknown k-sparse unit vector $v \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. There is an algorithm that on inputs S, k, and ϵ, runs in polynomial time and returns a unit vector $u \in \mathbb{R}^d$ such that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\delta/\rho})$.*

Interestingly, our algorithm for robust sparse PCA is a first-order method applied to a simple *convex* formulation of the problem. We view the existence of a convex formulation as an intriguing fact that, surprisingly, was not observed in prior work.

As we will discuss in Section 4, when the ground-truth distribution D is centered subgaussian with covariance $\Sigma = I + \rho vv^\top$, for an unknown k-sparse unit vector $v \in \mathbb{R}^d$, a set of $n = \widetilde{\Omega}(k^2 \log d/\epsilon^2)$ samples drawn from D is (k, ϵ, δ)-stable (Definition 2.2) with high probability for $\delta = O(\epsilon \log(1/\epsilon))$. Therefore, our algorithm outputs a vector that is $O(\sqrt{\epsilon \log(1/\epsilon)/\rho})$ close to the true direction v. The sample complexity in this case nearly matches the computational-statistical lower bound of $\Omega(k^2 \log d/\epsilon^2)$ [BR13] which holds even without corruptions. While the error guarantee of our algorithm is slightly worse compared to prior work [BDLS17, DKK+19b] for Gaussian data (we get $O(\sqrt{\delta/\rho})$ rather than $O(\delta/\rho)$), we note that our algorithm works for a broader family of distributions.

Prior Work on Robust Sparse Estimation. We provide a detailed summary of prior work for comparison. [BDLS17] obtained the first sample-efficient and polynomial-time algorithms for robust sparse mean estimation and robust sparse PCA. These algorithms succeed for Gaussian inliers and inherently use the ellipsoid method. The separation oracle required for the ellipsoid algorithm turns out to be another convex program — corresponding to an SDP to solve sparse PCA. As a consequence, the running time of these algorithms, while polynomially bounded, is impractically high. [LLC19] proposed an algorithm for robust sparse mean estimation via iterative trimmed hard thresholding, which can only tolerate a *sub-constant* fraction of corruptions. [DKK+19b] gave iterative spectral robust algorithms for sparse mean estimation and sparse PCA. These algorithms are still quite complex and are only shown to succeed under Gaussian inliers.

1.2 Overview of Our Approach

In this section, we give an overview of our approach for robust sparse mean estimation. At a very high level, we assign a nonnegative weight to each data point and try to find a good set of $(1-\epsilon)n$ samples. The constraint on the weight vector is that it represents at least a (fractional) set of $(1-\epsilon)$-portion of the input dataset. Formally, given n datapoints $(X_i)_{i=1}^n$, the goal is to find a weight vector $w \in \mathbb{R}^n$

[2] For two sets of samples S and T, we say S is an ϵ-corrupted version of T if $|S| = |T|$ and $|S \setminus T| \leq \epsilon|S|$.

such that $\mu_w = \sum_i w_i X_i$ is close to the true mean μ. The constraint on w is that it belongs to

$$\Delta_{n,\epsilon} = \left\{ w \in \mathbb{R}^n : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \tfrac{1}{(1-\epsilon)n} \ \forall i \right\},$$

which is the convex hull of all uniform distributions over subsets $S \subseteq [n]$ of size $|S| = (1-\epsilon)n$.

Let $\Sigma_w = \sum_i w_i (X_i - \mu_w)(X_i - \mu_w)^\top$ denote the weighted empirical covariance matrix. It is well-known that if one can find $w \in \Delta_{n,\epsilon}$ that minimizes the weighted empirical variance $v^\top \Sigma_w v$ for all k-sparse unit vectors v, then μ_w must be close to μ. Unfortunately, it is NP-Hard to find the sparse direction v with the largest variance. To get around this issue, [BDLS17] considered the following convex relaxation, minimizing the variance for convex combinations of sparse directions:

$$\min_w \max_{\mathrm{tr}(A)=1, \sum_{ij}|A_{ij}| \le k, A \succeq 0} (A \bullet \Sigma_w) \ . \tag{1}$$

Given w, the optimal A can be found using semidefinite programming (SDP). [ZJS20] observed that any stationary point w of (1) gives a good solution for robust sparse mean estimation. However, solving (1) requires convex programming to compute the gradient in each iteration. As explained in the proceeding discussion, our approach circumvents this shortcoming, leading to a formulation for which each gradient can be computed *using only basic matrix operations*.

In this work, we propose and analyze the following optimization formulation:

$$\min_w \ f(w) = \|\Sigma_w - I\|_{F,k,k} \quad \text{subject to} \ \ w \in \Delta_{n,\epsilon} \ ,$$

where $\|A\|_{F,k,k}$ is the Frobenius norm of the k^2 entries of A with largest magnitude, with the additional constraint that these k^2 entries are chosen from k rows with k entries in each row.

We prove that any stationary point of $f(w)$ yields a good solution for robust sparse mean estimation. Here we provide a brief overview of our proof (see Section 3 for more details). Given a weight vector w, we show that if w is not a good solution, then moving toward $w^\star$ (the weight vector corresponding to the uniform distribution on the clean input samples) will decrease the objective value. Formally, we will show that, for any $0 < \eta < 1$,

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta \Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top \ .$$

We can then take $\|\cdot\|_{F,k,k}$ norm on both sides (after subtracting I) and show that the third term can be essentially ignored. If the third term were not there, we would have

$$\begin{aligned} f((1-\eta)w + \eta w^\star) &= \left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_{F,k,k} \\ &\le (1-\eta)\|\Sigma_w - I\|_{F,k,k} + \eta \|\Sigma_{w^\star} - I\|_{F,k,k} = (1-\eta) f(w) + \eta f(w^\star) \ . \end{aligned}$$

Therefore, if w is a bad solution with $f(w)$ much larger than $f(w^\star)$, then w cannot be a stationary point because f decreases when we move from w to $(1-\eta)w + \eta w^\star$.

Remark 1.4. The technical overview for robust sparse PCA follows a similar high-level approach, but is somewhat more technical. It is deferred to Section 4.

Roadmap. In Section 2, we introduce basic notations and the deterministic stability conditions that we require on the good samples. We present our algorithms and analysis for robust sparse mean estimation in Section 3 and robust sparse PCA in Section 4. In Section 5, we evaluate our algorithm on synthetic datasets and show that it achieves good statistical accuracy under various noise models.

2 Preliminaries and Background

Notation. For a positive integer n, let $[n] = \{1, \ldots, n\}$. For a vector v, we use $\|v\|_0$, $\|v\|_1$, $\|v\|_2$, and $\|v\|_\infty$ for the number of non-zeros, the ℓ_1, ℓ_2, and ℓ_∞ norm of v respectively. Let I be the identity matrix. For a matrix A, we use $\|A\|_2$, $\|A\|_F$, $\mathrm{tr}(A)$ for the spectral norm, Frobenius norm, and trace of A respectively. For two vectors x, y, let $x^\top y$ denote their inner product. For two matrices A, B, we use $A \bullet B = \mathrm{tr}(A^\top B)$ for their entrywise inner product. A matrix A is said to be positive semidefinite (PSD) if $x^\top A x \ge 0$ for all x. We write $A \preceq B$ iff $(B - A)$ is PSD.

For a vector $w \in \mathbb{R}^n$, let $\mathrm{diag}(w) \in \mathbb{R}^{n\times n}$ denote a diagonal matrix with w on the diagonal. For a matrix $A \in \mathbb{R}^{n\times n}$, let $\mathrm{diag}(A) \in \mathbb{R}^n$ denote a column vector with the diagonal of A. For a vector $v \in \mathbb{R}^d$ and a set $S \subseteq [d]$, we write $v_S \in \mathbb{R}^d$ for a vector that is equal to v on S and zero everywhere else. Similarly, for a matrix $A \in \mathbb{R}^{d\times d}$ and a set $S \subseteq ([d] \times [d])$, we write A_S for a matrix that is equal to A on S and zero everywhere else.

For a vector v, we define $\|v\|_{2,k} = \max_{|S|=k} \|v_S\|_2$ to be the maximum ℓ_2-norm of any k entries of v. For a matrix A, we define $\|A\|_{F,k^2}$ to be the maximum Frobenius norm of any k^2 entries of A. Moreover, we define $\|A\|_{F,k,k}$ to be the maximum Frobenius norm of any k^2 entries with the extra requirement that these entries must be chosen from k rows with k entries in each row. Formally,

$$\|A\|_{F,k^2} = \max_{|Q|=k^2} \|A_Q\|_F \quad \text{and} \quad \|A\|^2_{F,k,k} = \max_{|S|=k} \sum_{i\in S} \|A_i\|^2_{2,k} \text{ where } A_i \text{ is } i\text{-th row of } A\ . \tag{2}$$

Sample Reweighting Framework. We use n for the number of samples, d for the dimension, and ϵ for the fraction of corrupted samples. For sparse estimation, we use k for the sparsity of the ground-truth parameters. We use $G^\star$ for the original set of n good samples. We use $S = G \cup B$ for the input samples after the adversary replaced ϵ-fraction of $G^\star$, where $G \subset G^\star$ is the set of remaining good samples and B is the set of bad samples (outliers) added by the adversary. Note that $|G| = (1-\epsilon)n$ and $|B| = \epsilon n$.

Given n samples $X_1, \ldots, X_n$, we write $X \in \mathbb{R}^{d\times n}$ as the sample matrix where the i-th column is X_i. For a weight vector $w \in \mathbb{R}^n$, we use $\mu_w = Xw = \sum_i w_i X_i$ for the weighted empirical mean and $\Sigma_w = X\,\mathrm{diag}(w)X - \mu_w\mu_w^\top = \sum_i w_i (X_i - \mu_w)(X_i - \mu_w)^\top$ for the weighted empirical covariance. Let $\Delta_{n,\epsilon}$ be the convex hull of all uniform distributions over subsets $S \subseteq [n]$ of size $|S| = (1-\epsilon)n$: $\Delta_{n,\epsilon} = \{w \in \mathbb{R}^n : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)n}\ \forall i\}$, In other words, every $w \in \Delta_{n,\epsilon}$ corresponds to a fractional set of $(1-\epsilon)n$ samples. We use $w^\star$ to denote the uniform distribution on G (the remaining good samples in S).

Deterministic Stability Conditions. For robust sparse mean estimation and robust sparse PCA, we require the following conditions respectively.

Definition 2.1 (Stability Conditions for Sparse Mean)**.** *A set of n samples $G^\star = (X_i)_{i=1}^n$ is said to be (k,ϵ,δ)-stable (w.r.t. a distribution with mean μ) iff for any weight vector $w \in \Delta_{n,2\epsilon}$, we have $\|\mu_w - \mu\|_{2,k} \le \delta$ and $\|\Sigma_w - I\|_{F,k,k} \le \delta^2/\epsilon$, where μ_w and Σ_w are the weighted empirical mean and covariance matrix respectively, and the $\|\cdot\|_{F,k,k}$ norm is defined in Equation* (2).

Definition 2.2 (Stability Conditions for Sparse PCA)**.** *A set of n samples $G^\star = (X_i)_{i=1}^n$ is (k,ϵ,δ)-stable (w.r.t. a centered distribution with covariance $I + \rho vv^\top$) iff for any weight vector $w \in \Delta_{n,2\epsilon}$, $\left\|M_w - (I + \rho vv^\top)\right\|_{F,2k^2} \le \delta$, where $M_w = \sum_i w_i X_i X_i^\top$ and the $\|\cdot\|_{F,2k^2}$ norm is defined in Equation* (2)*.*

First-Order Stationary Points. We give a formal definition of the notion of (approximate) first-order stationary point that we use in this paper.

Definition 2.3 (Approximate Stationary Points)**.** *Fix a convex set $\mathcal{K}$ and a differentiable function f. For $\gamma \ge 0$, we say that $x \in \mathcal{K}$ is a γ-stationary point of f iff the following condition holds: For any unit vector u where $x + \alpha u \in \mathcal{K}$ for some $\alpha > 0$, we have $u^\top \nabla f(x) \ge -\gamma$.*

We note that the objective functions studied in this paper are not everywhere differentiable. This is because, taking the $\|\cdot\|_{F,k,k}$ norm as an example, there can be ties in choosing the largest k^2 entries. When the function f is not differentiable, we use ∇f informally to denote an element of the sub-differential. We will show in Appendix C that, while f is not differentiable, it does have a nonempty subdifferential, as it can be written as the pointwise maximum of differentiable functions.

3 Robust Sparse Mean Estimation

In this section, we present our non-convex approach for robust sparse mean estimation. We will optimize the following objective, where $\|\cdot\|_{F,k,k}$ is defined in Equation (2):

$$\min_w\ f(w) = \|\Sigma_w - I\|_{F,k,k} \quad \text{subject to}\ \ w \in \Delta_{n,\epsilon}\ . \tag{3}$$

We will show that the objective function (3) has no bad stationary points (Theorem 3.1). In other words, *every* first-order stationary point of f yields a good solution for robust sparse mean estimation.

Our algorithm is stated in Algorithm 1. As a consequence of our landscape result (Theorem 3.1), we know that Algorithm 1 works *no matter how* we find a stationary point of f (because any stationary point works), so we intentionally did not specify how to find such a point. As a simple illustration, we show that (projected) gradient descent can be used to minimize f. The convergence analysis and iteration complexity are provided in Appendix C.

Algorithm 1: Robust sparse mean estimation.

Input: $k > 0$, $0 < \epsilon < \epsilon_0$, and an ϵ-corrupted set of samples $(X_i)_{i=1}^n$ drawn from a distribution with k-sparse mean μ. [3]

Output: a vector $\widehat{\mu}$ that is close to μ.

1: Find a first-order stationary point $w \in \Delta_{n,\epsilon}$ of the objective $\min_w f(w) = \|\Sigma_w - I\|_{F,k,k}$.

2: Return $\widehat{\mu} = (\mu_w)_Q$ where Q is a set of k entries of μ_w with largest magnitude.

Formally, we first prove that Algorithm 1 can output a vector $\widehat{\mu} \in \mathbb{R}^d$ that is close to μ in $\|\cdot\|_{2,k}$ norm, as long as the good samples satisfies the stability condition in Definition 2.1.

Theorem 3.1. *Fix $k > 0$, $0 < \epsilon < \epsilon_0$, and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k, ϵ, δ)-stable (as in Definition 2.1) w.r.t. a distribution with unknown k-sparse mean $\mu \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. Let $f(w) = \|\Sigma_w - I\|_{F,k,k}$. Let $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. Then, for any $w \in \Delta_{n,\epsilon}$ that is a γ-stationary point of $f(w)$, we have $\|\mu_w - \mu\|_{2,k} = O(\delta)$.*

Once we have a vector μ_w that is $O(\delta)$-close to μ in $\|\cdot\|_{2,k}$ norm, we can guarantee that a truncated version of μ_w (the output $\widehat{\mu}$ of Algorithm 1) is $O(\delta)$-close to μ in the ℓ_2-norm:

Lemma 3.2. *Fix two vectors x, y with $\|x\|_0 \le k$ and $\|x - y\|_{2,k} \le \delta$. Let z be a vector that keeps the k entries of y with largest absolute values and sets the rest to 0. We have $\|x - z\|_2 \le \sqrt{5}\delta$.*

Theorem 1.2 follows immediately from Theorem 3.1 and Lemma 3.2.

We can apply Theorem 1.2 to get an end-to-end result for subgaussian distributions. We show that the required stability conditions are satisfied with a small number of samples.

Lemma 3.3. *Fix $k > 0$ and $0 < \epsilon < \epsilon_0$. Let $G^\star$ be a set of n samples that are drawn i.i.d. from a subgaussian distribution with mean μ and covariance I. If $n = \Omega(k^2 \log d/\epsilon^2)$, then with probability at least $1 - \exp(-\Omega(k^2 \log d))$, $G^\star$ is (k, ϵ, δ)-stable (as in Definition 2.1) for $\delta = O(\epsilon \log(1/\epsilon))$.*

Combining Theorem 1.2 and Lemma 3.3, we know that given an ϵ-corrupted set of $O(k^2 \log d/\epsilon^2)$ samples drawn from a subgaussian distribution with k-sparse mean μ, the output of Algorithm 1 is $O(\epsilon\sqrt{\log(1/\epsilon)})$-close to μ in ℓ_2-norm.

In the rest of this section, we will prove Theorem 3.1. Omitted proofs in this section are in Appendix A.

We start with some intuition on why we choose our objective function (3). We would like to design $f(w) = g(\Sigma_w - I)$ to satisfy the following properties:

1. $g(\Sigma_w - I)$ is an upper bound on $v^\top(\Sigma_w - I)v$ for all k-sparse unit vectors $v \in \mathbb{R}^d$. This way, a small objective value implies that $\|\mu_w - \mu\|_{2,k}$ is small.
2. $g(\Sigma_{w^\star} - I)$ is small for $w^\star$ (the uniform distribution on G). This guarantees that a good w exists.
3. Triangle inequality on g. This allows us to upper bound the objective value when we move w toward $w^\star$ by the sum of $g(\cdot)$ of each term on the right-hand side:
$$\Sigma_{(1-\eta)w+\eta w^\star} - I = (1-\eta)(\Sigma_w - I) + \eta(\Sigma_{w^\star} - I) + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top .$$
4. $g(uu^\top)$ is close to $g(vv^\top)$ where v keeps only the k largest entries of u. We want to approximate μ in $\|\cdot\|_{2,k}$ norm, so intuitively $g(\Sigma_w - I)$ should depend only on the largest k entries of $(\mu_w - \mu)$.

[3] Without loss of generality we can assume that ϵ is given to the algorithm. This is because we can run a binary search to determine ϵ: if our guess of ϵ is too small, then the algorithm will output a w whose objective value $f(w)$ is much larger than it should be.

Our choice of $f(w) = g(\Sigma_w - I) = \|\Sigma - I\|_{F,k,k}$ is motivated by (and satisfies) all these properties.

Lemma 3.4. *Fix $A \in \mathbb{R}^{d\times d}$. We have $\left|v^\top A v\right| \le \|A\|_{F,k,k}$ for any k-sparse unit vector $v \in \mathbb{R}^d$.*

Lemma 3.5. *For any vector $v \in \mathbb{R}^d$, $\left\|vv^\top\right\|_{F,k,k} = \|v\|_{2,k}^2$.*

We now continue to present key technical lemmas for proving our main structural result (Theorem 3.1). Lemma 3.6 gives the weighted empirical covariance for a convex combination of two weight vectors.

Lemma 3.6. *Fix n samples $X_1, \dots, X_n \in \mathbb{R}^d$. Let $\overline{w}, \widehat{w} \in \mathbb{R}^n$ be two non-negative weight vectors with $\|\overline{w}\|_1 = \|\widehat{w}\|_1 = 1$. For any $\alpha, \beta \ge 0$ with $\alpha + \beta = 1$, letting $w = \alpha\overline{w} + \beta\widehat{w}$, we have*

$$\Sigma_w = \alpha\Sigma_{\overline{w}} + \beta\Sigma_{\widehat{w}} + \alpha\beta(\mu_{\overline{w}} - \mu_{\widehat{w}})(\mu_{\overline{w}} - \mu_{\widehat{w}})^\top .$$

Proof. Because $w = \alpha\overline{w} + \beta\widehat{w}$ and μ_w is linear in w, we have $\mu_w = \alpha\mu_{\overline{w}} + \beta\mu_{\widehat{w}}$. The lemma follows from the following calculations:

$$\begin{aligned}\Sigma_w = \sum_i w_i X_i X_i^\top - \mu_w\mu_w^\top &= \sum_i \alpha\overline{w}_i X_i X_i^\top - \alpha\mu_{\overline{w}}\mu_{\overline{w}}^\top + \sum_i \beta\widehat{w}_i X_i X_i^\top - \beta\mu_{\widehat{w}}\mu_{\widehat{w}}^\top \\ &\quad + \alpha\mu_{\overline{w}}\mu_{\overline{w}}^\top + \beta\mu_{\widehat{w}}\mu_{\widehat{w}}^\top - (\alpha\mu_{\overline{w}} + \beta\mu_{\widehat{w}})(\alpha\mu_{\overline{w}} + \beta\mu_{\widehat{w}})^\top \\ &= \alpha\Sigma_{\overline{w}} + \beta\Sigma_{\widehat{w}} + \alpha\beta(\mu_{\overline{w}} - \mu_{\widehat{w}})(\mu_{\overline{w}} - \mu_{\widehat{w}})^\top .\end{aligned}$$

The last step uses $\alpha - \alpha^2 = \beta - \beta^2 = \alpha\beta$ as $\alpha + \beta = 1$. □

Let $w^\star$ denote the uniform distribution on G, i.e., $w_i^\star = \frac{1}{(1-\epsilon)n}$ if $i \in G$ and $w_i^\star = 0$ otherwise. By Lemma 3.6 for any w, if we move toward $w^\star$, we have

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top .$$

We will show that we can essentially ignore the last rank-one term using Lemma 3.7.

Lemma 3.7. *Let $G^\star$ be a (k,ϵ,δ)-stable set of samples with respect to the ground-truth distribution with $0 < \epsilon \le \delta$. Let S be an ϵ-corrupted version of $G^\star$. Then, we have*

$$\left\|(\mu_w - \mu_{w_\star})(\mu_w - \mu_{w_\star})^\top\right\|_{F,k,k} \le 4\epsilon\left(\|\Sigma_w - I\|_{F,k,k} + O(\delta^2/\epsilon)\right) .$$

We are now ready to prove our main result (Theorem 3.1).

Proof of Theorem 3.1. Fix any weight vector $w \in \Delta_{n,\epsilon}$. We will show that if w is a bad solution, then $f(w)$ decreases if w moves toward $w^\star$, so w cannot be a stationary point.

Let c_1 be the constant in $O(\cdot)$ in Lemma 3.7. By Lemma 3.7, if $\|\mu_w - \mu\|_{2,k} \ge c_2\delta$ for a sufficiently large constant c_2, then $\|\Sigma_w - I\|_{F,k,k} \ge (\frac{c_2^2}{4} - c_1)\frac{\delta^2}{\epsilon} = \Omega(\frac{\delta^2}{\epsilon})$.

By Lemma 3.6, $\Sigma_{(1-\eta)w+\eta w^\star} - I = (1-\eta)(\Sigma_w - I) + \eta(\Sigma_{w^\star} - I) + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$.

Using the triangle inequality for $\|\cdot\|_{F,k,k}$, we have

$$\begin{aligned}\left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_{F,k,k} &\le (1-\eta)\|\Sigma_w - I\|_{F,k,k} \\ &\quad + \eta\|\Sigma_{w^\star} - I\|_{F,k,k} + \eta(1-\eta)\left\|(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top\right\|_{F,k,k} .\end{aligned}$$

We know that $\|\Sigma_{w^\star} - I\|_{F,k,k} \le \frac{\delta^2}{\epsilon}$ by the stability condition in Definition 2.1. By Lemma 3.7 and $\|\Sigma_w - I\|_{F,k,k} = \Omega(\delta^2/\epsilon)$, we can show that for all $0 < \eta < 1$,

$$\begin{aligned}f((1-\eta)w + \eta w^\star) &= \left\|\Sigma_{(1-\eta)w+\eta w^\star} - I\right\|_{F,k,k} \\ &\le (1-\eta)\|\Sigma_w - I\|_{F,k,k} + \tfrac{\eta\delta^2}{\epsilon} + 4\epsilon\eta\left(\|\Sigma_w - I\|_{F,k,k} + O(\tfrac{\delta^2}{\epsilon})\right) \\ &\le (1 - \eta + 4\epsilon\eta)\|\Sigma_w - I\|_{F,k,k} + (4c_1 + 1)\tfrac{\eta\delta^2}{\epsilon} \\ &\le (1 - \tfrac{\eta}{2})\|\Sigma_w - I\|_{F,k,k} = (1 - \tfrac{\eta}{2})f(w) .\end{aligned} \tag{4}$$

The last step uses $(\frac{1}{2}-4\epsilon)\left\|\Sigma_w - I\right\|_2 \geq (4c_1+1)\frac{\delta^2}{\epsilon}$ which holds if $\epsilon \leq 1/10$ and $c_2^2 \geq 164\,c_1 + 40$.

It follows immediately that w cannot be a stationary point. Let $u = \frac{w^\star - w}{\|w^\star - w\|_2}$ and $h = \eta\left\|w^\star - w\right\|_2$. We have $w + hu = (1-\eta)w + \eta w^\star \in \Delta_{n,\epsilon}$ because $\Delta_{n,\epsilon}$ is convex. Since $\left\|w^\star - w\right\|_2 = O(\sqrt{\epsilon/n})$,

$$u^\top \nabla f(w) = \lim_{h\to 0} \tfrac{f(w+hu)-f(w)}{h} \leq \lim_{\eta\to 0} \tfrac{-(\eta/2)f(w)}{\eta\|w^\star - w\|_2} \leq -\tfrac{\Omega(\delta^2/\epsilon)}{\|w^\star - w\|_2} \leq -\Omega(n^{1/2}\delta^2\epsilon^{-3/2})\ .$$

By Definition 2.3, we know w cannot be a γ-stationary point of f for some $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. $\square$

4 Robust Sparse PCA

We consider a spiked covariance model for sparse PCA. In this model, there is a direction $v \in \mathbb{R}^d$ with at most k nonzero entries. The good samples are drawn from a ground-truth distribution with covariance $\Sigma = I + \rho vv^\top$, where $\rho > 0$ is a parameter that intuitively measures the strength of the signal. We consider the more interesting case when $\rho \leq 1$ (if ρ is larger the problem becomes easier).

To solve the sparse PCA problem, we consider the following optimization problem, where $M_w = \sum_i w_i X_i X_i^\top$ and $\|A\|_{F,2k^2} = \max_{|Q|=2k^2} \|A_Q\|_F$:

$$\min_w\ f(w) = \|M_w - I\|_{F,2k^2}\ \text{ subject to }\ w \in \Delta_{n,\epsilon}. \tag{5}$$

The objective function minimizes the Frobenius norm of the largest $2k^2$ entries of a reweighted second-moment matrix M_w. Note that $f(w)$ is actually convex in w, because the matrix M_w is linear in w and the $\|\cdot\|_{F,2k^2}$ norm is convex.

Let R be the support of $vv^\top$. Intuitively, the k^2 entries in R could be large due to spiked covariance. By minimizing the norm of the largest $2k^2$ entries, we hope to make the entries outside of R very small. Our algorithm is given in Algorithm 2.

Algorithm 2: Robust sparse PCA.

Input: $k > 0$, $0 < \epsilon < \epsilon_0$, and an ϵ-corrupted set of samples $(X_i)_{i=1}^n$ drawn from a distribution with covariance $I + \rho vv^\top$ for a k-sparse unit vector v.
Output: a vector u that is close to v.

1: Find a first-order stationary point $w \in \Delta_{n,\epsilon}$ of the objective $\min_w f(w) = \|M_w - I\|_{F,2k^2}$.
2: Let $A = M_w - I$. Let Q be the k^2 entries of A with largest magnitude.
3: Return $u =$ the top eigenvector of $(A_Q + A_Q^\top)$.

Theorem 4.1. *Let $0 < \rho \leq 1$, $0 < \epsilon < \epsilon_0$, and $\delta > \epsilon$. Let $G^\star$ be a set of n samples that is (k,ϵ,δ)-stable (as in Definition 2.2) w.r.t. a centered distribution with covariance $I + \rho vv^\top$ for an unknown k-sparse unit vector $v \in \mathbb{R}^d$. Let $S = (X_i)_{i=1}^n$ be an ϵ-corrupted version of $G^\star$. Algorithm 2 outputs a vector u such that $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\delta/\rho})$.*

Theorem 1.3 is an immediate corollary of Theorem 4.1.

We can apply Theorem 4.1 to get an end-to-end result for subgaussian distributions. Algorithm 2 requires the stability conditions (Definition 2.2) of the original good samples $G^\star$. We show that these conditions are satisfied with a small number of samples.

Lemma 4.2. *Let $0 < \rho \leq 1$ and $0 < \epsilon < \epsilon_0$. Let D be a centered subgaussian distribution with covariance $I + \rho vv^\top$ for a k-sparse unit vector $v \in \mathbb{R}^d$. Let $G^\star$ be a set of $n = \Omega(k^2 \log d/\delta^2)$ samples drawn from D. Then then with probability at least $1 - \exp(-\Omega(k^2 \log d))$, $G^\star$ is (k,ϵ,δ)-stable (as in Definition 2.2) w.r.t. D for $\delta = O(\epsilon\log(1/\epsilon))$.*

Combining Theorem 4.1 and Lemma 4.2, given as input an ϵ-corrupted set of $n = \widetilde{\Omega}(k^2 \log d/\epsilon^2)$ samples drawn from a centered subgaussian distribution with covariance $I + \rho vv^\top$, Algorithm 2 returns a vector u with $\left\|uu^\top - vv^\top\right\|_F = O(\sqrt{\epsilon\log(1/\epsilon)/\rho})$.

We defer the proofs of Lemma 4.2 and Theorem 4.1 to Appendix B and give an overview of the proof of Theorem 4.1.

Proof Sketch of Theorem 4.1. We can use the stability conditions to upper bound the optimal objective value: note that for $w^\star$ (uniform distribution on the remaining good samples), we must have $\left\|M_{w^\star} - (I + \rho vv^\top)\right\|_{F,2k^2} \leq \delta$ by the stability conditions, therefore $\|M_{w^\star} - I\|_{F,2k^2} \leq \left\|M_{w^\star} - (I + \rho vv^\top)\right\|_{F,2k^2} + \left\|\rho vv^\top\right\|_{F,2k^2} \leq \rho + \delta$. Because the objective function $f(w)$ is convex, any stationary point w must be globally optimal and satisfies $f(w) \leq \rho + \delta$.

Fix a stationary point w and let $A = M_w - I$. Let R be the support of $vv^\top$ and let Q be the set of k^2 largest entries of A. The stability conditions implies for any w, the projection in the v direction must be large (formally $v^\top Av \geq \rho - \delta$). Because the objective function measures the norm of the largest $2k^2$ entries of A and it is not much larger than the norm of the largest k^2 entries, we can argue that A_R and A_Q are close, so $v^\top A_Q v \geq \rho - O(\delta)$.

Now A_Q is a matrix with Frobenius norm at most $\rho + \delta$ while $v^\top A_Q v \geq \rho - O(\delta)$. Together these imply that the norm of A_Q in direction orthogonal to $vv^\top$ is at most $O(\sqrt{\rho\delta})$, and then by standard matrix perturbation bounds we know the top eigenvector of $(A_Q + A_Q^\top)$ is $O(\sqrt{\delta/\rho})$ close to v.

5 Experiments

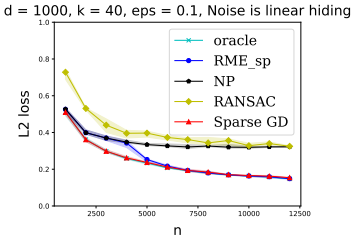
We perform an experimental evaluation of our robust sparse mean estimation algorithm on synthetic datasets with a focus on statistical accuracy (ℓ_2-distance between the output and the true sparse mean). We evaluate our algorithm (Sparse Gradient Descent, `Sparse GD`) on different noise models, and compare it to the following previous algorithms:

- `oracle`, which is told exactly which samples are inliers, and outputs their empirical mean,
- the robust sparse mean estimation algorithm `RME_sp` from [DKK+19b],
- `NP` (Naive Pruning), which removes samples far from the median and output the mean of the rest,
- `RANSAC`, which randomly selects half of the points and computes their mean. One solution is preferred to another if it has more points in a ball of radius $O(\sqrt{d})$ around it.

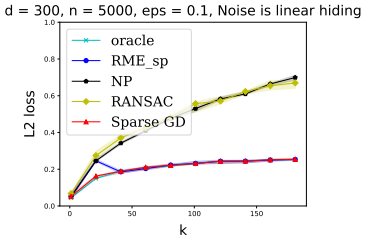
For algorithms that output non-sparse vectors, we take the largest k entries before measuring the ℓ_2 distance to the true mean. We evaluate the algorithms on various noise models:

- **Linear-hiding noise.** The inliers are drawn from $\mathcal{N}(0, I)$. Let S be a size k set. Then, half the outliers are drawn from $\mathcal{N}(1_S, I)$ and the other half are drawn from $\mathcal{N}(0, 2I - I_S)$.
- **Tail-flipping noise.** This noise model picks a k-sparse direction v and replaces the ϵ fraction of points farthest in the $-v$ direction with points in the $+v$ direction.
- **Constant-bias noise.** This model adds a constant to every coordinate of the outlier points. In Figure 3, we add 2 to every coordinate of every outlier point.

We ran our experiments on a computer with a 1.6 GHz Intel Core i5 processor and 8 GB RAM. We built on the codebase of [DKK+19b] [4] and implemented our new algorithm for the experiments. For each pair of algorithm and noise model, we repeat the same experiment 10 times and plot the median value of the measurements. We shade the interquartile region around the reported points in the figure as confidence intervals.



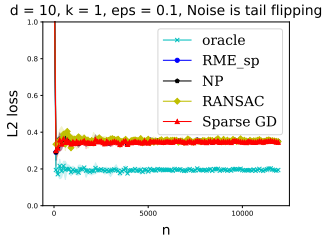
(a) Fix the sparsity k and change the number of samples n.



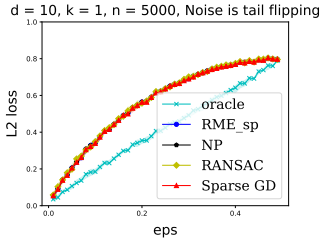
(b) Fix n and change the sparsity k.

Figure 1: The performance of various algorithms under linear-hiding noise. Notably, when the number of samples n or the sparsity k is small, our algorithm `Sparse GD` outperforms `RME_sp`.

[4] Available at: https://github.com/sushrutk/robust_sparse_mean_estimation, MIT license

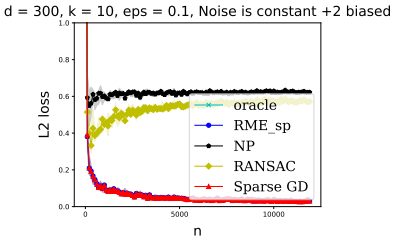


(a) Fix the sparsity k and change the number of samples n.

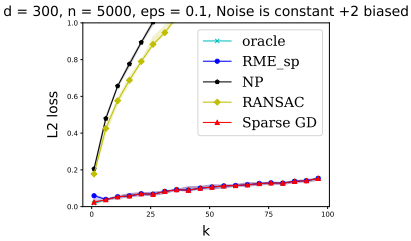


(b) Fix n, k and change the fraction of corruption ϵ.

Figure 2: The performance of various algorithms in the tail-flipping noise model.



(a) Fix the sparsity k and change the number of samples n.



(b) Fix n and change the sparsity k.

Figure 3: The performance of various algorithms in the constant-bias noise model.

Our experimental results are summarized in Figures 1, 2 and 3. For the linear-hiding and constant-bias noise models, we run two experiments: 1) fix the sparsity k and change the number of samples n, and 2) fix n and change k. For the tail-flipping noise model, we run two experiments: 1) fix the sparsity k and change the number of samples n, and 2) fix k and n and change the fraction of corruption ϵ.

In terms of statistical accuracy, our algorithm (`Sparse GD`), outperforms the filter-based `RME_sp` algorithm [DKK+19b] in the linear-hiding noise model when the number of samples or the sparsity is small, as shown in Figure 1. Our algorithm matches the performance of `RME_sp` under the tail flipping and constant-bias noise models, as shown in Figures 2 and 3.

Matching our theoretical results, our `Sparse GD` algorithm has accuracy $O(\epsilon\sqrt{\log(1/\epsilon)})$ and is within a constant factor of the $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ worst-case performance of `oracle`. In contrast, the naive algorithms `NP` and `RANSAC` both incur error that scales as $\epsilon\sqrt{k}$. The tail-flipping noise (Figure 2) illustrates that $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ error can occur no matter which algorithm is used (including `oracle`), because ϵ-fraction of the original good samples was removed.

Acknowledgments and Disclosure of Funding

Yu Cheng is supported in part by NSF Award CCF-2307106. Ilias Diakonikolas is supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), NSF Award AiTF-2006206, a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Rong Ge is supported by NSF Award CCF-1704656, NSF Award CCF-1845171 (CAREER), NSF Award CCF-1934964, a Sloan Research Fellowship, and a Google Faculty Research Award. Shivam Gupta is supported by NSF Award CCF-2008868, NSF Award AiTF-2006206, and the NSF AI Institute for Foundations of Machine Learning (IFML). Some of this work was done while Shivam Gupta was visiting the University of Wisconsin-Madison. Daniel M. Kane is supported by NSF Award CCF-1652862 (CAREER), NSF Award CCF-2107547, a Sloan Research Fellowship, and a grant from CasperLabs. Mahdi Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship, NSF Award CCF-1846369 (CAREER), NSF Award CCF-1813877, DARPA Learning with Less Labels (LwLL) and Fast Network Interface Cards (FastNICs) grants, and Faculty Research Awards from Google and Amazon.