A Sub-Quadratic Time Algorithm for Robust Sparse Mean Estimation

Ankit Pensia
IBM Research
ankitp@ibm.com

February 17, 2024

Abstract

We study the algorithmic problem of sparse mean estimation in the presence of adversarial outliers. Specifically, the algorithm observes a *corrupted* set of samples from $\mathcal{N}(\mu, \mathbf{I}_d)$, where the unknown mean $\mu \in \mathbb{R}^d$ is constrained to be k-sparse. A series of prior works has developed efficient algorithms for robust sparse mean estimation with sample complexity $\mathrm{poly}(k, \log d, 1/\epsilon)$ and runtime $d^2\mathrm{poly}(k, \log d, 1/\epsilon)$, where ϵ is the fraction of contamination. In particular, the fastest runtime of existing algorithms is quadratic in the dimension, which can be prohibitive in high dimensions. This quadratic barrier in the runtime stems from the reliance of these algorithms on the sample covariance matrix, which is of size d^2.

Our main contribution is an algorithm for robust sparse mean estimation which runs in *subquadratic* time using $\mathrm{poly}(k, \log d, 1/\epsilon)$ samples, with similar results for robust sparse PCA. Our results build on algorithmic advances in detecting weak correlations, a generalized version of the light-bulb problem by Valiant [Val15].

1 Introduction

Mean estimation, a fundamental unsupervised inference task studied in literature, may be described as follows: Given a family of distributions $\mathcal{P}$ over $\mathbb{R}^d$, the algorithm observes a set of i.i.d. points from an unknown $P \in \mathcal{P}$, with the goal of outputting $\widehat{\mu}$ such that, with high probability, $\|\widehat{\mu} - \mu\|_2$ is small. Although this framework is well-studied in the literature, the data observed in practice may deviate from the i.i.d. assumption and additionally may contain outliers. Crucially, these outliers can easily break standard off-the-shelf estimators, for example, sample mean, geometric median, and coordinate-wise median. To address this challenge, the field of robust statistics was initiated in the 1960s, aiming to develop algorithms that are robust to outliers [Hub64; ABHHRT72; HR09]. Before proceeding further, we formally define the contamination model we study in this paper.

Definition 1.1 (Strong Contamination Model)**.** Given a *corruption* parameter $\epsilon \in (0, 1/2)$ and a distribution P on uncorrupted samples, an algorithm obtains samples from P with *ϵ-contamination* as follows: (i) The algorithm specifies the number n of samples it requires. (ii) A set S of n i.i.d. samples from P is drawn but not yet shown to the algorithm. (iii) An arbitrarily powerful adversary then inspects S, before deciding to replace any subset of $\lceil \epsilon n \rceil$ samples with arbitrarily corrupted points ("outliers") to obtain the contaminated set T, which is then returned to the algorithm. We say T is an ϵ-corrupted version of S and a set of ϵ-corrupted samples from P.

Our focus will be on high-dimensional distributions, i.e., when the distribution P is over $\mathbb{R}^d$ for large d. Dealing with outliers becomes harder in high dimensions because *classical* outlier screening procedures (which otherwise work well in low dimensions) rely on the norm of the data points and are too coarse to distinguish outliers from inliers. Nevertheless, a long line of research, spurred by advances in [DKKLMS16; LRV16], has developed a systematic theory of handling outliers in high-dimensional robust statistics [DK23]. Notwithstanding this progress, major gaps persist in our *fine-grained* understanding of fast robust algorithms for data with additional structure.

Structured high-dimensional data distributions are ubiquitous in practice, e.g., natural images and sounds. Moreover, leveraging these underlying structures often dramatically improves algorithmic performance, e.g., in terms of error. A well-studied structure both in the theory and practice of high-dimensional statistics is *sparsity*, see, for example, the textbooks [EK12; HTW15; van16]. Consequently, we concentrate our efforts on *structured* mean estimation, where we assume that the underlying mean is sparse, i.e., an overwhelming majority of its coordinates are zero.

In light of the challenges posed by outliers above and the prevalence and importance of sparsity, we study the problem of *robust* sparse mean estimation. We say a vector $x \in \mathbb{R}^d$ is k-sparse if x has at most k non-zero entries. Our focus is on the practically relevant regime where k is much smaller than d, say, poly-logarithmic in d. We formally define robust sparse mean estimation below.

Problem 1.2 (Gaussian Robust Sparse Mean Estimation)**.** Let $\epsilon_0 \in (0, 1/2)$ be a sufficiently small constant. Given $\epsilon \in (0, \epsilon_0)$, sparsity $k \in \mathbb{N}$, and a set of ϵ-corrupted set of samples from $\mathcal{N}(\mu, \mathbf{I}_d)$ with an unknown k-sparse mean $\mu \in \mathbb{R}^d$, the goal is to output an estimate $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2$ is small with high probability.

Robust sparse mean estimation algorithms, efficient both in runtime and samples, were first developed in [BDLS17], with sample complexity $n = \mathrm{poly}(k, \log d, 1/\epsilon)$, runtime $\mathrm{poly}(d, n, 1/\epsilon)$, and near-optimal error $\|\widehat{\mu} - \mu\|_2 = \tilde{O}(\epsilon)$.[1] In particular, the sample complexity is only poly-logarithmic in the ambient dimension d, thereby permitting statistical inference with far fewer samples than the $\Omega(d)$ samples required by unstructured mean estimation. Therefore, for our algorithm for Problem 1.2, we set as the first requirement this sample complexity of $\mathrm{poly}(k, \log d, 1/\epsilon)$.

The focus of this work is to develop *fast* robust sparse mean estimation with the aforementioned sample complexity. Although the runtime of [BDLS17] is polynomial in dimension, their algorithm uses the ellipsoid algorithm (which in turn solves a semidefinite program) and hence is not practical in high dimensions. [DKKPS19] then developed a *spectral* algorithm with similar error guarantees and sample complexity and an improved runtime of $d^2\mathrm{poly}(k, \log d, 1/\epsilon)$. Subsequent papers have proposed many algorithmic improvements and generalizations to a wider class of distributions [ZJS22; CDKGGS22; DKKPP22; DKLP22]; see Section 1.3.

Despite this algorithmic progress, the fastest currently known algorithm for Problem 1.2 is that of [DKKPS19] with runtime scaling as d^2. This quadratic runtime of the algorithm can be prohibitive in high dimensions—the very setting that benefits most from sparsity (because of sample-efficiency). This quadratic dimension dependence is in stark contrast to the *non-robust* setting (i.e., the outlier-free regime), where there exists a simple (folklore) algorithm[2] with nearly-linear runtime, which is also minimax optimal. This motivates the following fundamental question:

Question 1.3. *Are there any nearly-linear time algorithms for robust sparse mean estimation?*

If we momentarily forgo sparsity (and the benefits that come along with it, e.g., the reduced sample complexity and interpretability) and focus on robust *dense* estimation, then positive answers

[1] In the presence of outliers, vanishing error is usually not possible. In our setting, this is because it is impossible to distinguish two isotropic Gaussian distributions that are $\Omega(\epsilon)$-far apart in the presence of ϵ-fraction of contamination.

[2] The algorithm computes the sample mean and thresholds entries to ensure sparsity, hence failing if there is even a single outlier. Moreover, natural attempts to make this algorithm robust, such as coordinate-wise median, incur an highly suboptimal error of $\Omega(\epsilon\sqrt{k})$.

are known to Question 1.3, see, e.g., [CDG19; DHL19; DKKLT22; DKPP22]. However, the sample complexities of the algorithms in these papers scale *linearly* with dimension,[3] which considerably exceeds our allowed budget of $\mathrm{poly}(k, \log d, 1/\epsilon)$ samples.

In fact, as alluded to earlier, existing attempts at answering Question 1.3 do not even break the *quadratic* runtime barrier. This is due to natural technical obstacles within current algorithms: to robustly estimate the mean, they crucially rely on the sample covariance matrix to detect outliers; but merely computing the sample covariance matrix costs $\Omega(d^2)$ time! Sparsity also precludes common tricks such as the power iteration to bypass explicitly writing the covariance matrix. Indeed, in certain parameter regimes, even detecting atypical values of the covariance matrix from samples is conjectured to require $\Omega(d^2)$ time [DS18]. This begs the question whether this quadratic gap is inherent:

Question 1.4. *Is there an algorithm for robust sparse mean estimation that runs in $d^{2-\Omega(1)}$ time and uses $\mathrm{poly}(k, \log d, 1/\epsilon)$ samples?*

The main result of our work is an affirmative answer to Question 1.4. We hope our answer paves the path for progress towards answering Question 1.3, which was highlighted as an important open problem in [Dia19; Che21; Dia23]. Our algorithm builds on advances in fast correlation detection algorithms by Valiant [Val15].

1.1 Our Results

We establish the following result:

Theorem 1.5 (Robust Sparse Mean Estimation in Subquadratic Time)**.** *Let the contamination rate be $\epsilon \in (0, \epsilon_0)$ for a small constant $\epsilon_0 \in (0, 1/2)$ and $k \in \mathbb{N}$ be the sparsity. Let T be an ϵ-corrupted set of n samples from $\mathcal{N}(\mu, \mathbf{I}_d)$ for an unknown k-sparse $\mu \in \mathbb{R}^d$. Then there is a randomized algorithm $\mathcal{A}$ that takes as input the corrupted set T, contamination rate ϵ, sparsity $k \in \mathbb{N}$, and a parameter $q \in \mathbb{N}$, and produces an estimate $\widehat{\mu}$ such that*

- *(Sample Complexity and Error) If $n \gtrsim (k^{2q} \log d)/\epsilon^{2q}$, then $\|\widehat{\mu} - \mu\|_2 \lesssim \epsilon\sqrt{\log(1/\epsilon)}$ with probability at least 0.9 over the randomness of the samples and the algorithm.*[4]
- *(Runtime) The algorithm runs in time at most $d^{1.62+\frac{3}{q}}\mathrm{poly}(\log(d), k^q, 1/\epsilon^q)$.*

Several remarks are in order. The error guarantee of Theorem 1.5, $O(\epsilon\sqrt{\log(1/\epsilon)})$, is nearly optimal even given infinite data and runtime.[5] The main contribution of Theorem 1.5 is the first algorithm for robust sparse mean estimation with runtime $d^{2-\Omega(1)}\mathrm{poly}(k/\epsilon)$ and sample complexity $\mathrm{poly}(k, \log d, 1/\epsilon)$ (by selecting $q \in \mathbb{N}$ to be a constant bigger than 9), thereby affirmatively answering Question 1.4. As q increases, the dependence of the runtime on the dimension approaches $d^{1.62}$. In particular, for a constant contamination rate ϵ, we may set q as large as $\gamma\left(\frac{\log d}{\log k}\right)$ for a small $\gamma \in (0, 1)$, and the algorithm retains sublinear (in d) sample complexity $d^{O(\gamma)}$ and subquadratic runtime $d^{1.62+O(\gamma)}k^{O(1/\gamma)}$. Finally, the sample complexity of Theorem 1.5 is (polynomially) larger than existing works; See Section 5 for further remarks.

[3] In fact, the overall runtime of these algorithms scales as $\tilde{\Theta}(nd) = \tilde{\Theta}(d^2/\epsilon^2)$, which is again quadratic in dimension.

[4] The success probability can be boosted to $1-\delta$ with a multiplicative increase of $\log(1/\delta)$ in the sample complexity and the runtime by repeating the procedure.

[5] The information-theoretic optimal error is $\Theta(\epsilon)$. Moreover, it is computationally hard to beat $\Theta\left(\epsilon\sqrt{\log(1/\epsilon)}\right)$ in the statistical query lower model [DKS17] and the low-degree polynomial tests [BBHLS21] under Definition 1.1.

We next focus on robust version of sparse principal component analysis (PCA). Sparse PCA is a fundamental estimation task in high-dimensional statistics [dGJL07; HTW15], in which the algorithm observes samples from $\mathcal{N}(0, \mathbf{I} + \eta vv^\top)$ for a k-sparse unit vector $v \in \mathbb{R}^d$. Similarly to sparse mean estimation, the standard algorithms for sparse PCA are not robust to outliers, and hence we study robust sparse PCA. Again, existing robust sparse PCA algorithms from [BDLS17; DKKPS19; CDKGGS22] require at least $\Omega(d^2)$ time. In contrast, we establish the following result:

Theorem 1.6 (Robust Sparse PCA in Subquadratic Time)**.** *Let T be an ϵ-corrupted set of samples from $\mathcal{N}(0, \mathbf{I}_d + \eta vv^\top)$ for $\eta \in (0,1)$ and a k-sparse unit vector. There is a randomized algorithm that takes as input the corrupted set T, contamination rate ϵ, sparsity $k \in \mathbb{N}$, spike η, and a parameter $q \in \mathbb{N}$, and produces an estimate $\widehat{v}$ such that*

- *(Sample Complexity and Error) If $n \gtrsim \mathrm{poly}((k^q \log d)/\epsilon^q)$, then $\|\widehat{v}\widehat{v}^\top - vv^\top\|_{\mathrm{Fr}} \lesssim \sqrt{\epsilon \log(1/\epsilon)/\eta}$ with probability at least $1 - \frac{1}{\mathrm{poly}(d)}$.*
- *(Runtime) The algorithm runs in time at most $d^{1.62 + \frac{3}{q}} \mathrm{poly}(n)$.*

This result gives the first subquadratic time algorithm for dimension-independent error, improving on the $\Omega(d^2)$ runtime of [BDLS17; DKKPS19; CDKGGS22]. We note that the error guarantee of Theorem 1.6 is sub-optimal by a polynomial factor of ϵ/η (like in [CDKGGS22]), since the information-theoretic optimal error is ϵ/η. Despite this (polynomially) larger error, Theorem 1.6 is the first subquadratic time algorithm for robust sparse PCA with any non-trivial error, say, less than 0.01. We defer detailed discussion to Section 4.

Our main technical ingredient in proving Theorems 1.5 and 1.6 is a result on detecting correlated coordinates of a high-dimensional distribution by [Val15]. We give an overview of Theorem 1.5 in Section 1.2, with details in Section 3, and defer Theorem 1.6 to Section 4.

1.2 Overview of Techniques

We begin by presenting a brief overview of the landscape of current robust sparse mean estimation algorithms, followed by challenges in using these approaches to obtain an $o(d^2)$ runtime, and then conclude by presenting our algorithm.

(Dense) Robust Mean Estimation Let μ' and $\mathbf{\Sigma}'$ be the sample mean and the sample covariance of the current (corrupted) dataset. The general guiding principle in robust *dense* mean estimation is to use $\mathbf{\Sigma}'$ to check if there are harmful outliers and iteratively remove them. Recall that inliers are sampled from an isotropic covariance distribution $\mathcal{N}(\mu, \mathbf{I}_d)$. Thus, if we take $\Theta(d/\epsilon^2)$ samples, then the variance of the inliers in any direction is $(1 \pm \tilde{O}(\epsilon))$. Moreover, the variance of any $(1 - \epsilon)$ fraction of inliers is $(1 \pm \tilde{O}(\epsilon))$.

The following are the key insights in developing algorithms for robust dense mean estimation: (i) Outliers cannot change the sample mean μ' of the data in any direction v without significantly increasing the covariance $\mathbf{\Sigma}'$ in the direction v, (ii) Given a direction of large variance v of the data (i.e., with variance larger than $1 + \tilde{\Omega}(\epsilon)$), one can reliably remove outliers by projecting the data onto v and thresholding appropriately, and (iii) In the dense setting, the directions of large variance correspond to leading eigenvectors of the covariance matrix $\mathbf{\Sigma}'$, and further they can be computed efficiently (in nearly-linear time) using power iteration. Thus, one can iteratively remove outliers as follows: compute (approximately) the leading eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{\Sigma}'$ and remove the samples that have large projections along the computed direction.

Adapting to Sparsity and Smaller Sample Complexity For robust *sparse* mean estimation, one can adapt the above strategy by focusing only on the sparse directions v. Indeed, (i) and (ii) above are straight-forward and the resulting sample complexity of the algorithm is $k\log(d)/\epsilon^2$ since we require concentration of the mean and the covariance only along k-sparse directions. However, the problem of computing the direction of the leading *sparse* eigenvalues of a matrix, $\max_{v:\|v\|_2=1,k\text{-sparse}} v^\top \mathbf{\Sigma}' v$, is computationally-hard in the worst case. Inspired by the literature on sparse PCA [dGJL07], [BDLS17] proposed the following convex relaxation[6]:

$$\sup_{\{\mathbf{A}:\mathbf{A}\succeq 0,\mathrm{tr}(\mathbf{A})=1,\|\mathbf{A}\|_1\le k\}} \langle \mathbf{A}, \mathbf{\Sigma}' - \mathbf{I}_d\rangle . \tag{1}$$

Given such a feasible $\mathbf{A}$ with value larger than $\tilde{\Omega}(\epsilon)$, one can remove outliers provided a larger sample complexity of $(k^2\log d)/\epsilon^2$.[7] Although the resulting algorithm is polynomial-time and the desired sample complexity $\mathrm{poly}(k, \log d, 1/\epsilon)$, the algorithm requires solving semidefinite programs (SDPs), for which the current algorithms require time superquadratic in dimension [JKLPS20].

Spectral Algorithm of [DKKPS19] To avoid solving the SDPs from the preceding paragraph, [DKKPS19] considers a different (and, in a sense, weaker) relaxation of sparse eigenvalues by relying on the distributional properties of Gaussians. Let $\mathcal{B}_{k^2\text{-sparse}}$ be the set of all (sparse) matrices $\mathbf{B}$ with Frobenius norm 1 and at most k^2 non-zero entries. Importantly, $\mathcal{B}_{k^2\text{-sparse}}$ contains all $vv^\top$ for all k-sparse unit vectors v. Their starting point is the observation is that for any $\mathbf{B}\in\mathcal{B}_{k^2\text{-sparse}}$, we have $\mathrm{Var}_{x\sim\mathcal{N}(0,\mathbf{I}_d)}(x^\top\mathbf{B}x) = 2\|\mathbf{B}\|_{\mathrm{Fr}}^2 = 2$. Thus, for a fixed $\mathbf{B}\in\mathcal{B}_{k^2\text{-sparse}}$, the empirical mean of $x^\top\mathbf{B}x$ over any $1-\epsilon$ fraction of inliers should be $\mathrm{tr}(\mathbf{B})\pm\tilde{O}(\epsilon)$. Moreover, standard uniform concentration arguments imply that this holds uniformly over $\mathbf{B}\in\mathcal{B}_{k^2\text{-sparse}}$ given $(k^2\log d)/\epsilon^2$ samples. Their key observation is that the resulting (non-convex) optimization problem

$$\max_{\mathbf{B}\in\mathcal{B}_{k^2\text{-sparse}}} \langle \mathbf{B}, \mathbf{\Sigma}' - \mathbf{I}_d\rangle \tag{2}$$

can be solved via standard matrix operations (despite being non-convex) without resorting to SDPs (as opposed to (1)); indeed, the optimal $\mathbf{B}$ corresponds to the top-k^2 values of $\mathbf{\Sigma}'-\mathbf{I}_d$ in magnitude, computable in $\tilde{O}(d^2)$ time given $\mathbf{\Sigma}'$.

Given such a feasible $\mathbf{B}\in\mathcal{B}_{k^2\text{-sparse}}$ that achieves $\langle\mathbf{B},\mathbf{\Sigma}'-\mathbf{I}_d\rangle\ge\tilde{\Omega}(\epsilon)$, [DKKPS19] also propose an efficient outlier-removal strategy. Overall, this yields an algorithm with runtime $d^2\mathrm{poly}(k/\epsilon)$ and sample complexity $(k^2\log d)/\epsilon^2$. While a significant improvement over [BDLS17], this still unfortunately falls short of our target runtime of $o(d^2)$.

The main challenge in extending [DKKPS19]'s algorithm to get an $o(d^2)$ runtime is that one needs to write down $\mathbf{\Sigma}'$ explicitly, which itself takes $\Omega(d^2)$ time. Moreover, there is no known analog of power iteration for sparse settings with provable guarantees (recall that in the dense setting, the power iteration can be implemented in nearly-linear runtime [SV14]). Our main technical insight is to use advances in fast algorithms for correlation detection initiated by [Val15]; see Section 2.4 for precise statements. Next, we explain why [Val15] is potentially useful in our setting, explain the challenges in a direction application of their result, and our proposed fix.

Fast Correlation Detection Algorithm To The Rescue Denote the correlation detection algorithm in [Val15] by $\mathcal{A}_{\mathrm{corr}}$. In our setting, this algorithm guarantees that given a $\rho\in(0,1)$ and

[6] For a matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$, $\|\mathbf{A}\|_1$ denotes the ℓ_1-norm of A when flattened as a d^2-dimensional vector

[7] This larger sample complexity, k^2 versus k, is due to the stronger concentration required by the relaxation.

a large $q \in \mathbb{N}$, it can find (off-diagonal) coordinate pairs (i, j) such that $|\mathbf{\Sigma}'_{i,j}| \geq \rho$ in subquadratic time *as long as* there are at most $O(d)$ off-diagonal coordinate pairs $(i,', j')$ such that $|\mathbf{\Sigma}'_{i',j'}| \geq \rho^q$; observe that $\rho^q \ll \rho$.

We now illustrate why such a subroutine may be useful. Suppose that $\mathcal{A}_{\text{corr}}$ returns many correlated coordinate pairs. Then the optimal value in (2) must be large (if we optimize over $\mathcal{B}_{k'\text{-sparse}}$ for some k' large enough[8]), and we can use those coordinates pairs to construct a k'-sparse $\mathbf{B}$ that can be used to remove outliers as before [DKKPS19]. If, on the other hand, $\mathcal{A}_{\text{corr}}$ returns only a few coordinate pairs, then we know that only these small set of coordinates are (potentially) corrupted, and we have reduced our problem to robust dense mean estimation on these coordinates; recall that in this setting, the sample mean is a good candidate for the remaining coordinates. Thus, a fast correlation detection algorithm leads to a subquadratic time algorithm to filter outliers (or declare victory) provided that there are only $O(d)$ coordinate pairs with correlation larger than ρ^q.

Challenges in Applying Fast Correlation Detection and A Proposed Fix A priori, it is unclear why there must be only $O(d)$ correlated coordinate pairs: indeed, the outliers are allowed to be dense (similar to inliers — recall that only the population mean is sparse), and, in the worst case, it is possible that they cause *all* coordinate pairs to be correlated (on the corrupted data). Thus, we need an alternative procedure to (i) detect if there are many ρ^q-correlated pairs and (ii) if so, find an alternative procedure to make progress.

Fortunately, it turns out that if there are $\Omega(d)$ correlated pairs, then a random pair has $\Omega(d/d^2) = \Omega(1/d)$ probability to be correlated. Hence, we can sample a relatively large — but subquadratic, say $\Theta(d^{1.5})$ — number of random pairs to estimate the true count of correlated pairs. If random sampling does not find many such pairs, then the true count would anyway have been small with high probability, and we may safely invoke [Val15]'s algorithm, solving the detection problem (i) above. On the other hand, if we do observe many (i.e., scaling polynomially with d) ρ^q-correlated pairs, then we know that the Frobenius norm of the largest $k' = \text{poly}(k)$ entries of $\mathbf{\Sigma}' - \mathbf{I}_d$ must be large enough, $\Omega(\rho^q \sqrt{k'})$. In other words, we can find a relatively sparse $\mathbf{B}' \in \mathcal{B}_{k'\text{-sparse}}$ such that $\langle \mathbf{\Sigma}' - \mathbf{I}_d, \mathbf{B}' \rangle \geq \tilde{\Omega}(\epsilon)$. Thus, we can iteratively remove outliers (or declare victory when safe to do so) irrespective of the number of correlated pairs. Finally, the larger sample complexity of algorithm comes from invoking the filter on k'-sparse matrices $\mathbf{B}$ for $k' \gg k$. We give a more detailed overview in Section 3.

1.3 Related Work

Our work is situated within the field of algorithmic robust statistics, and we refer the reader to [DK23] for an extensive exposition on the topic.

Robust Sparse Estimation Efficient algorithms for robust sparse mean estimation were first developed in [BDLS17], giving an algorithm to compute $\widehat{\mu}$ with sample complexity $\text{poly}(k, \log d, 1/\epsilon)$, runtime $\text{poly}(d, 1/\epsilon)$, and near-optimal error $\|\widehat{\mu} - \mu\|_2 = \tilde{O}(\epsilon)$. Their algorithm used the ellipsoid algorithm with a separation oracle that requires solving an SDP. Invoking the ellipsoid algorithm can be avoided by using [ZJS22] or through iterative filtering, but the resulting algorithm still requires solving multiple SDPs, which, as noted earlier, is inherently slow. Bypassing the use of SDPs, [DKKPS19] developed the first spectral algorithm for robust sparse estimation with runtime $d^2\text{poly}(k, \log d, 1/\epsilon)$. Another novel take on this problem was seen in [CDKGGS22], which gave

[8]Recall that we can interpret the maximum in (2) as Euclidean norm of the largest k^2 entries,which would be at least $\rho\sqrt{k^2} = \rho k$.

an optimization-based algorithm showing that first-order stationary points of a natural non-convex objective suffice. Although the resulting algorithm relies on simple matrix operations, the derived runtime is super-quadratic in dimension. In a different direction, [DKKPP22] and [DKLP22] developed algorithms for robust sparse mean estimation for a wider class of distributions: heavy-tailed distributions and light-tailed distributions with unknown covariance, respectively.

Robust sparse mean estimation is conjectured to have information-computation gaps [DKS17; BB20; DKKPP22]. Particularly, while there exist inefficient (d^k-time) algorithms for robust sparse mean estimation using $(k \log(d))/\epsilon^2$ samples, all polynomial time algorithms are conjectured to require $\Omega\left(k^2/\epsilon^2\right)$ samples [BB20].

Fast Algorithms for Robust Estimation Looking beyond polynomial runtime as the criterion of computational efficiency, a recent line of work has investigated *faster* algorithms for a variety of robust estimation tasks: mean estimation [CDG19; DHL19; DL22; LLVZ20; DKKLT22; DKPP22; DKPP23a], covariance estimation [CDGW19], principal component analysis [JLT20; DKPP23b], list-decoding [CMY20; DKKLT22], and linear regression [CATJFB20; DKPP23a]. The overarching goal in this line of work is to develop robust algorithms that have runtimes matching the corresponding non-robust off-the-shelf algorithms, thus reducing the computational overhead of robustness. However, none of these algorithms is tailored to sparsity and hence have sample complexity scaling (at least) linearly with the dimension.[9] *Our main contribution is the first subquadratic runtime algorithm for robust sparse mean estimation with sample complexity* $\mathrm{poly}(k \log d, 1/\epsilon)$.

Fast Correlation Detection Given n vectors in $\{\pm 1\}^d$ and two thresholds $1 \geq \rho > \tau > 0$, the correlation detection problem asks to find all coordinate pairs that have correlation at least ρ given that not too many pairs have correlation larger than τ. This problem is a generalization of the light bulb problem [Val88]. The first subquadratic algorithm for both these problems was given by [Val15], with further improvements and simplifications in [KKK18; KKKC20; Alm19]. Further algorithmic improvements for this task would likely also improve our runtime guarantees in Theorems 1.5 and 1.6.

2 Preliminaries

Notation For a random variable X, $\mathbb{E}[X]$ denotes its expectation. For a finite set S and a function $g : S \to \mathbb{R}^d$, we use $\mathbb{E}_S[g(X)]$ to denote $(\sum_{x \in S} g(X))/|S|$. We use $\mathrm{poly}(\cdots)$ to denote an expression that is polynomial in its arguments. The notations $\lesssim, \gtrsim, \asymp$ hide absolute constants.

For a vector $x \in \mathbb{R}^d$, we use $\|x\|_0$ and $\|x\|_2$ to denote the number of non-zero entries of x and the ℓ_2-norm of x, respectively. For a vector x and $k \in \mathbb{N}$, we define the $\|x\|_{2,k}$ norm as the maximum correlation between x and a unit k-sparse vector, i.e., $\|x\|_{2,k} := \sup_{v : \|v\|_0 \leq k} \langle v, x \rangle$. Estimation in $\| \cdot \|_{2,k}$ immediately yields an estimate that is close in ℓ_2 norm whenever μ is k-sparse:

Proposition 2.1 (Sparse estimation using $\| \cdot \|_{2,k}$ norm [CDKGGS22])**.** *Let $x \in \mathbb{R}^d, y \in \mathbb{R}^d$, where y is k-sparse. Let $J \subset [d]$ be the top-k coordinates of x in magnitude, breaking ties arbitrarily. Define $x' \in \mathbb{R}^d$ to be $x'_i = x_i$ if $i \in J$ and 0 otherwise. Then $\|x' - y\|_2 \leq 6\|x - y\|_{2,k}$.*

Thus, in the sequel, we solve the harder problem of estimating an *arbitrary* mean $\mu \in \mathbb{R}^d$ in the $\| \cdot \|_{2,k}$ norm.

We denote matrices by bold capital letters, e.g., $\mathbf{A}, \mathbf{\Sigma}$. We denote the $d \times d$ identity matrix by $\mathbf{I}_d$, omitting the subscript when clear. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_0$, and $\|\mathbf{A}\|_{\mathrm{Fr}}$, to denote the

[9] As a result, the dependence on the runtime again becomes quadratic because $nd = \Omega(d^2)$.

number of non-zero entries and the Frobenius norm, respectively. For matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimensions, $\langle \mathbf{A}, \mathbf{B} \rangle$ to denotes the trace inner product $\mathrm{tr}(\mathbf{A}^\top \mathbf{B})$.

For a subset $H \subset [d]$, and a vector $x \in \mathbb{R}^d$, define $(x)_H$ to be $|H|$-dimensional vector that corresponds to the coordinates in H. Similarly, for a matrix $\mathbf{A}$, we define $(\mathbf{A})_H$ to be the $|H| \times |H|$ matrix corresponding to coordinates in H. For a square matrix $\mathbf{A}$, we use $\mathrm{diag}(\mathbf{A})$ and $\mathrm{offdiag}(\mathbf{A})$ to denote its diagonal and offdiagonal, respectively.

For a finite set $T \subset \mathbb{R}^d$, we define μ_T and $\mathbf{\Sigma}_T$ to be the sample mean and the sample covariance of T, respectively.[10] When the set T is clear from context, for a coordinate pair $(i,j) \in [d] \times [d]$ with $i \neq j$, we denote the correlation between these coordinates on T as $\mathrm{corr}(i,j) := \left| \mathbf{\Sigma}'_{i,j} / \sqrt{\mathbf{\Sigma}'_{i,i}\mathbf{\Sigma}'_{j,j}} \right|$ for $\mathbf{\Sigma}' = \mathbf{\Sigma}_T$. For a $\rho \in (0,1)$, we say coordinates (i,j) are ρ-correlated if $\mathrm{corr}(i,j) \geq \rho$.

Robust sparse estimation requires checking whether the current covariance matrix $\mathbf{\Sigma}'$ has small quadratic forms, $v^\top(\mathbf{\Sigma}' - \mathbf{I})v$, for sparse unit vectors. For a matrix $\mathbf{A}$ and $k \in \mathbb{N}$, we define the sparse operator norm, $\|\mathbf{A}\|_{\mathrm{op},k} := \sup_{v: \|v\|_2 = 1, \|v\|_0 \leq k} |v^\top \mathbf{A} v|$. Since computing $\|\cdot\|_{\mathrm{op},k}$ is computationally hard, we look at the following relaxation from [DKKPS19]: For a matrix $\mathbf{A}$, define $\|\mathbf{A}\|_{\mathrm{Fr},k^2} := \sup_{\mathbf{B}: \|\mathbf{B}\|_{\mathrm{Fr}} = 1, \|\mathbf{B}\|_0 \leq k^2} \langle \mathbf{A}, \mathbf{B} \rangle$. It can be seen that $\|\mathbf{A}\|_{\mathrm{op},k} \leq \|\mathbf{A}\|_{\mathrm{Fr},k^2}$ since $\mathbf{B}$ could be all $\pm vv^\top$ for k-sparse unit vectors v. Moreover, $\|\mathbf{A}\|_{\mathrm{Fr},k^2}$ is the Euclidean norm of the largest k^2 entries (in magnitude) of $\mathbf{A}$.

Since we will routinely look at the projections of the points on a subset of coordinates, we formally define it below:

Definition 2.2 (Projection of Pairs of Coordinates)**.** Let $H_{\mathrm{pair}} \subset [d] \times [d]$ be a set of pair of coordinates such that $(i,i) \notin H_{\mathrm{pair}}$ for any $i \in [d]$. For an even $m \in [d^2]$, we define the operator Proj_m that takes any such H_{pair} and returns a set in $[d]$ that has a cardinality at most m as follows:

- If $|H_{\mathrm{pair}}| \leq \frac{m}{2}$, return $\{i : (i,j) \in H_{\mathrm{pair}} \text{ or } (j,i) \in H_{\mathrm{pair}}\}$.
- Otherwise, let any $m/2$ distinct elements of H_{pair} be $(i_1, i_2), \ldots, (i_{m-1}, i_m)$. Return $\{i_j : j \in [m]\}$.

When the subscript m is omitted, we take m to be d^2.

Informally, the operator returns a set H such that for any matrix $\mathbf{A}$, for small m, $\|(\mathbf{A})_H\|_{\mathrm{Fr}}^2 \geq \sum_{(i,j) \in H_{\mathrm{pair}}} \mathbf{A}_{i,j}^2$, while for larger m, $\|(\mathbf{A})_H\|_{\mathrm{Fr}}^2 \geq m \min_{(i,j) \in H_{\mathrm{pair}}} \mathbf{A}_{i,j}^2$.

2.1 Deterministic Condition on Inliers

A recurring notion in developing robust algorithms is that of *stability*, which stipulates that the first and second moment of the data not change much under removal of a small fraction of data points.

Definition 2.3 (Stability)**.** For an $\epsilon \in (0, 1/2)$, $\delta \geq \epsilon$, and sparsity $k \in \mathbb{N}$, we say a set $S \subset \mathbb{R}^d$ is (ϵ, δ, k)-stable with respect to $\mu \in \mathbb{R}^d$ if the following holds for any subset $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$: (i) $\|\mathbb{E}_{S'}[X - \mu]\|_{2,k} \leq \delta$, and (ii) $\left\| \mathbb{E}_{S'}[(X-\mu)(X-\mu)^\top] - \mathbf{I}_d \right\|_{\mathrm{Fr},k^2} \leq \delta^2/\epsilon$.

The following result gives a nearly-tight bound on the sample complexity required to ensure stability.

Lemma 2.4 (Stability Sample Complexity [CDKGGS22, Lemma 3.3])**.** *Let S be a set of n i.i.d. samples from a subgaussian distribution P over $\mathbb{R}^d$ such that P has (i) mean $\mu \in \mathbb{R}^d$, (ii) identity*

[10] Not to be confused with $(\mathbf{\Sigma})_H$ when $H \subset [d]$.

covariance, and (iii) satisfies the Hanson-Wright inequality; In particular, $\mathcal{N}(\mu, \mathbf{I}_d)$ satisfies all three properties. Then if $n \gtrsim (k^2(\log d)/\epsilon^2)$, then S is (ϵ, δ, k)-stable with high probability with respect to μ for $\delta = C\epsilon\sqrt{\log(1/\epsilon)}$ where C is a large absolute constant.

We note that the deterministic condition in Definition 2.3 is slightly stronger than [CDKGGS22]—the condition for the covariance—but their proof continues to work with the same sample complexity for Definition 2.3.[11]

2.2 Randomized Filtering

We will use the following template of filtering algorithm from [DK23, Section 2.4.2] (after a slight change in parameters). The following template for filtering has now become a standard in algoirthmic robust statistics.

Algorithm 1 Randomized Filtering

1: Let $T_1 \leftarrow T$
2: $i \leftarrow 1$
3: **while** $T_i \neq \emptyset$ and T_i does not satisfy the stopping condition $\mathcal{S}$ **do**
4: Get the scores $f : T_i \to \mathbb{R}_+$ satisfying $\sum_{x \in T_i \cap S} f(x) \leq \sum_{x \in T_i \setminus S} f(x)$ and $\max_{x \in T_i} f(x) > 0$
5: $T_{i+1} \leftarrow T_i$
6: **for** each $x \in T_i$ **do**
7: Remove the point x from T_{i+1} with probability $\frac{f(x)}{\max_{x \in T_i} f(x)}$
8: $i \leftarrow i + 1$
9: **return** T_i

These filtering algorithms have become a standard template in algorithmic robust statistics. Here, the stopping condition $\mathcal{S}$ can be a generic condition that can be evaluated in $\mathcal{T}_{\text{stopping}}$ time using some algorithm $\mathcal{A}_s$—it can be a randomized algorithm (using independent randomness from Line 6) that succeeds with probability $1 - \tau$. We also require that whenever the stopping condition is not satisfied and the set T_i is an 10ϵ-corruption of S, then the scores $f : T_i \to \mathbb{R}_+$ satisfying the guarantees of Line 4 can be computed in time $\mathcal{T}_{\text{score}}$.

Theorem 2.5 (Guarantee of Algorithm 1; [DK23, Theorem 2.17])**.** *If the above stopping conditions and filter conditions are met, then Algorithm 1 returns a set $T' \subseteq T$ such that, with probability at least $8/9 - \tau|T|$,*

1. *Algorithm 1 runs in time $O(|T|(\mathcal{T}_{stopping} + \mathcal{T}_{score} + |T|))$.*
2. *Each set $T_i \subseteq T$ observed throughout the run of the algorithm (which includes T') is an 10ϵ-corruption of S.*
3. *T' satisfies the stopping condition $\mathcal{S}$.*

2.3 Certificate Lemma and Frobenius Norm Filtering

The following standard certificate lemma guides the algorithmic design: if the sample covariance matrix of the corrupted data is roughly isotropic, then the sample mean is a good estimate.

[11] We remark that [CDKGGS22, Lemma 3.3] claims to establish the deterministic condition for all isotropic subgaussian distributions. However, their proof crucially uses Hanson-Wright inequality, which does not hold for arbitrary isotropic subgaussian distributions.

Lemma 2.6 (Sparse Certificate Lemma, see, e.g., [BDLS17])**.** *Let T be an ϵ-corrupted version of S, where S is (ϵ, δ, k)-stable with respect to μ (cf. Definition 2.3). Then $\|\mu_T - \mu\|_{2,k} \lesssim \delta + \sqrt{\epsilon \|\mathbf{\Sigma}_T - \mathbf{I}_d\|_{\mathrm{op},k}}$.*

We now state the guarantee of filtering procedures, where the goal is to filter outliers from a corrupted set T. In dense mean estimation, the most common filters are based on the scores of the form $(v^\top(x - \mu_T))^2$ for a direction of large variance v; this filter is guaranteed to succeed as long as the covariance matrix $\mathbf{\Sigma}_T$ is far from the identity in operator norm. In our setting, we will need a stronger filter guarantee that is guaranteed to succeed under the weaker condition that the covariance (submatrix) matrix is far from the (submatrix) identity in the Frobenius norm; Observe that the operator norm of (the corresponding submatrix of) $\mathbf{\Sigma}_T - \mathbf{I}_d$ can be much smaller. The following lemma corresponds to the above situation and simplifies [DKKPS19, Steps 6-10 of Algorithm 1]:

Lemma 2.7 (Sparse Filtering Lemma)**.** *Let $\epsilon \in (0, \epsilon_0)$ for a small absolute constant ϵ_0. Let T be an ϵ-corrupted version of S, where S is (ϵ, δ, k)-stable with respect to μ. Let $H \subset [d]$ be such $\|(\mathbf{\Sigma}_T - \mathbf{I})_H\|_{\mathrm{Fr}} = \lambda$ for $\lambda \gtrsim \delta^2/\epsilon$ and $|H| \leq k$. There exists an algorithm $\mathcal{A}$ that takes T, H, ϵ, and δ and returns scores $f : T \to \mathbb{R}_+$ so that $\sum_{x \in S \cap T} f(x) \leq \sum_{x \in T \setminus S} f(x)$, i.e., the sum of scores over inliers is less than that of outliers, and $\max_{x \in T} f(x) > 0$. Moreover, the algorithm runs in time $d \cdot \mathrm{poly}(|H||T|)$.*

These scores can be used to filter points from T such that on expectation over the algorithm's randomness, more outliers are removed than inliers [DK23]. We give a proof of Lemma 2.7 in Appendix A.

Thus, if $\mathbf{\Sigma}_T - \mathbf{I}_d$ has large (sparse) Frobenius norm, then we can make progress by removing outliers. The contribution to this norm from the *diagonal* entries can be calculated efficiently in $O(dn)$ time, and if large, then can also be used to remove outliers. Thus, we will assume that the corrupted set has already been pre-processed to satisfy the following:

Condition 2.8 (Preprocessing)**.** Let T be an ϵ-corrupted version of S, where S is (ϵ, δ, k)-stable. Suppose T satisfies $\|\mathrm{diag}(\mathbf{\Sigma}_T - \mathbf{I}_d)\|_{\mathrm{Fr},k^2} \leq \min\left(O(\delta^2/\epsilon), 0.5\right)$.

For completeness, we give details in Appendix A.2. The next result, also proved in Appendix A.2, shows that any further small modifications of the preprocessed sets retains small sparse operator norm.

Claim 2.9. *Let C be a large enough constant $C > 0$. Let $T'' \subset T'$ be two $O(\epsilon)$-contamination of S such that S is an $(C\epsilon, \delta, k)$-stable with respect to μ. Suppose that $\|\mathrm{diag}(\mathbf{\Sigma}_{T'} - \mathbf{I}_d)\|_{\mathrm{op},k} \lesssim \delta^2/\epsilon$. Then $\|\mathrm{diag}(\mathbf{\Sigma}_{T'} - \mathbf{I}_d)\|_{\mathrm{op},k} \lesssim \delta^2/\epsilon$*

2.4 Detecting Correlation in Subquadratic Time

We will use [Val15, Theorem 2.1] that can detect ρ-correlated coordinates in subquadratic time if there are not too many τ-correlated coordinates for $\tau \ll \rho$, say ρ^3.

Theorem 2.10 (Fast Correlation Detection [Val15])**.** *Let $\rho \in (0, 1)$ be strong correlation threshold and $\tau \in (0, 1)$ be margin threshold with $\rho > 12\tau$. Let T be a set of n vectors in $\mathbb{R}^d$ such that there are at most s τ-correlated coordinate pairs. Then, there is an algorithm that takes ρ, τ, T as input, and, with probability $1 - o(1)$, will output all ρ-coordinate pairs. Additionally, the runtime of the algorithm is $\left(sd^{0.62} + d^{1.62 + 2.4\frac{\log(4/\rho)}{\log(1/3\tau)}}\right) \mathrm{poly}(n, \log d, 1/\tau)$.*

The version above follows from [Val15, Theorem 2.1] for the binary vectors using standard reductions, for example, [Val15, Lemma 4.1]. For completeness, we state the guarantees of [Val15, Theorem 2.1] and show the reduction in Appendix A.

3 Robust Sparse Mean Estimation in Subquadratic Time

In this section, we explain our main technical contribution: a fast algorithm for robust sparse mean estimation under the stability condition.

Theorem 3.1 (Robust Sparse Mean Estimation in Subquadratic Time)**.** *Let c be a small enough absolute constant and C be a large enough absolute constant. Consider the corruption rate $\epsilon \in (0, \epsilon_0)$, where ϵ_0 is a small enough absolute constant. Let $k \in \mathbb{N}$ be the sparsity parameter and $q \in \mathbb{N}$ the correlation decay parameter with $q \geq 3$. Let T be an ϵ-corrupted version of a set S, where S satisfies $(C\epsilon, \delta, k')$-stability with respect to μ for $k' := \frac{(Ck)^q}{(\delta^2/\epsilon)^{q-1}}$ and $\delta^2/\epsilon \leq c$. Then there is a randomized algorithm (Algorithm 5) that takes as inputs T, ϵ, δ, k, and q and produces an estimate $\widehat{\mu}$ such that, with a probability at least $1 - 1/d^2$ over the randomness of the algorithm, we have the following guarantees:*

- *(Error) $\|\widehat{\mu} - \mu\|_{2,k} \lesssim \delta$.*
- *(Runtime) The algorithm runs in time at most $d^{1.62+\frac{3}{q}} \mathrm{poly}\left(|T|, \log d, k^q, 1/\epsilon^q\right)$.*

Organization In Section 3, we highlight the key technical challenges and our proposed fix. We record the guarantees of the key subroutines in Section 3.2. The proof of Theorem 3.1 is given in Section 3.3, and we finally show in Section 3.4 how Theorem 3.1 implies Theorem 1.5.

3.1 Algorithmic Blueprint of Theorem 3.1

To establish Theorem 3.1, we start with the following blueprint for robust sparse mean estimation, with the aim of implementing it in $o(d^2)$ time.

Algorithm 2 Algorithmic Blueprint

```
1: Compute ‖Σ_T − I_d‖_{Fr,k^2} approximately.
2: while ‖Σ_T − I_d‖_{Fr,k^2} is large do
3:     Let H be the corresponding coordinates with large Frobenius norm.
4:     Filter T using H in Lemma 2.7.
5: Output the sample mean μ_T.
```

The problem in implementing this blueprint naively is that Steps 1 and 3 in Algorithm 2 take d^2 time. However, these are the only two bottlenecks: Each filtering step takes only $d\mathrm{poly}(k, n) = d\mathrm{poly}(k/\epsilon)$ time, and there are at most $n = \mathrm{poly}(k/\epsilon)$ iterations. As we describe below, our goal in this section is to use Theorem 2.10 to speed up the Steps 1 and 3.

Usefulness of Fast Correlation Detection We will run Theorem 2.10 to identify the off-diagonal indices $(i, j) \in [d] \times [d]$ that are correlated. Recall that Theorem 2.10 takes two arguments ρ (the threshold for strong correlation) and τ (the margin threshold). Suppose we fix ρ to be small, roughly $\frac{\delta^2}{\epsilon \mathrm{poly}(k)}$. Then Theorem 2.10 reports back all ρ-correlated coordinate pairs. For the time to be subquadratic in d, we require the number of τ-correlated pairs to be small. Let $H_{\mathrm{pair}} \subset [d] \times [d]$

be the set of coordinate pairs in the output, and define H to be the set of all coordinates that appear in H_{pair}; $\text{Proj}(H_{\text{pair}})$ from Definition 2.2 to be more formal. Then, one of the following two cases must be true:

- $|H|$ is small ($\text{poly}(k/\epsilon)$): Since coordinates in $H^{\complement}$ have correlations at most ρ, then we know that $\|(\mathbf{\Sigma}_T - \mathbf{I})_{H^{\complement}}\|_{\text{Fr},k^2} \leq k\rho$, which can be made less than δ^2/ϵ for ρ small enough ($\rho \lesssim \delta^2/(k\epsilon)$). Thus, the sample mean on $H^{\complement}$ is a good estimate in $\ell_{2,k}$ norm (cf. Lemma 2.6). On H, we can use a dense mean estimation algorithm, which would be fast as $|H| = \text{poly}(k/\epsilon)$.
- $|H|$ is large: Since there are many coordinate pairs with correlation at least ρ, we can filter and iterate as follows: If we take any $H' \subset H$ of size k', then each row and column in $(\mathbf{\Sigma}_T - \mathbf{I})_{H'}$ has an entry larger than ρ (in absolute value), and thus $\|(\mathbf{\Sigma}_T - \mathbf{I})_{H'}\|_{\text{Fr}} \geq \rho\sqrt{k'}$. By taking k' large enough (larger than $\delta^4/(\rho^2\epsilon^2)$), the resulting quantity will be bigger than δ^2/ϵ, allowing us to filter if the inliers satisfy stability with parameter k' (cf. Lemma 2.7).

Thus, we can implement Steps 1 and 3 in Algorithm 2 fast using Theorem 2.10, so long as (i) inliers satisfy (ϵ, δ, k') stability, (ii) we choose $\rho^2 \asymp \frac{\delta^4}{\epsilon^2 k^2}$ and $k' \asymp \frac{\delta^4}{\epsilon^2\rho^2} \asymp k^2$, and (iii) *there are not too many τ-correlated pairs.*

Challenges in Using Fast Correlation Detection Suppose we set $\tau = \rho^q$ for some $q \geq 3$.[12] Looking at Theorem 2.10, we obtain a subquadratic time algorithm *only* if s, the number of τ-correlated pairs for $\tau := \rho^q$, is much smaller than $d^{1.38}$; In fact, we will impose s to be less than d so that it is not the dominant factor in the runtime. A priori, there is no reason for there to be at most d coordinate pairs (out of d^2 pairs) that are τ-correlated. Thus, we need a way to detect this situation and find an alternative way to make progress.

Proposed Solutions: Efficient Detection and Filtering We begin with the detection procedure. If there are $\Omega(d)$ many τ-correlated pairs, then a pair sampled uniformly at random has a probability of $\Omega(d^{-1})$ of being τ-correlated. Thus, if we check many random coordinate pairs, superlinear but subquadratic, then we can accurately guess s.

In particular, let U be the number of τ-correlated pairs that were observed out of m random pairs (sampled with replacement). Then, U is distributed roughly as $\text{Ber}(m, s/d^2)$. Binomial concentration implies that $s \lesssim (d^2U/m) + (\log d)(d^2/m)$ with probability at least $(1 - 1/d^2)$. Taking m to be $d^{1.5}$, we see that $s \lesssim \sqrt{d}U + \sqrt{d}\log d$. Thus, we obtain a fast (randomized) check to see if s is less than d than runs in $mn = d^{1.5}\text{poly}(k/\epsilon)$ time: simply check if $U \leq \sqrt{d}$. Thus, it remains to ensure that we can make progress when U is large, in particular, $\Omega(\sqrt{d})$.

Let H'_{pair} be the τ-correlated coordinates that were observed in the above procedure; $U = |H'_{\text{pair}}|$. Crucially, we are in the regime when $|H'_{\text{pair}}| \geq \sqrt{d}$. We want to use a small subset of the coordinates in H'_{pair} to filter outliers. Let H be k''-sized set of coordinates that appear in H'_{pair}; formally, $H := \text{Proj}_{k''}(H'_{\text{pair}})$. Thus, each row and column in $(\mathbf{\Sigma}_T - \mathbf{I})_H$ has an entry larger than τ in absolute value, implying that $\|(\mathbf{\Sigma}_T - \mathbf{I})_H\|_{\text{Fr}} \geq \tau\sqrt{k''}$. Therefore, we can use this H to filter using Lemma 2.7 as long as the original set is also (ϵ, δ, k'')-stable and $\sqrt{k''}\tau \gg \delta^2/\epsilon$, i.e., $k'' \asymp \frac{\delta^4}{\epsilon^2\tau^2} \asymp \frac{\delta^4}{\epsilon^2\rho^{2q}} \asymp \frac{\delta^4}{\epsilon^2(\delta^4/k\epsilon^2)^q} \asymp \frac{k^{2q}}{(\delta^4/\epsilon^2)^{q-1}}$.

Observe that we require the inliers to satisfy (ϵ, δ, k'')- stability for $k'' = \text{poly}(k^q/\epsilon^q)$. By Lemma 2.4, a set of $(k'')^2/\epsilon^2$ many i.i.d. points will satisfy this stability condition, giving us the sample complexity.

[12] We choose this parameterization because the runtime of Theorem 2.10 depends on $\log(1/\rho)/\log(1/\tau) = 1/q$.

3.2 Sparse Certificates and Filters in Subquadratic Time

We now give formal guarantees of the key procedures outlined above. First, consider the procedure that randomly samples the coordinates to estimate the number of weakly-correlated coordinates.

Lemma 3.2. *Let $T \subset \mathbb{R}^d$ be a multiset with covariance matrix Σ'. Let $m \in N$ be the sampling parameter. Let $J_* \subset [d] \times [d]$ be the off-diagonal coordinate pairs such that $|\mathbf{\Sigma}'_{i,j}| \geq \tau$. Then Algorithm 3 takes τ, m, and T as input and returns a set $J \subset J_*$ in time $O(m|T|)$ such that with probability $1 - 1/d^2$, $|J| \geq \frac{m|J_*|}{4d^2} - 16 \log d$.*

Algorithm 3 RANDOMLYCHECKCOORDINATES

1: Let $H_{\text{pair}} \subset [d] \times [d]$ of size m, with $(i,j) \in H_{\text{pair}}$ sampled i.i.d. from off-diagonal elements.
2: Let $J \leftarrow \{(i,j) \in H_{\text{pair}} : |\mathbf{\Sigma}'_{i,j}| \geq \tau\}$ for $\mathbf{\Sigma}' = \mathbf{\Sigma}_T$
3: **return** J.

Proof. To show correctness, we shall use the following concentration inequality for Binomials: If $X \sim \text{Ber}(n,p)$, then with probability $1-\delta$, $|\sqrt{X} - \sqrt{np}| \leq 2\sqrt{\log(1/\delta)}$. See, for example, [PW23, Equations (15.21) and (15.22)]. In particular, with probability $1 - 1/d^4$, $\sqrt{X} \geq \sqrt{np} - 4\sqrt{\log d}$, which implies $X \geq 0.25np - 16 \log d$.[13]

The probability that a single pair in H_{pair} has correlation of magnitude at least τ is exactly $(|J_*|/d(d-1))$, and thus $|J| \sim \text{Ber}(m, |J_*|/(d(d-1)))$. Therefore, applying the Binomial concentration, with probability $1 - 1/d^2$, it holds that $|J| \geq \frac{m|J_*|}{4d^2} - 16 \log d$. The claim about the runtime is immediate.

□

Combining Algorithm 3 with Theorem 2.10, Algorithm 4 either returns a small subset of coordinates with large Frobenius norm or indicates when all tiny subsets have small Frobenius norm.

Proposition 3.3 (Subroutine to Identify Corrupted Coordinates Algorithm 4)**.** *Suppose we are given a corrupted set T, Frobenius threshold $\kappa \in \mathbb{R}$, correlation threshold $\rho \in (0,1)$, margin threshold $\tau \in (0, \rho/12)$, sampling parameter $m \in \mathbb{N}$, and correlation count $s \in \mathbb{N}$. Suppose that each diagonal entry of $\mathbf{\Sigma}_T$ lies in $[1/2, 3/2]$ and $\frac{c_1 d^2}{m}\left(\frac{\kappa^2}{\tau^2} + \log d\right) \leq s$.*

Then Algorithm 4 takes as input $T, \kappa, \rho, \tau, m, s$ as input and satisfies the following with probability $1 - 1/\text{poly}(d)$:

- *It either outputs a set $H \subset [d]$ with $|H| \leq \frac{\kappa^2}{\tau^2}$ and $\| (\text{offdiag}(\mathbf{\Sigma}_T))_H \|_{\text{Fr}} \geq \kappa/2$.*
- *Else, it outputs "⊥". If it outputs "⊥", then $\|\text{offdiag}(\mathbf{\Sigma}_T)\|_{\text{Fr},k^2}$ is at most $2\kappa + 2\rho k$.*

Moreover, the algorithm runs in $O\Big(m + sd^{0.62} + d^{1.62 + 3\frac{\log(4/\rho)}{\log(1/3\tau)}}\Big)\text{poly}(n, \log d, 1/\tau)$ time.

Proof. Let $\mathbf{A}$ be the matrix with diagonal entries zero and off-diagonal entry (i,j) equal to $\frac{\mathbf{\Sigma}'_{i,j}}{\sqrt{\mathbf{\Sigma}'_{i,i}\mathbf{\Sigma}'_{j,j}}}$. Since diagonal entries of $\mathbf{\Sigma}_T$ lie in $[0.5, 1.5]$, the entries of $\mathbf{A}$ and $\text{offdiag}(\mathbf{\Sigma}_T) = \text{offdiag}(\mathbf{\Sigma}_T - \mathbf{I}_d)$ have the same signs and have the magnitude up to a factor of 2.

[13] We use that if $a, b, c \in \mathbb{R}_+$ then $a \geq b - c$ implies that $a^2 \geq b^2/4 - c^2$. The proof is as follows: if $b \geq 2c$ then $a \geq b/2$ and thus $a \geq b^2/4 \geq b^2/4 - c^2$; otherwise $b^2/4 \leq c^2$ and thus $a^2 \geq 0 \geq b^2/4 - c^2$ holds trivially.

Algorithm 4 Main Subroutine

Input: Frobenius threshold $\kappa \in \mathbb{R}_+$, a finite set $T \subset \mathbb{R}^d$ such that diagonal entries of Σ_T lie in $[1/2, 2]$, correlation threshold $\rho \in (0, 1)$, weak correlation threshold τ, sampling parameter $m \in \mathbb{N}$. We require the parameters to satisfy

$$\frac{c_1 d^2}{m}\left(\frac{\kappa^2}{\tau^2} + \log d\right) \leq s. \tag{3}$$

Output: With high probability, output either (i) a set $H \subset [d]$ with $|H| \leq k''$ and $\|(\text{offdiag}(\mathbf{\Sigma}_T)_H\|_{\text{Fr}} \geq 0.5\kappa$ or (ii) "$\perp$"; If it outputs "$\perp$", then $\|(\text{offdiag}(\mathbf{\Sigma}_T)_H\|_{\text{Fr}} \leq 2\kappa + 2k\rho$.

1: Let H'_{pair} be the output of Random Subsampling Algorithm (Algorithm 3) with m and τ
2: **if** $|H'_{\text{pair}}| \geq \kappa^2/\tau^2$ **then**
3:     Set $H_1 \leftarrow \text{Proj}_{\kappa^2/\tau^2}(H_{\text{proj}})$
4:     **return** H_1
5: **else**
6:     Let $H_{\text{pair}} \subseteq [d] \times [d]$ be the output of Theorem 2.10 with τ, ρ, s on T
7:     **if** $|H_{\text{pair}}| > \kappa^2/\rho^2$ **then**
8:         $H_3 \leftarrow \text{Proj}_{\kappa^2/\rho^2}(H_{\text{proj}})$
9:         **return** H_3
10:     **else**
11:         Let $H_2 \leftarrow \text{Proj}(H_{\text{pair}})$
12:         **if** $\|(\text{offdiag}(\mathbf{\Sigma}_S))_{H_2}\|_{\text{Fr}} \geq \kappa$ **then**
13:             **return** H_2
14:         **else**
15:             **return** "$\perp$"

Thus, at the cost of constant factor, we will upper bound $\|\mathbf{A}\|_{\text{Fr},k^2}$ and lower bound $\|(\mathbf{A})_H\|_{\text{Fr}}$ instead of dealing with $\text{offdiag}(\mathbf{\Sigma}_T - \mathbf{I}_d)$.

We will use the notation from Algorithm 4. Let s_* be the number of off-diagonal coordinates of $\mathbf{A}_*$ that have absolute value larger than τ. Our algorithm will be correct on the event when both Theorem 2.10 and Lemma 3.2 succeed, which happens with high probability. For the rest of this proof, we condition on both of these algorithms succeeding.

We first consider the case when $s_* > s$. Then, the lower bound on s in the statement (cf. (3)), coupled with Lemma 3.2, implies that

$$\begin{aligned} |H'_{\text{pair}}| &> \frac{m s_*}{4d^2} - 16\log d > \frac{ms}{4d^2} - 16\log d \\ &> \frac{m}{4d^2}\left(\frac{c_1 d^2}{m}\left(\frac{\kappa^2}{\tau^2} + \log d\right)\right) - 16\log d > \frac{c_1 \kappa^2}{2\tau^2}, \end{aligned}$$

where we use c_1 is large enough, say $c_1 \geq 100$. By definition of $H_1 = \text{Proj}_{\kappa^2/\tau^2}(H_{\text{proj}})$, there are at least $\lfloor \kappa^2/\tau^2 \rfloor$ entries in $(\mathbf{A})_{H_1}$ with absolute value at least τ, and thus $(\mathbf{A})_{H_1}$ has Frobenius norm at least $\sqrt{\kappa^2/\tau^2}\tau = \kappa$.

Consider the alternate case when $s_* < s$. If $|H'_{\text{pair}}|$ happens to be large, then the same argument as above implies the correctness and the runtime. Thus, in the rest of this proof, we consider the case when we enter Line 5. Since $s_* \leq s$, Theorem 2.10 runs in the desired time and finds all off-diagonal coordinate pairs of A, collected in H_{pair}, with entries bigger than ρ in the promised time. If H_{pair} has more than κ^2/ρ^2 entries, then by definition of $H_3 = \text{Proj}_{\kappa^2/\rho^2}(H_{\text{pair}})$, the same argument as above implies that the Frobenius norm of $(\mathbf{A})_{H_3}$ is at least κ.

If $|H_{\text{pair}}| < \kappa^2/\rho^2$, then H_2 is defined to contain all coordinates that appear in H_{pair}. If Line 12 succeeds, then the algorithm returns a subset of coordinates satisfying the desired conditions (small size and Frobenius norm larger than κ). Suppose the if condition is not satisfied (and we return "$\perp$"). We argue that $\|\mathbf{A}\|_{\text{Fr},k^2}$ is small enough. Write $\mathbf{A} = \mathbf{B} + \mathbf{B}'$, where $\mathbf{B}_{i,j}$ is non-zero only when $i \in H_2, j \in H_2$ with values equal to $\mathbf{A}_{i,j}$ and zero otherwise ($\mathbf{B}'$ is then defined to be $\mathbf{A} - \mathbf{B}$). By definition of H_2, we know each entry of $\mathbf{B}'$ is at most ρ in absolute value. By triangle inequality, $\|\mathbf{A}\|_{\text{Fr},k^2} \le \|\mathbf{B}\|_{\text{Fr},k^2} + \|\mathbf{B}'\|_{\text{Fr},k^2} = \|(\mathbf{A})_{H_2}\|_{\text{Fr},k^2} + \|\mathbf{B}'\|_{\text{Fr},k^2} \le \|(\mathbf{A})_{H_2}\|_{\text{Fr}} + k\rho \le \kappa + k\rho$. The accuracy guarantee follows by noting that the entries of $\mathbf{A}$ and $\mathbf{\Sigma} - \mathbf{I}$ are within a factor of 2. Finally, the runtime guarantees are immediate by Theorem 2.10 and Lemma 3.2. □

3.3 Proof of Theorem 3.1

The complete algorithm is given below:

Algorithm 5 Main Algorithm

Input: corruption rate $\epsilon \in (0, 1)$, stability parameter $\delta \in (0, 1)$, corrupted set $T \subset \mathbb{R}^d$, correlation-threshold $\rho \in (0, 1)$, correlation-decay $q \in \mathbb{N}$, sparsity $k \in \mathbb{N}$, sampling parameter $m \in \mathbb{N}$. We require T to be an ϵ-corrupted version of an $(C\epsilon, \delta, k')$ stable set with respect to μ for $k' := \frac{(Ck)^q}{(\delta^2/\epsilon)^{q-1}}$ and $\delta^2/\epsilon \le c$ for a small absolute constant $c > 0$.

Output: $\widehat{\mu} \in \mathbb{R}^d$ such that, with high probability, $\|\widehat{\mu} - \mu\|_{2,k} \lesssim \delta$.

1: $T' \leftarrow$ Filter T using Claim A.2
▷ Preprocessing along the diagonals to ensure Condition 2.8
2: $i \leftarrow 1$
3: $T_i \leftarrow T'$.
4: $H \leftarrow$ output of Proposition 3.3 with inputs: corrupted set T_i, Frobenius threshold $\kappa = C'\delta^2/\epsilon$ for a large constant C', correlation threshold $\rho = (\delta^2/\epsilon)/k$, margin threshold $\tau = (\rho/12)^q$, sampling parameter $m \asymp d(\kappa^2/\tau^2 + \log d)$, and correlation count $s = d$
▷ Algorithm 4
5: **while** $T_i \neq \emptyset$ and $H \neq$ "$\perp$" **do**
6:     Get the scores $f : T_i \to \mathbb{R}_+$ from Lemma 2.7 with inputs T_i, H, ϵ, and δ
7:     $T_{i+1} \leftarrow$ Filter T_i using the scores f similar to Algorithm 1
8:     $i \leftarrow i + 1$
9:     Update H as above
10: **return** the sample mean of T_i

We now present the proof of its correctness:

Proof of Theorem 3.1. The first step of Algorithm 5 is the fast preprocessing step from Claim A.2, which takes at most $\tilde{O}(dk^2|T|^2)$ time and removes not too many inliers with high probability In particular, the returned set T' has diagonal values in $[1/2, 2]$. In fact, Claim 2.9 implies that the subsequent sets T_i's will satisfy this property, at least until we remove more than $\Omega(\epsilon)$ fraction of points, which is the regime of interest anyway (cf. Theorem 2.5). Thus, Proposition 3.3 will be applicable.

We will follow the standard proof template of randomized filtering algorithm from Theorem 2.5, with the stopping condition provided by Proposition 3.3. The choice of parameters in Line 4 are such that whenever Proposition 3.3 does not return "$\perp$" (and T_i is 10ϵ-corruption of S), it returns a set H satisfying the guarantees of Lemma 2.7. To see this, observe that whenever Algorithm 4 outputs a subset $H \subset [d]$, then $|H| \leq \kappa^2/\tau^2$ and $\|(\mathbf{\Sigma}_{T_i} - \mathbf{I}_d)_H\|_{\text{Fr}} > \|(\text{offdiag}(\mathbf{\Sigma}_{T_i}))_H\|_{\text{Fr}} > \kappa/4 \gtrsim \delta^2/\epsilon$. Moreover, whenever Proposition 3.3 returns "$\perp$", then $\|(\mathbf{\Sigma}_{T_i} - \mathbf{I}_d)_H\|_{\text{Fr},k^2} \lesssim \kappa + k\rho \lesssim \delta^2/\epsilon$ since $\kappa \asymp \delta^2/\epsilon$ and $\rho \asymp \delta^2/\epsilon$. The choice of sparsity parameter in the stability of inliers is k', which is larger than κ^2/τ^2, because the choice of κ above and $\tau = (\rho/12)^q$ imply that κ^2/τ^2 is of the order $(\delta^2/\epsilon)/(c'\rho^q) \asymp (\delta^2/\epsilon)/(((c'\delta^2/\epsilon)/k)^q) \asymp \frac{k^q}{C^q(\delta^2/\epsilon)^{q-1}}$. Thus, the scores generated by H using Lemma 2.7 give more weights to outliers than inliers. Theorem 2.5 now implies that the final set T'', with probability at least 0.6, satisfies (i) T'' is an $O(\epsilon)$-contamination of S, (ii) T'' is an $O(\epsilon)$ contamination of T', where T' satisfies Condition 2.8, and (iii) $\|\text{offdiag}(\mathbf{\Sigma}_{T''} - \mathbf{I}_d)_H\|_{\text{Fr},k^2} \lesssim \delta^2/\epsilon$.

We now argue that $\|\mathbf{\Sigma}_{T''} - \mathbf{I}_d\|_{\text{op},k} \lesssim \delta^2/\epsilon$, and not just the off-diagonal terms. We follow the following inequalities using triangle inequality:

$$\begin{aligned}\|\mathbf{\Sigma}_{T''} - \mathbf{I}_d\|_{\text{op},k} &\leq \|\text{diag}\left(\mathbf{\Sigma}_{T''} - \mathbf{I}_d\right)\|_{\text{op},k} + \|\text{offdiag}\left(\mathbf{\Sigma}_{T''} - \mathbf{I}_d\right)\|_{\text{op},k} \\ &\lesssim \delta^2/\epsilon + \|\text{offdiag}\left(\mathbf{\Sigma}_{T''} - \mathbf{I}_d\right)\|_{\text{Fr},k^2} \\ &\qquad\qquad \text{(using Claim 2.9 and } T' \text{ satisfies Condition 2.8)} \\ &\lesssim \delta^2/\epsilon + \delta^2/\epsilon \lesssim \delta^2/\epsilon\,.\end{aligned}$$

Applying Lemma 2.6, the final output of the algorithm, sample mean of T'' will be $O(\delta)$ close to the true mean μ, implying the correctness of the procedure.

It remains to show the choice of the parameters m, s, and q lead to fast runtimes. We take $\tau = (\rho/12)^q$ and $s = d$. Finally, we take $m \asymp (d^2/s) \cdot (\kappa^2/\tau^2 + \log d) \leq (d \log d)(\kappa^2/\tau^2)$, which satisfies the parameter constraints in Proposition 3.3 (cf. (3)). Letting $n = |T|$, the resulting runtime of a single application of Proposition 3.3 is thus at most

$$\begin{aligned}A &= \left(m + sd^{0.62} + d^{1.62 + 3\frac{\log(4/\rho)}{\log(1/3\tau)}}\right)\text{poly}(n, \log d, 1/\tau) \\ &\leq \left(d^{1.62 + 3\frac{\log(4/\rho)}{\log((\rho/4)^q)}}\right)\text{poly}(n, \log d, 1/\rho^q) \\ &\leq \left(d^{1.62 + \frac{3}{q}}\right)\text{poly}(n, \log d, k^q, 1/\epsilon^q)\,.\end{aligned}$$

Since there are at most n iterations, the claim on the total runtime follows. Finally, the success probability of the algorithm can be boosted from 0.55 to $1 - \delta$ by repeating the algorithm $\ell = \log(1/\delta)$ times and outputting the estimate that is $O(\delta)$-close (in the $\|\cdot\|_{2,k}$ norm) to the majority of the ℓ estimates. It only adds a multiplicative factor of ℓ and an additive term of $d\ell^2$ to the runtime. □

3.4 Proof of Theorem 1.5

We now explain how Theorem 3.1 implies Theorem 1.5.

Proof of Theorem 1.5. First, observe that by applying Proposition 2.1, the estimation guarantee of Theorem 3.1 can be translated from $\|\cdot\|_{2,k}$ norm into $\|\cdot\|_2$ norm by hard-thresholding the estimator. We now turn our attention to the sample complexity. Let S be a set of n i.i.d. samples from $\mathcal{N}(\mu, I)$. If $n \geq ((k')^2 \log d)/\epsilon^2$, then Lemma 2.4 implies that S is (ϵ, δ, k') stable with respect to μ for $\delta \lesssim \epsilon\sqrt{\log(1/\epsilon)}$. Plugging in the value of δ and $k' \leq \frac{(O(k))^q}{\epsilon^{q-1}}$ in Theorem 3.1 yields a sample complexity of at most $n \lesssim \frac{(k')^2 \log d}{\epsilon^2} \lesssim (O(k/\epsilon))^{2q} \log d$. The claim on the runtime in Theorem 1.5 follows from Theorem 3.1 along with the bound on the sample complexity. □

4 Robust Sparse PCA

In this section, we show that the ideas from fast correlation detection can also be useful for robust sparse PCA. Our main result in this section is the result below:

Theorem 1.6 (Robust Sparse PCA in Subquadratic Time)**.** *Let T be an ϵ-corrupted set of samples from $\mathcal{N}(0, \mathbf{I}_d + \eta vv^\top)$ for $\eta \in (0,1)$ and a k-sparse unit vector. There is a randomized algorithm that takes as input the corrupted set T, contamination rate ϵ, sparsity $k \in \mathbb{N}$, spike η, and a parameter $q \in \mathbb{N}$, and produces an estimate $\widehat{v}$ such that*

- *(Sample Complexity and Error) If $n \gtrsim \mathrm{poly}((k^q \log d)/\epsilon^q)$, then $\|\widehat{v}\widehat{v}^\top - vv^\top\|_{\mathrm{Fr}} \lesssim \sqrt{\epsilon \log(1/\epsilon)/\eta}$ with probability at least $1 - \frac{1}{\mathrm{poly}(d)}$.*
- *(Runtime) The algorithm runs in time at most $d^{1.62 + \frac{3}{q}} \mathrm{poly}(n)$.*

The result above provides the first subquadratic time algorithm for robust sparse PCA that has error independent of d and k. Similar to the literature on robust sparse mean estimation, existing algorithms for robust sparse PCA with similar dimension-independent error took $\Omega(d^2)$ time. However, the error guarantee of Theorem 1.6 is not optimal: the error guarantee above is $(\epsilon \log(1/\epsilon))/\eta$, same as [CDKGGS22], as opposed to the near-optimal error of $(\epsilon^2 \mathrm{polylog}(1/\epsilon))/\eta$ in [BDLS17; DKKPS19].

We remark that Theorem 1.6 does not follow directly from Theorem 1.5. In particular, a standard reduction relies on the fact that the mean of the random variable $\mathbf{Y} - \mathbf{I}_d$ is exactly $\eta vv^\top$ for $\mathbf{Y} = xx^\top$, where $x \sim \mathcal{N}(0, \mathbf{I} + \eta vv^\top)$, i.e., (robust) sparse PCA reduces to (robust) sparse mean estimation. However, $\mathbf{Y}$ is a d^2-dimensional object and thus a naive application of Theorem 1.5 would yield a super-quadratic runtime, in fact, $(d^2)^{1.62} = \Omega(d^3)$. Moreover, the covariance of $\mathbf{Y}$ is not isotropic and thus it is unclear if samples from Y would satisfy the stability condition from Definition 2.3. In the rest of this section, we show that a more whitebox analysis of the main ideas from Theorem 1.5, in particular, Proposition 3.3 can yield a subquadratic runtime.

Organization This section is organized as follows: We list the deterministic conditions for robust PCA in Section 4.1. Section 4.2 contains the results pertaining to the certificate lemma, filtering, and dense estimation algorithm for sparse PCA. Finally, the algorithm and its proof under a generic stability condition is given in Section 4.3. Finally, we present the proof of Theorem 1.6 in Section 4.4.

4.1 Deterministic Conditions

For a set T, we use the notation $\overline{\mathbf{\Sigma}}_T$ to denote the second moment matrix $\mathbb{E}_T[xx^\top]$—not to be confused with the covariance matrix $\mathbf{\Sigma}_T$.

Definition 4.1 (Stability Condition for PCA)**.** For the contamination rate $\epsilon \in (0, 1/2)$, error parameter $\gamma \geq \epsilon$, spike strength $\eta \in \mathbb{R}_+$, and sparsity $k \in \mathbb{N}$, we say a set $S \subset \mathbb{R}^d$ is $(\epsilon, \gamma, k, \eta)$-pca-stable with respect to a k-sparse unit vector $v \in \mathbb{R}^d$ and spike strength η if

1. For any subset $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$: $\left\| \overline{\mathbf{\Sigma}}_{S'} - (\mathbf{I}_d + \eta vv^\top) \right\|_{\mathrm{Fr}, k^2} \leq \gamma$
2. For all subsets $H \subset [d]$ with $|H| \leq k$, the set $\{(x)_H : x \in S\}$ is (ϵ, γ)-covariance stable in the sense of [DK23, Definition 4.5] (with respected to an appropriate flattening of $\mathbf{I}_{|H|}$).

The second stability condition allows us to perform covariance estimation in Frobenius norm (and hence stronger than principal component analysis) but only when the support is known.

The following result lists the sample complexity of Definition 4.1.

Lemma 4.2 (Sample Complexity). *Let S be a set of n i.i.d. samples from $\mathcal{N}(0, \mathbf{I} + \eta vv^\top)$ for $\eta \in (0,1)$ and a k-sparse unit vector v. Then if $n \gtrsim \mathrm{poly}(k, \log(d/\delta), 1/\epsilon)$, then S is $(\epsilon, \gamma, k, \eta)$-pca-stable with respect to v and η for $\gamma = O(\epsilon \log(1/\epsilon))$.*

Proof. We sketch the argument here. The first condition is identical to [CDKGGS22, Definition 2.2] and [CDKGGS22, Lemma 4.2] establishes a sample complexity of $\frac{k^2 \log(d/\delta)}{\epsilon^2}$.

For the second condition, we do an admittedly loose analysis. Fixing a subset H, it amounts to showing a stability condition for a $|H|$-dimensional Gaussian distribution, for which [DK23, Proposition 4.2] establishes a sample complexity of $\mathrm{poly}(k \log(1/\delta)/\epsilon)$ for the set to be (ϵ, γ)-covariance stable for $\gamma = O(\epsilon \log(1/\epsilon))$ with probability $1 - \delta$.[14] Since there are at most d^k possible choices of H, a union bound shows that the second condition in Definition 4.1 holds with probability $1 - \delta$ with sample complexity $\mathrm{poly}(k \log(d^k/\delta)/\epsilon) = \mathrm{poly}(k \log(d/\delta)/\epsilon)$. □

4.2 Filters, Certificates, and Dense Estimation

Dense Estimation Algorithm We shall use the following algorithm to estimate the spike when we have identified the support of the spike.

Lemma 4.3 (Dense Covariance Estimation Algorithm; Implication of [DK23, Theorem 4.6]). *Let $\epsilon \in (0, \epsilon_0)$ for a small absolute constant ϵ_0. Let T be an ϵ-corrupted version of S, where S is $(\epsilon, \gamma, k, \eta)$-pca-stable with respect to v (cf. Definition 4.1) and $\eta \in (0,1)$. Let H be a k-sparse subset of $[d]$. There exists an algorithm $\mathcal{A}$ that takes as inputs corrupted set T, contamination rate ϵ, error parameter γ, spike η, and the sparse support set H and outputs a k-sparse vector u, supported on H, such that $\left\| \left(\mathbf{I}_d + \eta uu^\top\right)_H - \left(\mathbf{I}_d + \eta vv^\top\right)_H \right\| \lesssim \gamma$. Moreover, the algorithm runs in time $d\mathrm{poly}(k, |T|/\epsilon)$.*

Proof. Let $T' \subset \mathbb{R}^k$ and $S' \subset \mathbb{R}^k$ denote the projections of T and S, respectively, on H. Let $\mathbf{\Sigma}' = (\mathbf{I}_d + \eta vv^\top)_H = \mathbf{I}_{|H|} + \eta' v' v'^\top$ for $v' = (v)_H / \|(v)_H\|_2$ and $\eta' = \eta \|(v)_H\|_2^2$.

[DK23, Theorem 4.6] gives an estimate $\widehat{\mathbf{\Sigma}}$ such that $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}'\|_{\mathrm{Fr}} \lesssim \gamma \|\mathbf{\Sigma}'\|_{\mathrm{op}} \lesssim \gamma$, which can be calculated in time $d\mathrm{poly}(k|T|/\epsilon)$; here $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of a matrix. That is, $\|(\widehat{\mathbf{\Sigma}} - \mathbf{I}) - \eta' v' v'^\top\|_{\mathrm{Fr}} \lesssim \gamma$. Letting $\mathbf{A}$ be the best one-rank approximation of $(\widehat{\mathbf{\Sigma}} - \mathbf{I})$ in the Frobenius norm, we see that

$$\|\mathbf{A} - \eta' v' v'^\top\|_{\mathrm{Fr}} \le \|\mathbf{A} - (\widehat{\mathbf{\Sigma}} - \mathbf{I})\|_{\mathrm{Fr}} + \|(\widehat{\mathbf{\Sigma}} - \mathbf{I}) - \eta' v' v'^\top\|_{\mathrm{Fr}} \le 2\|(\widehat{\mathbf{\Sigma}} - \mathbf{I}) - \eta' v' v'^\top\|_{\mathrm{Fr}} \lesssim \gamma,$$

where the second inequality follows from the fact that $\mathbf{A}$ is the best rank-one approximation. Moreover, the symmetry of $\mathbf{A}$ implies that $\mathbf{A}$ must be of the form $\lambda' u' u'^\top$ for a unit vector u' and $\lambda' \in \mathbb{R}$. We thus obtain

$$\|\lambda' u' u'^\top - \eta' v' v'^\top\|_{\mathrm{Fr}} \lesssim \gamma\,.$$

Consequently, $|\lambda' - \eta'| \lesssim \gamma$ by Weyl's inequality. If $\lambda' \le 0$, we set $u = 0$. Otherwise, we set $u = \sqrt{\lambda'/\eta} u'$. We now calculate the approximation error. If $\lambda' \ge 0$, then the resulting error in our estimate is $\left\| \left(\mathbf{I}_d + \eta uu^\top\right)_H - \left(\mathbf{I}_d + \eta vv^\top\right)_H \right\|_{\mathrm{Fr}} = \left\| \lambda' uu^\top - \eta' v' v'^\top \right\|_{\mathrm{Fr}} \lesssim \gamma$ by the approximation guarantee. If $\lambda' < 0$, then η' must also be $O(\gamma)$ because $\eta' \le \lambda' + |\lambda' - \eta'| \lesssim \gamma$. In this case, the approximation error is η', which is also $O(\gamma)$ as required. □

[14] Although the proof of [DK23, Proposition 4.2] does not explicitly write the dependence on δ, but it is immediate from the proof.

Certificate Lemma for Sparse PCA The following result shows that if the covariance matrix looks roughly isotropic on a subset of coordinates $H^{\complement}$, then the spike vector v places most of its mass on the complement, H. The benefit of this stopping condition is that it does not depend on the unknown spike vector v.

Lemma 4.4 (Sparse Certificate Lemma for PCA)**.** *Let T be an ϵ-corrupted version of S, where S is $(\epsilon, \gamma, k, \eta)$-pca-stable with respect to a k-sparse unit vector $v \in \mathbb{R}^d$ and spike strength $\eta \in (0,1)$. Let $H \subset [d]$ be such that $\|(\overline{\mathbf{\Sigma}}_T - \mathbf{I})_{H^{\complement}}\|_{\mathrm{op},k} = O(\gamma)$, then $\|vv^\top - (v)_H(v)_H^\top\|_{\mathrm{Fr}}^2 = O\left((\gamma/\eta) + (\gamma/\eta)^2\right)$.*

Proof. Let z be the unit vector along $(v)_{H^{\complement}}$, which is at most k-sparse because v is k-sparse. Since z is supported on $H^{\complement}$, the assumption on $\overline{\mathbf{\Sigma}}_T$ and $H^{\complement}$ implies that $\left|z^\top \left(\overline{\mathbf{\Sigma}}_T - \mathbf{I}\right) z\right| \lesssim \gamma$, which further gives us that $z^\top \overline{\mathbf{\Sigma}}_T z = 1 \pm O(\gamma)$. Using the stability of S, we see $z^\top \overline{\mathbf{\Sigma}}_T z \geq (1-\epsilon) z^\top \overline{\mathbf{\Sigma}}_{S\cap T} z \geq (1-\epsilon)\left(z^\top \left(\mathbf{I} + \eta vv^\top\right) z - \gamma\right)$. Defining $\eta' := \eta (v^\top z)^2 = \eta \|v_{H^{\complement}}\|_2^2$, we have that $z^\top \overline{\mathbf{\Sigma}}_T z \geq (1-\epsilon)(1 + \eta' - \gamma)$. Combining this with the aforementioned upper bounds on $z^\top \overline{\mathbf{\Sigma}}_T z$, we obtain that $(1-\epsilon)(1+\eta'-\gamma) \leq 1 + O(\gamma)$. Thus, $\eta' = O(\gamma + \epsilon) = O(\gamma)$ by using $\epsilon \leq \gamma$. Therefore, $\|(v)_{H^{\complement}}\|_2^2 = \eta'/\eta \lesssim \gamma/\eta$. Finally, the triangle inequality implies that

$$\|vv^\top - (v)_H(v)_H\|_{\mathrm{Fr}} \leq \|(v)_{H^{\complement}}\|_2^2 + 2\|(v)_{H^{\complement}}\|_2 \lesssim \max\left(\sqrt{(\gamma/\eta)}, (\gamma/\eta)\right).$$

□

Filter for Sparse PCA We use the following result to filter outliers when we identify a small set of coordinates with a large variance.

Lemma 4.5 (Sparse PCA Filter)**.** *Let $\epsilon \in (0, \epsilon_0)$ for a small absolute constant ϵ_0 and let C be a large enough absolute constant. Let T be an ϵ-corrupted version of S, where S is $(\epsilon, \gamma, k, \eta)$-stable with respect to v (cf. Definition 4.1) and $\eta \in (0,1)$. Let $H \subset [d]$ be such $\left\|(\mathbf{\Sigma}_T - \mathbf{I} - \eta vv^\top)_H\right\|_{\mathrm{Fr}} = \lambda$ for $\lambda \geq 8C\gamma$ and $|H| \leq k$.*

Then there exists an algorithm $\mathcal{A}$ that takes T, H, ϵ, γ, and η and returns a score mapping $f : T \to \mathbb{R}_+$ such that the sum of inliers' scores is less than outliers': $\sum_{x\in S\cap T} f(x) \leq \sum_{x \in T\setminus S} f(x)$ and $\max_{x\in T} f(x) > 0$. Moreover, the algorithm runs in time $d\mathrm{poly}(k, |T|)$.

We give the proof of the result above in Appendix B.1.

Preprocessing Condition for Sparse PCA Similar to robust sparse mean estimation, our algorithm tracks the contribution to the diagonal terms and off-diagonal terms separately. The following conditions mirrors Condition 2.8.

Condition 4.6 (Idealistic Condition for PCA)**.** Let T be an ϵ-corrupted version of S, where S is $(\epsilon, \gamma, k, \eta)$-stable with respect to a k-sparse unit vector v and spike strength η (cf. Definition 4.1). We have an $H_1 \subset [d]$ and $|H_1| \lesssim k^2$ such that T satisfies $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_T - \mathbf{I}_d)_{H_1^{\complement}}\|_{\mathrm{Fr},k^2} \lesssim \gamma$.

The next result gives an efficient algorithm to ensure the above condition:

Claim 4.7. *Let $\epsilon \in (0, \epsilon_0)$ and $\gamma \in (0, \gamma_0)$ for small constants $\epsilon_0 \in (0, 1/2)$, $\gamma_0 \in (0,1)$. Let sparsity $k \in \mathbb{N}$. Let C be a large enough constant and T be an ϵ-corrupted set S where S is $(C\epsilon, \gamma, k', \eta)$-pca-stable with respect to an unknown k-sparse unit vector $v \in \mathbb{R}^d$, $\eta \in (0,1)$, $\gamma \geq \epsilon$, and $k' = C'k^2$ for a large enough constant $C' > 0$. There is a randomized algorithm $\mathcal{A}$ that takes as input the corrupted set T, contamination rate ϵ, sparsity $k \in \mathbb{N}$, and a parameter $q \in \mathbb{N}$, and returns a set $T' \subset T$ and $H_1 \subset [d]$ in time $O(d\mathrm{poly}(k, |T|))$ such that with probability* 0.9

1. T' *is an* $O(\epsilon)$*-contamination of* S.

2. *Each diagonal entry of* $\overline{\mathbf{\Sigma}}_{T'} \in [1/2, 4]$.

3. T' *and* H_1 *satisfy* Condition 4.6, *i.e.,* $|H_1| \lesssim k^2$ *and* $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_{T'} - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{Fr},k^2} \lesssim \gamma$.

The proof of this simple claim is given in Appendix B.2.

Similar to Claim 2.9, the next result shows that all large subsets of a set satisfying Claim 4.7 are close to identity in the sparse operator norm.

Claim 4.8. *Let* C *be a large enough constant* $C > 0$ *and* $k, k' \in \mathbb{N}$. *Let* $T'' \subset T'$ *be two* $O(\epsilon)$*-contamination of* S *such that* S *is an* $(C\epsilon, \gamma, k', \eta)$*-pca-stable with respect to* v *and* $\eta \in (0, 1)$. *Let* $H \subset [d]$ *be a small subset* $|H| \le k$ *such that* $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_{T'} - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{Fr},k^2} \lesssim \gamma$. *Then* $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_{T''} - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{op},k} \lesssim \gamma$.

4.3 Subquadratic Time Algorithm For Sparse PCA

We now establish the main technical result of this section:

Theorem 4.9 (Subquadratic Time Algorithm for Robust Sparse PCA under Stability)**.** *Let* $\epsilon \in (0, \epsilon_0)$ *and* $\gamma \in (0, \gamma_0)$ *for small constants* $\epsilon_0 \in (0, 1/2)$, $\gamma_0 \in (0, 1)$. *Let* C *be a large enough constant and* $k \in \mathbb{N}$ *be the sparsity parameter and* $q \in \mathbb{N}$ *the correlation decay parameter with* $q \ge 3$. *Let* T *be an* ϵ*-corrupted set* S *where* S *is* $(C\epsilon, \gamma, Ck', \eta)$*-pca-stable with respect to an unknown* k*-sparse unit vector* $v \in \mathbb{R}^d$, $\eta \in (0, 1)$, *and* $\gamma \ge \epsilon$ *for* $k' = \left(\frac{C^q k^{2q}}{\gamma^{2q-2}}\right)$.

There is a randomized algorithm $\mathcal{A}$ *that takes as input the corrupted set* T, *contamination rate* ϵ, *stability parameter* γ, *sparsity* $k \in \mathbb{N}$, *and a parameter* $q \in \mathbb{N}$, *and produces an estimate* $\widehat{v}$ *such that*

- *(Error) Then* $\|\widehat{v}\widehat{v}^\top - vv^\top\|_{\mathrm{Fr}} \lesssim \sqrt{\frac{\gamma}{\eta}}$ *with high probability over the randomness of the samples and the algorithm.*
- *(Runtime) The algorithm runs in time at most* $d^{1.62 + \frac{3}{q}} \mathrm{poly}\left(|T| \log d, (k/\epsilon)^q\right)$.

The complete algorithm is given below: We now present the proof of its correctness:

Proof of Theorem 4.9. We can assume that $\gamma \le \eta$, otherwise we can simply output any unit vector, whose error will be at most $O(1) = O(\sqrt{\gamma/\eta})$.

We will use the same template of filtering-based algorithms from Theorem 2.5, with the filtering subroutines provided by Lemma 4.5.

The first step of Algorithm 6 is the preprocessing step from Claim 4.7, which takes at most $\tilde{O}(d\mathrm{poly}(k, |T|))$ time and removes not too many inliers with high probability. In particular, the diagonal entries of $\overline{\mathbf{\Sigma}}_{T'}$ lie in $[1/2, 4]$; Moreover, all the subsequent sets T_i's will continue to satisfy this property by Claim 4.8 as long as we have removed at most $O(\epsilon)$ fraction of points (since γ is small enough). We shall use the output H_1 generated by Claim 4.7 in the end.

We will use a fine-grained result from Proposition 3.3. Observe that Proposition 3.3 can be used not just for the covariance $\mathbf{\Sigma}_T$ but also for $\overline{\mathbf{\Sigma}}_T$, i.e., without centering, which is what we will use in this proof. We shall invoke Proposition 3.3 to identify all the coordinate pairs with correlation larger than $\rho = \gamma/k$. We then set $\tau = (\rho/12)^q$ and use the Frobenius threshold $\kappa = C'\gamma$ for a constant $C > 100$. Defining $k' := \kappa^2/\tau^2$ and $k'' := \kappa^2/\rho^2$, we note that $k' \ge k'' = C'\gamma^2/(\gamma/k)^2 = C'k^2 \ge 2k^2$. We will show that the set H allows us to filter points using Lemma 4.3. The proof of Proposition 3.3 reveals that it returns H such that either

Algorithm 6 Robust Sparse PCA Algorithm

Input: corruption rate $\epsilon \in (0,1)$, stability parameter $\gamma \in (0,1)$, corrupted set $T \subset \mathbb{R}^d$, correlation-threshold $\rho \in (0,1)$, correlation decay $q \in \mathbb{N}$, sparsity $k \in \mathbb{N}$, spike strength $\eta \in (0,1)$, sampling parameter $m \in \mathbb{N}$. We require T to be an ϵ-corrupted version of an $(C\epsilon, \delta, Ck', \eta)$ stable set with respect to μ for $k' := \left(\frac{C^q k^{2q}}{\gamma^{2q-2}}\right)$ and $\gamma \ll 1$.

Output: $\widehat{v} \in \mathbb{R}^d$ such that, with high probability, $\|\widehat{v}\widehat{v} - vv^\top\|_{\mathrm{Fr}} \lesssim \gamma$.

$H_1, T' \leftarrow$ be the outputs of Claim 4.7 on T

$\triangleright$ T' is an $O(\epsilon)$-contamination of S, $0.5\mathbf{I}_d \preceq \mathrm{diag}(\overline{\mathbf{\Sigma}}) \preceq 2\mathbf{I}_d$, $|H_1| \lesssim k^2$ and $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_T - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{Fr},k^2} \lesssim \sqrt{\gamma/\eta}$

$i \leftarrow 1$

$T_i \leftarrow T'$.

$H \leftarrow$ output of Proposition 3.3 with inputs: corrupted set T_i, Frobenius threshold $\kappa = \Theta(\gamma)$, correlation threshold $\rho = \gamma/k$, margin threshold $\tau = (\rho/12)^q$, sampling parameter $m \asymp d(\kappa^2/\tau^2 + \log d)$, and correlation count $s = d$

$\triangleright$ Algorithm 4

while $T_i \neq \emptyset$ and $H \neq$ "$\perp$" and $|H| \leq \kappa^2/\rho^2$ **do**

  Get the scores $f : T_i \to \mathbb{R}_+$ from Lemma 4.5 with inputs T_i, H, ϵ, and δ

  $T_{i+1} \leftarrow$ Filter T_i using the scores f.

  $i \leftarrow i + 1$

  Update H as above

Let $H_{\mathrm{end}} \leftarrow H \cup H_1$

$u \leftarrow$ output of Lemma 4.3 with inputs T_i, ϵ, γ, η, and H_{end}

(Case I) $|H| = k' = \kappa^2/\tau^2$ and for each coordinate in $i \in H$, there exists a $j \in H$ such that (i,j) is τ-correlated

Let T'' be the current iterate of the corrupted data set. Since $\eta vv^\top$ is a k^2-sparse matrix and $\mathrm{offdiag}\left(\overline{\mathbf{\Sigma}}_T - \mathbf{I}_d\right)$ has at least k' entries with magnitude at least $\Theta(\tau)$[15], their difference must have at least $k' - k^2 \geq k'/2 = \kappa^2/\tau^2$ entries with magnitude $\Theta(\tau)$. Thus, the sparse Frobenius norm of their difference must be large, i.e.,

$$\left\|\left(\overline{\mathbf{\Sigma}}_{T''} - \mathbf{I}_d - \eta vv^\top\right)_H\right\|_{\mathrm{Fr}} \gtrsim \kappa \gtrsim \gamma\,.$$

Given such an H, we filter outliers using Lemma 4.3.

(Case II) $|H| \geq \kappa^2/\rho^2$ and for each coordinate in $i \in H$, there exists a $j \in H$ such that (i,j) is ρ-correlated

By the same argument as above, the matrix $\overline{\mathbf{\Sigma}}_{T''} - \mathbf{I}_d - \eta vv^\top$ has at least $k'' - k^2 \geq k''/2 = \kappa^2/(2\rho^2)$ entries with magnitude $\Theta(\rho)$. Thus, $\left\|\left(\overline{\mathbf{\Sigma}}_{T''} - \mathbf{I}_d - \eta vv^\top\right)_H\right\|_{\mathrm{Fr}} \gtrsim \kappa \gtrsim \gamma$. Again, we filter outliers using Lemma 4.3.

(Case III) $|H| \leq \kappa^2/\rho^2$ and no coordinate pair in $H^\complement$ is ρ-correlated.

Let T'' be the current iterate of the corrupted data set. Since $H \subset H_{\mathrm{end}}$ and the coordinates in $H^\complement$ are at most ρ-correlated, it follows that the coordinates in $H_{\mathrm{end}}^\complement$ are

[15] Recall that each diagonal entry of $\overline{\mathbf{\Sigma}}_{T''}$ is $\Theta(1)$ and thus τ-correlation implies that the corresponding entry in $\overline{\mathbf{\Sigma}}_{T''}$ is also $\Theta(\tau)$.

also at most ρ-correlated.

$$\left\|\operatorname{offdiag}\left(\left(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d\right)_{H_{\mathrm{end}}^{\complement}}\right)\right\|_{\mathrm{Fr},k^2} \le \left\|\operatorname{offdiag}\left(\left(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d\right)_{H^{\complement}}\right)\right\|_{\mathrm{Fr},k^2} \lesssim \rho k \lesssim \gamma\,. \tag{4}$$

We now want to combine the above guarantee on the closeness along the offdiagonals with the guarantee on the closeness along the diagonals from H_1. In particular, H_{end} satisfies that

$$\left\|\operatorname{diag}\left(\left(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d\right)_{H_{\mathrm{end}}^{\complement}}\right)\right\|_{\mathrm{op},k} \le \left\|\operatorname{diag}\left(\left(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d\right)_{H_1^{\complement}}\right)\right\|_{\mathrm{op},k} \lesssim \gamma\,, \tag{5}$$

where T'' has small sparse operator norm because it close to the preprocessed set T' and thus Claim 4.8 is applicable. Combining (4) and (5), we obtain

$$\left\|(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d)_{H_{\mathrm{end}}^{\complement}}\right\|_{\mathrm{op},k} \le \left\|\operatorname{diag}\left(\left(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d\right)_{H_{\mathrm{end}}^{\complement}}\right)\right\|_{\mathrm{op},k} + \left\|\operatorname{offdiag}\left(\left(\overline{\boldsymbol{\Sigma}}_{T''}-\mathbf{I}_d\right)_{H_{\mathrm{end}}^{\complement}}\right)\right\|_{\mathrm{op},k} \lesssim \gamma\,. \tag{6}$$

The size of H_{end} is at most $O(k^2)+k' \le 2k'$. Since S satisfies stability with $2k'$, Lemma 4.4 then implies that the spike v is mostly contained in H_{end}. By invoking the dense PCA algorithm on H_{end}, Lemma 4.3 estimates the spike $(v)_{H_{\mathrm{end}}}(v)_{H_{\mathrm{end}}}^\top$ with $uu^\top$, with u supported on H_{end}. Combining, we obtain:

$$\begin{aligned}\left\|vv^\top - uu^\top\right\|_{\mathrm{Fr}} &\le \left\|vv^\top - (v)_{H_{\mathrm{end}}}(v)_{H_{\mathrm{end}}}^\top\right\|_{\mathrm{Fr}} + \left\|uu^\top - (v)_{H_{\mathrm{end}}}(v)_{H_{\mathrm{end}}}^\top\right\|_{\mathrm{Fr}} \\ &\lesssim \sqrt{\gamma/\eta}+\gamma/\eta \lesssim \sqrt{\gamma/\eta},\end{aligned}$$

where the first term is bounded using Lemma 4.4 with (6) and the second term is bounded using Lemma 4.3.

(Case IV) $H =$ "$\perp$".

The same argument as the previous case holds.

Thus, the output of the algorithm is $O(\sqrt{\gamma/\eta})$ close as required. It remains to show the choice of the parameters m, s, and q lead to the claimed runtime. We take $\tau = (\rho/12)^q$ and $s = d$. Finally, we take $m \asymp (d^2/s)\cdot(\kappa^2/\tau^2+\log d) \le (d\log d)(\kappa^2/\tau^2)$, which satisfies the parameter constraints in Proposition 3.3 (cf. (3)). Letting $n = |T|$, the resulting runtime of a single application of Proposition 3.3 is thus at most

$$\begin{aligned} A &= \left(m + sd^{0.62} + d^{1.62+3\frac{\log(4/\rho)}{\log(1/3\tau)}}\right)\operatorname{poly}(n,\log d, 1/\tau) \\ &\le \left(d^{1.62+3\frac{\log(4/\rho)}{\log((\rho/4)^q)}}\right)\operatorname{poly}(n,\log d,1/\rho^q) \\ &\le \left(d^{1.62+\frac{3}{q}}\right)\operatorname{poly}(n,\log d,k^q,1/\epsilon^q)\,. \end{aligned}$$

As the iteration count is bounded by n, the total runtime is at most nA. □

4.4 Proof of Theorem 1.6

Proof of Theorem 1.6 using Theorem 4.9. By Theorem 4.9, it suffices to establish that a set of n i.i.d. samples from $\mathcal{N}(0,\mathbf{I}+\eta vv^\top)$ for a k-sparse unit vector v, with high probability, is $(\epsilon,\gamma,k',\eta)$-pca-stable for $\gamma \lesssim \epsilon\log(1/\epsilon)$ and $k' := (\frac{C^q k^{2q}}{\gamma^{2q-2}})$, where C is a large absolute constant. Lemma 4.2 gives a bound on the sample complexity, stating that it suffices to take $n = \operatorname{poly}(k',\log(d),1/\epsilon)$ many samples, thus establishing Theorem 1.6. □

5 Discussion

In this article, we presented the first subquadratic time algorithm for robust sparse mean estimation. We now discuss some related open problems and avenues for improvement. First, the sample complexity of Theorem 1.5 is polynomially larger than $k^2(\log d)/\epsilon^2$—the sample complexity of existing (quadratic runtime) algorithms.[16] Bridging this gap is an important problem. Second, Theorem 1.5 is specific to isotropic structured distributions such as Gaussians (more broadly, isotropic distributions P with mean μ that satisfy $\mathrm{Var}_{x\sim P}((x-\mu)^\top \mathbf{A}(x-\mu)) \lesssim \|\mathbf{A}\|_{\mathrm{Fr}}^2$ for symmetric matrices $\mathbf{A}$). Indeed, both [DKKPS19; CDKGGS22] rely on the isotropy and the aforementioned variance structure of quadratic polynomials to avoid solving SDPs that appear in [BDLS17]. Developing a similarly fast algorithm for unstructured distributions (with sparse means) is still open to our knowledge. Third, because Theorem 1.5 relies on [Val15], which in turn relies on fast matrix multiplication, the resulting algorithm may not offer practical benefits for moderate dimensions; see the discussion in [Val15]. We believe overcoming these limitations is an important practically-motivated question. Finally, Question 1.3 remains open.

Acknowledgements

We are grateful to Sushrut Karmalkar, Jasper Lee, Thanasis Pittas, and Kevin Tian for discussions on robust sparse mean estimation. We also thank Ilias Diakonikolas and Daniel Kane for their many insights on robustness.

[16] The sample complexity of $\tilde{\Theta}(k^2/\epsilon^2)$ is also conjectured to be near-optimal among computationally-efficient algorithms [BB20].

References

[ABHHRT72] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances*. Princeton, NJ, USA: Princeton University Press, 1972. [1]

[Alm19] J. Alman. "An Illuminating Algorithm for the Light Bulb Problem". *Proc. 2nd Symposium on Simplicity in Algorithms (SOSA)*. 2019. [7]

[BB20] M. Brennan and G. Bresler. "Reducibility and Statistical-Computational Gaps from Secret Leakage". *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020. [7, 23]

[BBHLS21] M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. "Statistical query algorithms and low-degree tests are almost equivalent". *Proc. 34th Annual Conference on Learning Theory (COLT)*. 2021. [3]

[BDLS17] S. Balakrishnan, S. S. Du, J. Li, and A. Singh. "Computationally Efficient Robust Sparse Estimation in High Dimensions". *Proc. 30th Annual Conference on Learning Theory (COLT)*. 2017. [2, 4–6, 10, 17, 23]

[CATJFB20] Y. Cherapanamjeri, E. Aras, N. Tripuraneni, M. I. Jordan, N. Flammarion, and P. L. Bartlett. "Optimal Robust Linear Regression in Nearly Linear Time". *arXiv* abs/2007.08137 (2020). [7]

[CDG19] Y. Cheng, I. Diakonikolas, and R. Ge. "High-Dimensional Robust Mean Estimation in Nearly-Linear Time". *Proc. 30th Annual Symposium on Discrete Algorithms (SODA)*. 2019. [3, 7]

[CDGW19] Y. Cheng, I. Diakonikolas, R. Ge, and D. P. Woodruff. "Faster Algorithms for High-Dimensional Robust Covariance Estimation". *Proc. 32nd Annual Conference on Learning Theory (COLT)*. 2019. [7]

[CDKGGS22] Y. Cheng, I. Diakonikolas, D. M. Kane, R. Ge, S. Gupta, and M. Soltanolkotabi. "Outlier-Robust Sparse Estimation via Non-Convex Optimization". *Advances in Neural Information Processing Systems 35 (NeurIPS)*. 2022. [2, 4, 6–9, 17, 18, 23]

[Che21] Y. Cheng. *High-Dimensional Robust Statistics: Faster Algorithms and Optimization Landscape*. Robustness in High-dimensional Statistics and Machine Learning at IDEAL Institute. See timestamp 22:00 in the talk. 2021. [3]

[CMY20] Y. Cherapanamjeri, S. Mohanty, and M. Yau. "List decodable mean estimation in nearly linear time". *Proc. 61st IEEE Symposium on Foundations of Computer Science (FOCS)*. 2020. [7]

[dGJL07] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. "A Direct Formulation for Sparse PCA Using Semidefinite Programming". *SIAM Review* 49.3 (2007). [4, 5]

[DHL19] Y. Dong, S. B. Hopkins, and J. Li. "Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection". *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 2019. [3, 7]

[Dia19] Ilias Diakonikolas. *Computational-Statistical Tradeoffs and Open Problems*. STOC 2019 Tutorial: Recent Advances in High-Dimensional Robust Statistics. See page number 36 in http://www.iliasdiakonikolas.org/stoc19-tutorial/Tradeoffs-and-Open-Problems.pdf. 2019. [3]

[Dia23] I. Diakonikolas. *Algorithmic Robust Statistics*. Statistical thinking in the age of AI : robustness, fairness and privacy (Meeting in Mathematical Statistics). Available online at https://youtu.be/HKm0L2Cy69Y?t=3527. 2023. [3]

[DK23] I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023. [2, 6, 9, 10, 17, 18, 27]

[DKKLMS16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. "Robust Estimators in High Dimensions without the Computational Intractability". *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*. 2016. [2]

[DKKLT22] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. "Clustering Mixture Models in Almost-Linear Time via List-Decodable Mean Estimation". *Proc. 54th Annual ACM Symposium on Theory of Computing (STOC)*. 2022. [3, 7]

[DKKPP22] I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas. "Robust Sparse Mean Estimation via Sum of Squares". *Proc. 35th Annual Conference on Learning Theory (COLT)*. 2022. [2, 7]

[DKKPS19] I. Diakonikolas, D. M. Kane, S. Karmalkar, E. Price, and A. Stewart. "Outlier-Robust High-Dimensional Sparse Estimation via Iterative Filtering". *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 2019. [2, 4–6, 8, 10, 17, 23]

[DKLP22] I. Diakonikolas, D. M. Kane, J. C. H. Lee, and A. Pensia. "Outlier-Robust Sparse Mean Estimation for Heavy-Tailed Distributions". *Advances in Neural Information Processing Systems 35 (NeurIPS)*. 2022. [2, 7]

[DKPP22] I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. "Streaming Algorithms for High-Dimensional Robust Statistics". *Proc. 39th International Conference on Machine Learning (ICML)*. 2022. [3, 7]

[DKPP23a] I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. "Near-Optimal Algorithms for Gaussians with Huber Contamination: Mean Estimation and Linear Regression". *Advances in Neural Information Processing Systems 36 (NeurIPS)*. 2023. [7]

[DKPP23b] I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. "Nearly-Linear Time and Streaming Algorithms for Outlier-Robust PCA". *Proc. 40th International Conference on Machine Learning (ICML)*. 2023. [7]

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures". *Proc. 58th IEEE Symposium on Foundations of Computer Science (FOCS)*. 2017. [3, 7]

[DL22] J. Depersin and G. Lecué. "Robust Subgaussian Estimation of a Mean Vector in Nearly Linear Time". *The Annals of Statistics* 50.1 (2022). [7]

[DS18] Y. Dagan and O. Shamir. "Detecting correlations with little memory and communication". *Proc. 31st Annual Conference on Learning Theory (COLT)*. 2018. [3]

[EK12] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012. [2]

[HR09] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, 2009. [1]

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. 2015. [2, 4]

[Hub64] P. J. Huber. "Robust Estimation of a Location Parameter". *The Annals of Mathematical Statistics* 35.1 (Mar. 1964). [1]

[JKLPS20] H. Jiang, T. Kathuria, Y. T. Lee, S. Padmanabhan, and Z. Song. "A Faster Interior Point Method for Semidefinite Programming". *Proc. 61st IEEE Symposium on Foundations of Computer Science (FOCS)*. 2020. [5]

[JLT20] A. Jambulapati, J. Li, and K. Tian. "Robust sub-gaussian principal component analysis and width-independent schatten packing". *Advances in Neural Information Processing Systems 33 (NeurIPS)*. 2020. [7]

[KKK18] M. Karppa, P. Kaski, and J. Kohonen. "A Faster Subquadratic Algorithm for Finding Outlier Correlations". *ACM Trans. Algorithms* 14.3 (2018). [7, 30]

[KKKC20] M. Karppa, P. Kaski, J. Kohonen, and P. Ó Catháin. "Explicit Correlation Amplifiers for Finding Outlier Correlations in Deterministic Subquadratic Time". *Algorithmica* 82.11 (2020). [7]

[LLVZ20] Z. Lei, K. Luh, P. Venkat, and F. Zhang. "A Fast Spectral Algorithm for Mean Estimation with Sub-Gaussian Rates". *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020. [7]

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. "Agnostic Estimation of Mean and Covariance". *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*. 2016. [2]

[PW23] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023. [13]

[SV14] S. Sachdeva and N. K. Vishnoi. "Faster Algorithms via Approximation Theory". *Foundations and Trends® in Theoretical Computer Science* 9 (2014). [5]

[Val15] G. Valiant. "Finding Correlations in Subquadratic Time, with Applications to Learning Parities and the Closest Pair Problem". *Journal of the ACM* 62.2 (2015). [1, 3–7, 10, 11, 23, 27, 30]

[Val88] L. G. Valiant. "Functionality in Neural Nets". *Proc. of the Seventh AAAI National Conference on Artificial Intelligence*. AAAI Press, 1988. [7]

[van16] S. van de Geer. *Estimation and Testing Under Sparsity*. École d'Été de Probabilités de Saint-Flour. Springer, 2016. [2]

[ZJS22] B. Zhu, J. Jiao, and J. Steinhardt. "Robust Estimation via Generalized Quasi-Gradients". *Information and Inference: A Journal of the IMA* (2022). [2, 6]

A Details Deferred from Section 2

In this section, we include details that were omitted from Section 2. Appendix A.1 provides the proof of Lemma 2.7. Appendix A.2 ensures the preprocessing condition listed in Condition 2.8. Finally, Appendix A.3 gives further details about the correlation detection algorithm from [Val15].

A.1 Defining Sparse Scores

In this section, we give the proof of Lemma 2.7.

Lemma 2.7 (Sparse Filtering Lemma)**.** *Let $\epsilon \in (0, \epsilon_0)$ for a small absolute constant ϵ_0. Let T be an ϵ-corrupted version of S, where S is (ϵ, δ, k)-stable with respect to μ. Let $H \subset [d]$ be such $\|(\mathbf{\Sigma}_T - \mathbf{I})_H\|_{\mathrm{Fr}} = \lambda$ for $\lambda \gtrsim \delta^2/\epsilon$ and $|H| \leq k$. There exists an algorithm $\mathcal{A}$ that takes T, H, ϵ, and δ and returns scores $f : T \to \mathbb{R}_+$ so that $\sum_{x \in S \cap T} f(x) \leq \sum_{x \in T \setminus S} f(x)$, i.e., the sum of scores over inliers is less than that of outliers, and $\max_{x \in T} f(x) > 0$. Moreover, the algorithm runs in time $d \cdot \mathrm{poly}(|H||T|)$.*

Proof. As a first step, we simply project the data points along the coordinates in H, and by abusing the notation, call the projected set T. In the remainder of this proof, $\mathbf{I}$ refers to $\mathbf{I}_{|H|}$. Computing this projected set takes at most $d|T||H|$ time.

Let $\lambda = \|\mathbf{\Sigma}_T - \mathbf{I}\|_{\mathrm{Fr}}$. Define $\mathbf{A} = (\mathbf{\Sigma}_T - \mathbf{I})/\|\mathbf{\Sigma}_{T_1} - \mathbf{I}\|_{\mathrm{Fr}}$ so that $\mathbf{A}$ maximizes the trace inner product with $\mathbf{\Sigma}_T - \mathbf{I}$ over unit Frobenius norm matrices. Define the function $g(x) := (x - \mu_T)^\top \mathbf{A}(x - \mu_T) - \mathrm{tr}(\mathbf{A})$. We first compute the average of $g(x)$ over the T below:

$$\mathbb{E}_T[g(x)] = \mathbb{E}_T\left[\left\langle (x - \mu_T)(x - \mu_T)^\top - \mathbf{I}, \mathbf{A} \right\rangle\right] = \langle \mathbf{\Sigma}_T - \mathbf{I}, \mathbf{A} \rangle = \lambda \,.$$

By abusing notation again, we use S to denote the projection of the inliers S along the coordinates in H; Note that the projected set also inherits (ϵ, δ, k)-stability and the $\|\cdot\|_{\mathrm{Fr},k^2}$ reduces to the standard Frobenius norm. Thus, for any large subset of $S' \subset S$ with $|S'| \geq (1-\epsilon)|S|$, we use Lemma 2.6 to obtain the following:

$$\begin{aligned}
|\mathbb{E}_{S'}[g(x)]| &= \left|\mathbb{E}_{S'}\left[\left\langle (x - \mu_T)(x - \mu_T)^\top - \mathbf{I}, \mathbf{A} \right\rangle\right]\right| \\
&= \left|\langle \mathbf{\Sigma}_{S'} - \mathbf{I}, \mathbf{A} \rangle + 2(\mu - \mu_T)^\top \mathbf{A}(\mu - \mu_{S'}) + (\mu - \mu_T)^\top \mathbf{A}(\mu - \mu_T)\right| \\
&\leq \|\mathbf{\Sigma}_{S'} - \mathbf{I}\|_{\mathrm{Fr}} + 2\|\mu - \mu_T\|_2 \|\mathbf{A}\|_{\mathrm{Fr}} \|\mu - \mu_{S'}\|_2 + \|\mu - \mu_T\|_2^2 \|\mathbf{A}\|_{\mathrm{Fr}}^2 \\
&\lesssim \delta^2/\epsilon + 2(\delta + \sqrt{\epsilon\lambda})\delta + (\delta + \sqrt{\epsilon\lambda})^2 && \text{(using stability and Lemma 2.6)} \\
&\lesssim \delta^2/\epsilon + 3\delta^2 + 4\delta\sqrt{\epsilon\lambda} + \epsilon\lambda \\
&\lesssim \delta^2/\epsilon + 7\delta^2 + 2\epsilon\lambda && \text{(using } 2ab \leq a^2 + b^2\text{)} \\
&\lesssim \delta^2/\epsilon + \epsilon\lambda, && (7)
\end{aligned}$$

where we used that $\epsilon \leq 1$. The following helper result, which is a slight generalization of [DK23, Proposition 2.19] from non-negative g's to real-valued g's will be useful.

Claim A.1. *Let $h : S \to \mathbb{R}$ be a real-valued function on a finite set S. Further suppose that $|\mathbb{E}_{S'}[h(X)]| \leq \tau$ for all sets $S' \subset S$ with $|S'| \geq (1-\epsilon)|S|$. Then for $f'(x) = h(x)\mathbf{1}_{h(x) \geq 3\tau/\epsilon}$, we have that $\mathbb{E}_S[f'(x)] \leq 3\tau$.*

Proof. We shall do it in two steps. First, we show the following: for all subsets $S'' \subset S$ with $|S''| \leq \epsilon|S|$.

$$\mathbb{E}_{S''}[\max(h(X), 0)] \leq \tau + 2\tau/\epsilon \,. \tag{8}$$

To that end, for all sets S'' with $|S''| = \epsilon|S|$, the triangle inequality implies

$$|\mathbb{E}_{S''}[h(X)]| = \left|\frac{\mathbb{E}_S[h(X)] - (1-\epsilon)\mathbb{E}_{S\setminus S''}[h(X)]}{\epsilon}\right| \le 2\tau/\epsilon. \tag{9}$$

Let S_* be the top $\epsilon|S|$ entries of S in the increasing order of $h(\cdot)$, not just in the absolute value. Then establishing (8) is equivalent to establish an upper upper bound on $\frac{1}{|S_*|}\sum_{x\in S_*}\max(h(X),0)$. Thus, if all the entries of S_* are bigger than 0, then (8) follows by (9). We shall show that $h(x)$ on S_* is lower bounded by $-\tau$. Under this condition, we see that the desired result follows similarly by (9): $\frac{1}{|S_*|}\sum_{x\in S_*}\max(h(X),0) \le \frac{1}{|S_*|}\sum_{x\in S_*}(h(X)+\tau) \le \tau + 2\tau/\epsilon$. We now establish that $\min_{x\in S_*} h(x) \ge -\tau$. If there exists an $x \in S_*$ with $h(X) \le -\tau$, then the average of $S \setminus S_*$ must be less than $-\tau$, contradicting the assumption that $|\mathbb{E}_{S\setminus S_*}[h(X)]| \le \tau$. Thus, we have established (8).

Given (8), we see that the fraction of points with $h(x) \ge 3\tau/\epsilon$ must be less than $\epsilon|S|$. Otherwise, the conditional average over those points would be at least than $3\tau/\epsilon \ge 2\tau/\epsilon + \tau$, contradicting (8). Therefore, the function f is non-zero only on at most ϵ-fraction of S. Therefore, the non-negativity of f implies that

$$\frac{1}{|S|}\sum_{x\in S} f(X) \le \frac{1}{|S|}\max_{S''\subset S:|S''|\le\epsilon|S|}\sum_{x\in S''} f(X) \le \frac{1}{|S|}\max_{S''\subset S:|S''|\le\epsilon|S|}\sum_{x\in S''}\max(h(X),0) \le \epsilon(\tau + 2\tau/\epsilon), \tag{10}$$

where we use non-negativity of f and (8). □

Let $R := C'(\delta^2/\epsilon + \epsilon\lambda)$ for a large enough constant C' be the bound from (7). Combining this with the claim above, we see that defining $f(x)$ to be $g(x)\mathbf{1}_{x\ge 3R/\epsilon}$, the sum of scores over inliers is small:

$$\sum_{x\in S\cap T} f(x) \le \sum_{x\in S} f(x) \le 3R|S| = 3C'(\delta^2/\epsilon + \epsilon\lambda)|S| \le 0.25\lambda|S|\,, \tag{11}$$

where we use that $3C'\epsilon \le 1/8$ and $3C'\delta^2/\epsilon \le \lambda/8$.

On the other hand, $\mathbb{E}_{x\in T\setminus S}[f(x)]$ must be large argued as argued below. We observe that $f(x) \ge g(x) - 3R/\epsilon$, and thus applying (7) to $T \cap S$, which is of size at least $(1-\epsilon)|T| = (1-\epsilon)|S|$, we obtain

$$\begin{aligned}\sum_{x\in T\setminus S} f(x) \ge \sum_{x\in T\setminus S}(g(x) - 3R/\epsilon) &= \left(\left(\sum_{x\in T} g(x)\right) - \left(\sum_{x\in T\cap S} g(x)\right)\right) - (|T\setminus S|3R/\epsilon)\\ &\ge (\lambda|T|) - ((|T\cap S|)R) - \epsilon|T|\,(3R/\epsilon)\\ &\ge \lambda|T| - 4|T|R \ge \lambda|T|/2\,,\end{aligned} \tag{12}$$

where the last inequality follows if we show that $R \le \lambda/8$. Indeed, this follows if $C'\delta^2/\epsilon \le \lambda/16$ and $C'\epsilon \le 1/16$. Combining (11) and (12), we get the desired result; the claim on $\max_{x\in T} f(x)$ follows from (12). The complete algorithm is given below:

□

A.2 Preprocessing: Proofs of Claims 2.9 and A.2

In this section, we outline how to ensure Condition 2.8 quickly using Algorithm 1 and Lemma 2.7.

Algorithm 7 Quadratic Scores

$T \leftarrow \{(x)_H : x \in T\}$ ▷ Projection onto H
Let $\mathbf{A}$ be the matrix $(\mathbf{\Sigma}_T - \mathbf{I})/\|\mathbf{\Sigma}_T - \mathbf{I}\|_{\mathrm{Fr}}$ ▷ The matrix such that $\langle \mathbf{A}, \mathbf{\Sigma}_T - \mathbf{I}\rangle = \|\mathbf{\Sigma}_T - \mathbf{I}\|_{\mathrm{Fr}}$
Define $g(x) := (x-\mu_T)^\top \mathbf{A}(x-\mu_T) - \mathrm{tr}(\mathbf{A})$
Define $f(x)$ to be $g(x)$ if $g(x) \geq 3C'\left(\frac{\delta^2}{\epsilon^2} + \lambda\right)$ otherwise 0
Return f

Condition 2.8 (Preprocessing)**.** Let T be an ϵ-corrupted version of S, where S is (ϵ, δ, k)-stable. Suppose T satisfies $\|\mathrm{diag}(\mathbf{\Sigma}_T - \mathbf{I}_d)\|_{\mathrm{Fr},k^2} \leq \min\left(O(\delta^2/\epsilon), 0.5\right)$.

Given the sparse filtering lemma (Lemma 2.7), we can simply filter along the diagonals to ensure Condition 2.8 as shown below.

Claim A.2. *Let $\epsilon \in (0, \epsilon_0)$ for a small absolute constant ϵ_0. Let c be a small enough absolute constant and C be a large enough constant. Let T be an ϵ-corrupted version of S, where S is (ϵ, δ, k^2)-stable with respect to μ such that $\delta^2/\epsilon \leq c$. Then there is a randomized algorithm $\mathcal{A}$ that takes as inputs ϵ, δ, and k such that it outputs a set $T' \subseteq T$ such that with probability at least $8/9$: (i) T' is at most 10ϵ-corruption of S and (ii) $\|\mathrm{diag}\left(\mathbf{\Sigma}_{T'} - \mathbf{I}_d\right)\|_{\mathrm{Fr},k^2} \leq C\delta^2/\epsilon \leq 0.1$, and (iii) the algorithm runs in time $\tilde{O}(|T|dk^2 + |T|^2 d)$.*

Proof. We can simply invoke Algorithm 1 with the stopping condition on the set T_i set to

$$\|\mathrm{diag}\left(\mathbf{\Sigma}_{T_i} - \mathbf{I}_d\right)\|_{\mathrm{Fr},k} \leq C\delta^2/\epsilon$$

for a constant C large enough. To evaluate this stopping condition, we can simply compute the matrix $\mathrm{diag}\left(\mathbf{\Sigma}_{T'} - \mathbf{I}_d\right)$ in $O(d|T|)$ time. The associated $\|\cdot\|_{\mathrm{Fr},k^2}$ can be easily calculated by computing the Euclidean norm of its largest k^2 entries, again computable in $\tilde{O}(dk^2)$ time.

If the stopping condition is not satisfied, Lemma 2.7 returns the required scores. Thus, we get the desired guarantees on the set T from Theorem 2.5. □

We now give the proof of Claim 2.9.

Claim 2.9. *Let C be a large enough constant $C > 0$. Let $T'' \subset T'$ be two $O(\epsilon)$-contamination of S such that S is an $(C\epsilon, \delta, k)$-stable with respect to μ. Suppose that $\|\mathrm{diag}(\mathbf{\Sigma}_{T'} - \mathbf{I}_d)\|_{\mathrm{op},k} \lesssim \delta^2/\epsilon$. Then $\|\mathrm{diag}(\mathbf{\Sigma}_{T'} - \mathbf{I}_d)\|_{\mathrm{op},k} \lesssim \delta^2/\epsilon$*

Proof. First we note that the lower bound on the sparse eigenvalues follow rather directly as shown below. We make use of the equality $\mathbf{\Sigma}_{T''} := \frac{1}{|T''|^2}\sum_{x,y\in T''}(x-y)(x-y)^\top$. For any sparse unit vector v, we use the stability condition applied to $S \cap T''$ to obtain the following:

$$\begin{aligned}
v^\top \mathbf{\Sigma}_{T''} v &= \frac{1}{|T''|^2}\sum_{x,y\in T''}(v^\top(x-y))^2 \geq \frac{1}{|T''|^2}\sum_{x,y\in S\cap T''}(v^\top(x-y))^2 && \text{(using non-negativity)}\\
&= \frac{|S\cap T''|^2}{|T''|^2} v^\top \mathbf{\Sigma}_{S\cap T''} v \geq (1-O(\epsilon)) v^\top \mathbf{\Sigma}_{S\cap T''} v\\
&= (1-O(\epsilon))\left(v^\top\left(\mathbb{E}_{S\cap T''}\left[(X-\mu)(X-\mu)^\top\right]\right)v - \left(v^\top(\mu-\mu_{S\cap T''})\right)^2\right)\\
&\geq (1-O(\epsilon))\left(1 - O(\delta^2/\epsilon) - O(\delta^2)\right) && \text{(using stability)}\\
&\geq 1 - O(\delta^2/\epsilon), && (13)
\end{aligned}$$

where the last inequality uses $\epsilon \le \delta$. For the upper bound, we observe that for any matrix $\mathbf{A}$, $\|\mathrm{diag}(\mathbf{A})\|_{\mathrm{op},k}$ is attained by 1-sparse unit vectors v, i.e., $\|\mathbf{A}\|_{\mathrm{op},1} = \|\mathrm{diag}(\mathbf{A})\|_{\mathrm{op},k}$. Thus, for any $T'' \subset T'$ and that $|T'| \le |T''|(1+O(\epsilon))$: for any 1-sparse unit vector v,

$$v^\top \mathbf{\Sigma}_{T''} v = \frac{1}{|T''|^2} \sum_{x,y\in T''} (v^\top(x-y))^2 \le \frac{|T'|^2 2}{|T''|^2} \frac{1}{|T'|^2} \sum_{x,y\in T'} (v^\top(x-y))^2 \qquad \text{(using nonnegativity)}$$

$$= \frac{|T'|^2 2}{|T''|^2} v^\top \mathbf{\Sigma}_{T'} v \le (1+O(\epsilon)) v^\top \mathbf{\Sigma}_{T'} v \le (1+O(\epsilon))(1+O(\delta^2/\epsilon)) = 1 + O(\delta^2/\epsilon)\,, \qquad (14)$$

where we use that for 1-sparse unit vectors v, $v^\top \mathbf{\Sigma}_{T'} v = v^\top \mathrm{diag}\left(\mathbf{\Sigma}_{T'}\right) v$. Combining Equations (13) and (14) for all 1-sparse unit vectors v, we obtain that $\|\mathrm{diag}\left(\mathbf{\Sigma}_{T''} - \mathbf{I}\right)\|_{\mathrm{op},k} = O(\delta^2/\epsilon)$. □

A.3 Fast Correlation Detection

In this section, we show how to obtain Theorem 2.10 from [Val15, Theorem 2.1].

Theorem A.3 (Robust Correlation Detection For Boolean Vectors in Subquadratic Time [Val15, Theorem 2.1])**.** *Consider a set of n' vectors in $\{-1,1\}^{d'}$ and constants $\rho, \tau \in [0,1]$ with $\rho > \tau$ such that for all but at most s pairs u, v of distinct vectors, $|u^\top v|/\|u\|_2\|v\|_2 \le \tau$. There is an algorithm that, with probability $1 - o(1)$, will output all pairs of vectors whose normalized inner product is least ρ. Additionally, the runtime of the algorithm is*

$$\left(s d' n'^{0.62} + n'^{1.62 + 2.4 \frac{\log(1/\rho)}{\log(1/\tau)}} \right) \mathrm{poly}\left(\log n, 1/\tau\right) .$$

An improved algorithm with better runtime was then provided in [KKK18, Corollary 1.8], but we choose the version above for its simplicity. We provide the proof of Theorem 2.10, the version we used in this work, from Theorem A.3 using standard arguments below:

Proof of Theorem 2.10 from [Val15, Theorem 2.1]. Let $X \in \mathbb{R}^{d\times n}$ denote the matrix with the columns of X denoting the centered vectors of T. That is, if $T = \{z_1, \ldots, z_n\}$, then the i-th column of X is equal to $z_i - \mu_T$. Let X_i denote the i-th row of the matrix X. For $i, j \in [d] \times [d]$, the correlation between i-th and j-th coordinate on T, $\mathrm{corr}(i,j)$, is equal to $\left| \frac{X_i^\top X_j}{\|X_i\|_2 \|X_j\|_2} \right|$. Thus, we would like to apply Theorem A.3 with the rows of X (thus $n' = d$ and $d' = n$). However, X' is not a binary matrix.

A standard reduction allows us to compute a binary matrix that preserves the correlation between the rows of X. Let $G \in \mathbb{R}^{n\times m}$ be a matrix with independent $\mathcal{N}(0,1)$ entries. Let $Y = XG$ be in $\mathbb{R}^{d\times m}$ and let $Y' = \mathrm{sgn}(Y) \in \mathbb{R}^{d\times m}$, where sgn is applied elementwise. Thus, Y' is a boolean matrix as required in Theorem A.3. The following arguments show that the correlation between the rows of Y is preserved for m large enough. Let Y_i, Y_i' denote the i-th row of the matrices Y, Y'.

Lemma A.4 ([Val15, Lemma 4.1])**.** *If $m \ge 10\log(n)/\gamma^2$, then with probability $1 - o(1)$, we have that for all $i \ne j \in [n]$, we have that*

$$\left| \frac{\langle Y_i', Y_j' \rangle}{\|Y_i'\|_2 \|Y_j'\|_2} - \frac{2}{\pi} \arcsin\left(\frac{\langle X_i, X_j \rangle}{\|X_i\|_2 \|X_j\|_2} \right) \right| \le \gamma.$$

In particular, if the original correlation is less than τ in the absolute value, then the corresponding correlation in Y' in the absolute value is at most $(2/\pi)\arcsin(\tau)+\gamma \leq (4/\pi)\tau+\gamma \leq 2\tau+\gamma$, where we use $|\arcsin(x)| \leq |2x|$. Similarly, if the original correlation is at least ρ in absolute value, then the new correlation is at least $(2/\pi)\arcsin(\rho)-\gamma \geq (2/\pi)\rho-\gamma \geq \rho/2-\gamma$. Choosing $\gamma = \tau$, the new matrix Y' satisfies the guarantees of Theorem A.3 with $n' = d, d' = m = 10\log(n)/\tau^2, \tau' = 3\tau, \rho' = \rho/4$ (using $\tau \leq \rho/4$). The time taken to compute Y and Y' is at most ndm. Thus, the total runtime is at most

$$\left(nd\frac{\log n}{\gamma^2} + s\frac{\log n}{\tau^2}d^{0.62} + d^{1.62+2.4\frac{\log(4/\rho)}{\log(1/3\tau)}}\mathrm{poly}\left(\log n, 1/\tau\right)\right)$$
$$\leq \left(sd^{0.62} + d^{1.62+2.4\frac{\log(4/\rho)}{\log(1/3\tau)}}\right)\cdot \mathrm{poly}\left(n(\log d)/\gamma^2\right) .$$

The probability of success can be boosted using repetition, if needed. □

B Details Deferred from Section 4

In this section, we give the proofs of Lemma 4.5 and Claims 4.7 and 4.8.

B.1 Proof of Lemma 4.5

Lemma 4.5 (Sparse PCA Filter). *Let $\epsilon \in (0,\epsilon_0)$ for a small absolute constant ϵ_0 and let C be a large enough absolute constant. Let T be an ϵ-corrupted version of S, where S is (ϵ,γ,k,η)-stable with respect to v (cf. Definition 4.1) and $\eta \in (0,1)$. Let $H \subset [d]$ be such $\left\|(\mathbf{\Sigma}_T - \mathbf{I} - \eta vv^\top)_H\right\|_{\mathrm{Fr}} = \lambda$ for $\lambda \geq 8C\gamma$ and $|H| \leq k$.*

Then there exists an algorithm $\mathcal{A}$ that takes T, H, ϵ, γ, and η and returns a score mapping $f : T \to \mathbb{R}_+$ such that the sum of inliers' scores is less than outliers': $\sum_{x\in S\cap T} f(x) \leq \sum_{x \in T\setminus S} f(x)$ and $\max_{x\in T} f(x) > 0$. Moreover, the algorithm runs in time $d\mathrm{poly}(k,|T|)$.

Proof. We use the same ideas from the proof of Lemma 2.7 and similarly assume that T and S already correspond to the projected coordinates. The challenge in applying the idea as is lie in the uncertainty about v. We thus use Lemma 4.3 to first estimate v using the returned vector u which satisfies that $\|\eta uu^\top - \eta vv^\top\|_{\mathrm{Fr}} \lesssim \gamma$.[17] We give the algorithm below.

Algorithm 8 PCA Filter

$T \leftarrow \{(x)_H : x \in T\}$ ▷ Projection onto H
$u \leftarrow$ be output of dense covariance estimation algorithm Lemma 4.3
Let $\mathbf{A}$ be the matrix $(\mathbf{\Sigma}_T - \mathbf{I} - \eta uu^\top)/\|\mathbf{\Sigma}_{T_1} - \mathbf{I} - \eta uu^\top\|_{\mathrm{Fr}}$
Define $g(x) := (x-\mu_T)^\top \mathbf{A}(x-\mu_T) - \mathrm{tr}(\mathbf{A})$
Define $f(x)$ to be $g(x)$ if $g(x) \geq \Omega(\gamma/\epsilon)$ otherwise 0
Return f

Let $\lambda = \left\|(\mathbf{\Sigma}_T - \mathbf{I} - \eta vv^\top)_H\right\|_{\mathrm{Fr}}$ and define $\mathbf{A} = (\mathbf{\Sigma}_T - \mathbf{I} - \eta uu^\top)/\|\mathbf{\Sigma}_{T_1} - \mathbf{I} - \eta uu^\top\|_{\mathrm{Fr}}$. Define the function $g(x) := x^\top \mathbf{A} x - \langle \mathbf{I} + \eta uu^\top, \mathbf{A}\rangle$. Computing the average of $g(x)$ over the T, we obtain

$$\mathbb{E}_T[g(x)] = \mathbb{E}_T\left[\left\langle xx^\top - \mathbf{I} - \eta uu^\top, \mathbf{A}\right\rangle\right] = \langle \overline{\mathbf{\Sigma}}_T - \mathbf{I} - \eta uu^\top, \mathbf{A}\rangle = \|\overline{\mathbf{\Sigma}}_T - \mathbf{I} - \eta uu^\top\|_{\mathrm{Fr}}$$

[17] Observe that v here corresponds to $(v)_H$ because of the projection to H. Hence v is no longer a unit vector.

$$\geq \|\overline{\mathbf{\Sigma}}_T - \mathbf{I} - \eta vv^\top\|_{\mathrm{Fr}} - \|\eta uu^\top - \eta vv^\top\|_{\mathrm{Fr}} = \lambda - O(\gamma) \geq 3\lambda/4, \tag{15}$$

where we use that $\lambda \gtrsim \gamma$. Using the stability of S (observe that on the projected set, the $\|\cdot\|_{\mathrm{Fr},k^2}$ norm becomes the usual Frobenius norm), for any large subset of $S' \subset S$ with $|S'| \geq (1-\epsilon)|S|$, the closeness between u and v implies the following:

$$\begin{aligned}|\mathbb{E}_{S'}[g(x)]| &= \left|\mathbb{E}_{S'}\left[\left\langle xx^\top - \mathbf{I} - \eta uu^\top, \mathbf{A}\right\rangle\right]\right| = \left|\left\langle \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} - \eta uu^\top, \mathbf{A}\right\rangle\right| \\ &= \left|\left\langle \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} - \eta vv^\top, \mathbf{A}\right\rangle + \left\langle \eta vv^\top - \eta uu^\top, \mathbf{A}\right\rangle\right| \leq \left\|\overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} - \eta vv^\top\right\|_{\mathrm{Fr}} + \eta\left\|vv^\top - uu^\top\right\|_{\mathrm{Fr}} \\ &\leq \gamma + O(\gamma) \leq C\gamma \end{aligned} \tag{16}$$

for a large enough absolute constant C. Now define $f(x) := g(x)\mathbf{1}_{g(x)\geq 3C\gamma/\epsilon}$, i.e., $f(x)$ is equal to $g(x)$ if $g(x) \geq 3C\gamma/\epsilon$ and 0 otherwise. By Claim A.1,

$$\sum_{x\in S\cap T} f(x) \leq \sum_{x\in S} f(x) \leq 3C\gamma|S| \leq |T|\lambda/4\,, \tag{17}$$

where we use that $\lambda \gtrsim C\gamma$.

On the other hand, $\mathbb{E}_{x\in T\setminus S}[f(x)]$ must be large argued as argued below. Applying (16) to $T\cap S$, which is of size at least $(1-\epsilon)|T| = (1-\epsilon)|S|$, we obtain

$$\begin{aligned}\sum_{x\in T\setminus S} f(x) &\geq \sum_{x\in T\setminus S}(g(x) - 3C\gamma/\epsilon) = \left(\left(\sum_{x\in T} g(x)\right) - \left(\sum_{x\in T\cap S} g(x)\right)\right) - (|T\setminus S|3C\gamma/\epsilon) \\ &\geq (3\lambda|T|/4) - ((|T\cap S|)(C\gamma)) - \epsilon|T|\,(3C\gamma/\epsilon) \\ &\geq (3\lambda|T|/4) - (4|T|C\gamma) \\ &\geq \lambda|T|/2\,, \end{aligned} \tag{18}$$

where we use (15) and $\lambda \gtrsim \gamma$. The desired conclusions follow from (17) and (18). □

B.2 Proofs of Claims 4.7 and 4.8

Claim 4.7. *Let $\epsilon \in (0,\epsilon_0)$ and $\gamma \in (0,\gamma_0)$ for small constants $\epsilon_0 \in (0,1/2)$, $\gamma_0 \in (0,1)$. Let sparsity $k \in \mathbb{N}$. Let C be a large enough constant and T be an ϵ-corrupted set S where S is $(C\epsilon,\gamma,k',\eta)$-pca-stable with respect to an unknown k-sparse unit vector $v \in \mathbb{R}^d$, $\eta \in (0,1)$, $\gamma \geq \epsilon$, and $k' = C'k^2$ for a large enough constant $C' > 0$. There is a randomized algorithm $\mathcal{A}$ that takes as input the corrupted set T, contamination rate ϵ, sparsity $k \in \mathbb{N}$, and a parameter $q \in \mathbb{N}$, and returns a set $T' \subset T$ and $H_1 \subset [d]$ in time $O(d\mathrm{poly}(k,|T|))$ such that with probability* 0.9

1. *T' is an $O(\epsilon)$-contamination of S.*
2. *Each diagonal entry of $\overline{\mathbf{\Sigma}}_{T'} \in [1/2,4]$.*
3. *T' and H_1 satisfy Condition 4.6, i.e., $|H_1| \lesssim k^2$ and $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_{T'} - \mathbf{I}_d)_{H_1^{\complement}}\|_{\mathrm{Fr},k^2} \lesssim \gamma$.*

Proof. We will filter the set using Lemma 4.5 following the template of Algorithm 1 until the second and the third conditions are met.

Starting with the second condition, the lower bound on the diagonal entries follows by the fact that T contains an ϵ-fraction of $S\cap T$ and diagonal entries of $\overline{\mathbf{\Sigma}}_{S\cap T}$ are at least $1-\gamma$. Since ϵ and γ are small enough, $(1-\epsilon)(1-\gamma) \geq 1/2$. We now focus on establishing the upper bound, for each coordinate $i \in [d]$, the true variance is at most $(1+\eta\|v\|_\infty^2) \leq 2$, where $\|\cdot\|_\infty$ denotes the ℓ_∞ norm.

Thus, if for any coordinate $i \in [d]$, the empirical variance of T is larger than 3, which is bigger than $1 + \eta + \lambda$, $H := \{i\}$ satisfies the condition of Lemma 4.5. Thus, we can filter points until all the empirical variances are less than 3 following the template of Theorem 2.5.

We now turn our attention to the third condition. For a vector $x \in \mathbb{R}^d$, let the function $g : \mathbb{R}^d \to \mathbb{R}^d$ be the coordinate-wise square of its input. Thus, for $X \sim \mathcal{N}(0, \mathbf{I} + \eta vv^\top)$, we have that $\mu := \mathbb{E}[g(X)] = u + \eta g(v)$, where $u \in \mathbb{R}^d$ denotes the all ones vector. Therefore, the required condition in Condition 4.6 can also be written equivalently as $\| \left(\mu_{f(T)} - u\right)_{H_1^\complement} \|_{2,k^2} = O(\gamma)$. Let $f(T)$ and $f(S)$ denote the sets transformed by f, i.e., $f(T) := \{f(x) : x \in T\}$ and $f(S) := \{f(x) : x \in S\}$. We first compute $\mu_{f(T)} - u$, which takes $O(d|T|)$ time. Let $J \subset [d]$ denote the coordinates i for which the i-th coordinate of $\mu_{f(T)} - u$ is bigger than γ/k in absolute value. Let J_* denote the support of the sparse spike vector v.

Consider the case when J is a large enough set: $|J| > k'$. Then let $H \subset J$ be any subset of size k'. Then $|H \setminus J_*| > |H| - |J_*| = k' - k \geq k'/2$, And thus we have that

$$\begin{aligned}
\|(\mathbf{\Sigma}_T - \mathbf{I}_d - \eta vv^\top)_H\|_{\mathrm{Fr},k'^2} &\geq \|(\mathbf{\Sigma}_T - \mathbf{I}_d - \eta vv^\top)_{H\setminus J_*}\|_{\mathrm{Fr},k'^2} \geq \\
&\geq \|(\mathbf{\Sigma}_T - \mathbf{I}_d)_{H\setminus J_*}\|_{\mathrm{Fr},k'^2} \\
&\geq \|\mathrm{diag}\left((\mathbf{\Sigma}_T - \mathbf{I}_d)_{H\setminus J_*}\right)\|_{\mathrm{Fr},k'^2} \\
&\geq (\gamma/k)\sqrt{(k'/2)} \gtrsim \gamma\,,
\end{aligned}$$

since $k' \gtrsim k^2$. Thus, we have obtained a set H that satisfies the guarantees of Lemma 4.5, which allows us to filter using the template of Algorithm 1.

If on the other hand, J happens to be small, then we return $H_1 = J$ and since $\mu_{f(T)}$ is (γ/k)-close to u (that is, all-ones vector) on $H_1^\complement$, we obtain

$$\|\mathrm{diag}(\mathbf{\Sigma}_T - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{Fr},k^2} = \| \left(\mu_{f(T)} - u\right)_{H_1^\complement} \|_{2,k^2} \leq k\gamma/k \leq \gamma.$$

□

Claim 4.8. *Let C be a large enough constant $C > 0$ and $k, k' \in \mathbb{N}$. Let $T'' \subset T'$ be two $O(\epsilon)$-contamination of S such that S is an $(C\epsilon, \gamma, k', \eta)$-pca-stable with respect to v and $\eta \in (0,1)$. Let $H \subset [d]$ be a small subset $|H| \leq k$ such that $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_{T'} - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{Fr},k^2} \lesssim \gamma$. Then $\|\mathrm{diag}(\overline{\mathbf{\Sigma}}_{T''} - \mathbf{I}_d)_{H_1^\complement}\|_{\mathrm{op},k} \lesssim \gamma$.*

Proof. Our proof strategy will be similar to that of Claim 2.9, and we refer the reader to the proof of Claim 2.9 for more details. Starting with the lower bound on the sparse eigenvalues of $\overline{\mathbf{\Sigma}}_{T''}$, we note that for any k'-sparse unit vector u, the stability condition applied to $S \cap T''$ applies

$$u^\top \overline{\mathbf{\Sigma}}_{T''} u \geq \frac{|S \cap T''|}{|T''|} u^\top \mathbf{\Sigma}_{S\cap T''} u \geq (1 - O(\epsilon))(u^\top (I + \eta vv^\top)u - \gamma) \geq (1 - O(\epsilon))(1 - O(\gamma)) \geq 1 - O(\gamma)$$

since $\epsilon \leq \gamma$. Applying this inequality for 1-sparse unit vectors u, we obtain that for any unit vector z: $-z^\top \mathrm{diag}(\overline{\mathbf{\Sigma}}_T - \mathbf{I})z \leq C\gamma$ for a large constant $C > 0$.

Turning towards upper bound, we proceed as follows: for any 1-sparse unit vector u supported on $H_1^\complement$,

$$u^\top \mathrm{diag}\left(\mathbf{\Sigma}_{T''}\right) u = u^\top \mathbf{\Sigma}_{T''} u = \frac{1}{|T''|} \sum_{x\in T''} (u^\top x)^2 \leq (1 + O(\epsilon)) u^\top \mathbf{\Sigma}_{T'} u \leq (1 + O(\epsilon))(1 + O(\gamma)) = 1 + O(\gamma)\,.$$

Overall, we obtain the desired guarantee of $\left\| \mathrm{diag}\left((\mathbf{\Sigma}_{T''} - \mathbf{I})_{H_1^\complement} \right) \right\|_{\mathrm{op},k} = O(\gamma)$.

□