Outlier-Robust PCA: The High Dimensional Case

Huan Xu, Constantine Caramanis, *Member*, and Shie Mannor, *Senior Member*

***Abstract*— Principal Component Analysis plays a central role in statistics, engineering and science. Because of the prevalence of corrupted data in real-world applications, much research has focused on developing robust algorithms. Perhaps surprisingly, these algorithms are unequipped – indeed, unable – to deal with outliers in the *high dimensional setting* where the number of *observations* is of the same magnitude as the number of *variables* of each observation, and the data set contains some (arbitrarily) corrupted observations. We propose a High-dimensional Robust Principal Component Analysis (HR-PCA) algorithm that is efficient, robust to contaminated points, and easily kernelizable. In particular, our algorithm achieves maximal robustness – it has a breakdown point of 50% (the best possible) while all existing algorithms have a breakdown point of *zero*. Moreover, our algorithm recovers the optimal solution *exactly* in the case where the number of corrupted points grows sub linearly in the dimension.**

***Index Terms*— Statistical Learning, Dimension Reduction, Principal Component Analysis, Robustness, Outlier**

I. Introduction

The analysis of very high dimensional data – data sets where the dimensionality of each observation is comparable to or even larger than the number of observations – has drawn increasing attention in the last few decades [1], [2]. Individual observations can be curves, spectra, images, movies, behavioral characteristics or preferences, or even a genome; a single observation's dimensionality can be astronomical, and, critically, it can equal or even outnumber the number of samples available. Practical high dimensional data examples include DNA Microarray data, financial data, climate data, web search engine, and consumer data. In addition, the nowadays standard "Kernel Trick" [3], a pre-processing routine which non-linearly maps the observations into a (possibly infinite dimensional) Hilbert space, transforms virtually every data set to a high dimensional one. Efforts to extend traditional statistical tools (designed for the low dimensional case) into this high-dimensional regime are often (if not generally) unsuccessful. This fact has stimulated research on formulating fresh data-analysis techniques able to cope with such a "dimensionality explosion."

Principal Component Analysis (PCA) is perhaps one of the most widely used statistical techniques for dimensionality reduction. Work on PCA dates back to the beginning of the 20^{th} century [4], and has become one of the most important techniques for data compression and feature extraction. It is widely used in statistical data analysis, communication theory, pattern recognition, image processing and far beyond [5]. The standard PCA algorithm constructs the optimal (in a least-square sense) subspace approximation to observations by computing the eigenvectors or Principal Components (PCs) of the sample covariance or correlation matrix. Its broad application can be attributed to primarily two features: its success in the classical regime for recovering a low-dimensional subspace even in the presence of noise, and also the existence of efficient algorithms for computation. Indeed, PCA is nominally a non-convex problem, which we can, nevertheless, solve, thanks to the magic of the SVD which allows us to *maximize* a convex function. It is well-known, however, that precisely because of the quadratic error criterion, standard PCA is exceptionally fragile, and the quality of its output can suffer dramatically in the face of only a few (even a vanishingly small fraction) grossly corrupted points. Such non-probabilistic errors may be present due to data corruption stemming from sensor failures, malicious tampering, or other reasons. Attempts to use other error functions growing more slowly than the quadratic that might be more robust to outliers, result in non-convex (and intractable) optimization problems.

In this paper, we consider a high-dimensional counterpart of Principal Component Analysis (PCA) that is robust to the existence of *arbitrarily corrupted* or contaminated data. We start with the standard statistical setup: a low dimensional signal is (linearly) mapped to a very high dimensional space, after which point high-dimensional Gaussian noise is added, to produce points that no longer lie on a low dimensional subspace. At this point, we deviate from the standard setting in two important ways: (1) *a constant fraction of the points are arbitrarily corrupted* in a perhaps non-probabilistic manner. We emphasize that these "outliers" can be entirely arbitrary, rather than from the tails of any particular distribution, e.g., the noise distribution; we call the remaining points "authentic"; (2) the number of data points is of the same order as (or perhaps considerably smaller than) the dimensionality. As we discuss below, these two points confound (to the best of our knowledge) all tractable existing Robust PCA algorithms.

A fundamental feature of the high dimensionality is that the noise is large in some direction, with very high probability, and therefore definitions of "outliers" from classical statistics are of limited use in this setting. Another important property of this setup is that the signal-to-noise ratio (SNR) can go to

The work of H. Xu was supported by the National University of Singapore startup grant R-265-000-384-133. The work of C. Caramanis was supported by U.S. National Science Foundation under Grants EFRI-0735905, EECS-1056028, and Defence Threat Reduction Agency under grant HDTRA 1-08-0029. The work of S. Mannor was supported by Israel Science Foundation (contract 890015). Preliminary versions of these results have appeared in part, in The Proceedings of the 46th Annual Allerton Conference on Control, Communication, and Computing, and at the 23rd international Conference on Learning Theory (COLT).

H. Xu is with the Department of Mechanical Engineering, National University of Singapore, Singapore. email: (mpexuh@nus.edu.sg).

C. Caramanis is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA email: (caramanis@mail.utexas.edu).

S. Mannor is with the Department of Electrical Engineering, Technion, Israel. email: (shie@ee.technion.ac.il).

zero, as the ℓ_2 norm of the high-dimensional Gaussian noise scales as the square root of the dimensionality. In the standard (i.e., low-dimensional case), a low SNR generally implies that the signal cannot be recovered, even without any corrupted points.

The Main Result

Existing algorithms fail spectacularly in this regime: to the best of our knowledge, there is no algorithm that can provide any nontrivial bounds on the quality of the solution in the presence of even a vanishing fraction of corrupted points. In this paper we do just this. We provide a novel robust PCA algorithm we call High Dimensional PCA (HR-PCA). HR-PCA is efficient (performing at most n, the number of samples, rounds of PCA), and robust with provable nontrivial performance bounds with up to *up to 50% arbitrarily corrupted points*. If that fraction is vanishing (e.g., n samples, $\sqrt{n}$ outliers), then HR-PCA guarantees perfect recovery of the low-dimensional subspace providing optimal approximation of the authentic points. Moreover, our algorithm is easily kernelizable. This is the first algorithm of its kind: tractable, maximally robust (in terms of breakdown point – see below) and asymptotically optimal when the number of authentic points scales faster than the number of corrupted points.

The proposed algorithm performs a PCA and a random removal alternately. Therefore, in each iteration a candidate subspace is found. The random removal process guarantees that with high probability, one of candidate solutions found by the algorithm is "close" to the optimal one. Thus, comparing all solutions using a (computational efficient) one-dimensional robust variance estimator leads to a "sufficiently good" output. Alternatively, our algorithm can be shown to be a randomized algorithm giving a constant factor approximation to the non convex projection pursuit algorithm.

Organization and Notation

The paper is organized as follows: In Section II we discuss past work and the reasons that classical robust PCA algorithms fail to extend to the high dimensional regime. In Section III we present the setup of the problem, and the HR-PCA algorithm. We also provide finite sample and asymptotic performance guarantees. Section IV is devoted to the kernelization of HR-PCA. We provide some numerical experiment results in Section V. The performance guarantees are proved in Section VI. Some technical details in the derivation of the performance guarantees are postponed to the appendix.

Capital letters and boldface letters are used to denote matrices and vectors, respectively. A $k \times k$ identity matrix is denoted by I_k. For $c \in \mathbb{R}$, $[c]^+ \triangleq \max(0, c)$. We let $\mathcal{B}_d \triangleq \{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leq 1\}$, and $\mathcal{S}_d$ be its boundary. We use a subscript $(\cdot)$ to represent order statistics of a random variable. For example, let $v_1, \ldots, v_n \in \mathbb{R}$, and $f : \mathbb{R} \mapsto \mathbb{R}$. Then $v_{(1)}, \ldots, v_{(n)}$ is a permutation of $v_1, \ldots, v_n$, and $f(v)_{(1)}, \ldots, f(v)_{(n)}$ is a permutation of $f(v_1), \ldots, f(v_n)$, both in non-decreasing order. The operator $\vee$ and $\wedge$ are used to represent the maximal and the minimal value of the operands, respectively. For example, $x \vee y = \max(x, y)$. The standard asymptotic notations $o(\cdot), O(\cdot), \Theta(\cdot), \omega(\cdot)$ and $\Omega(\cdot)$ are used to lighten notations. Throughout the paper, "with high probability" means with probability (jointly on sampling and the randomness of the algorithm) at least $1 - Cn^{-10}$ for some absolute constant C. Indeed that the exponent -10 is arbitrary, and can readily changed to any fixed integer with all the results still hold.

II. Relation to Past Work

In this section, we discuss past work and the reasons that classical robust PCA algorithms fail to extend to the high dimensional regime.

Much previous robust PCA work focuses on the traditional robustness measurement known as the "breakdown point" [6]: the percentage of corrupted points that can make the output of the algorithm *arbitrarily* bad. To the best of our knowledge, no other algorithm can handle *any constant fraction of outliers* with a lower bound on the error in the high-dimensional regime. That is, the best-known breakdown point for this problem is zero. As discussed above, we show that the algorithm we provide has breakdown point of 50%, which is the best possible for any algorithm. In addition to this, we focus on providing explicit bounds on the performance, for all corruption levels up to the breakdown point.

In the low-dimensional regime where the observations significantly outnumber the variables of each observation, several robust PCA algorithms have been proposed (e.g., [7]–[16]). These algorithms can be roughly divided into two classes: (i) The algorithms that obtain a robust estimate of the covariance matrix and then perform standard PCA. The robust estimate is typically obtained either by an outlier rejection procedure, subsampling (including "leave-one-out" and related approaches) or by a robust estimation procedure of each element of the covariance matrix; (ii) So-called *projection pursuit* algorithms that seek to find directions $\{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$ maximizing a robust variance estimate of the points projected to these d dimensions. Both approaches encounter serious difficulties when applied to high-dimensional data-sets, as we explain.

One of the fundamental challenges tied to the high-dimensional regime relates to the relative magnitude of the signal component and the noise component of even the authentic samples. In the classical regime, most of the authentic points must have a larger projection along the true (or optimal) principal components than in other directions. That is, the noise component must be smaller than the signal component, for many of the authentic points. In the high dimensional setting entirely the opposite may happen. As a consequence, and in stark deviation from our intuition from the classical setting, in the high dimensional setting, all the authentic points may be far from the origin, far from each other, and nearly perpendicular to all the principal components. To explain this better, consider a simple generative model for the *authentic points*: $\mathbf{y}_i = A\mathbf{x}_i + \mathbf{v}_i$, $i = 1, \ldots, n$ where A is a $p \times d$ matrix, $\mathbf{x}$ is drawn from a zero mean symmetric random variable, and $\mathbf{v} \sim \mathcal{N}(0, I_p)$. Let us suppose that for n the number of points, p the ambient dimension, and $\sigma_A = \sigma_{\max}(A)$ the largest singular value of A, we have: $n \approx p \gg \sigma_A$ and

also much bigger than d, the number of principal components. Then, standard calculation shows that $\sqrt{\mathbb{E}(\|A\mathbf{x}\|_2^2)} \leq \sqrt{d}\sigma_A$, while $\sqrt{\mathbb{E}(\|\mathbf{v}\|_2^2)} \approx \sqrt{p}$, and in fact there is sharp concentration of the Gaussian about this value. Thus we may have $\sqrt{\mathbb{E}(\|\mathbf{v}\|_2^2)} \approx \sqrt{p} \gg \sqrt{d}\sigma_A \geq \sqrt{\mathbb{E}(\|A\mathbf{x}\|_2^2)}$: the magnitude of the noise may be vastly larger than the magnitude of the signal.

While this observation is simple, it has severe consequences. First, Robust PCA techniques based on some form of outlier rejection or anomaly detection, are destined to fail. The reason is that in the ambient (high dimensional) space, since the noise is the dominant component of even the authentic points, it is essentially impossible to distinguish a corrupted from an authentic point.

Two criteria are often used for to determine a point being an outlier, namely, points with large Mahalanobis distance or points with large Stahel-Donoho outlyingness. The Mahalanobis distance of a point $\mathbf{y}$ is defined as

$$D_M(\mathbf{y}) = \sqrt{(\mathbf{y} - \overline{\mathbf{y}})^\top S^{-1} (\mathbf{y} - \overline{\mathbf{y}})},$$

where $\overline{\mathbf{y}}$ is the sample mean and S is the sample covariance matrix. Stahel-Donoho outlyingness is defined as:

$$u_i \triangleq \sup_{\|\mathbf{w}\|=1} \frac{|\mathbf{w}^\top \mathbf{y}_i - \mathrm{med}_j(\mathbf{w}^\top \mathbf{y}_j)|}{\mathrm{med}_k |\mathbf{w}^\top \mathbf{y}_k - \mathrm{med}_j(\mathbf{w}^\top \mathbf{y}_j)|}.$$

Both the Mahalanobis distance and the Stahel-Donoho (S-D) outlyingness are extensively used in existing robust PCA algorithms. For example, Classical Outlier Rejection, Iterative Deletion and various alternatives of Iterative Trimmings all use the Mahalanobis distance to identify possible outliers. Depth Trimming [17] weights the contribution of observations based on their S-D outlyingness. More recently, the ROBPCA algorithm proposed in [18] selects a subset of observations with least S-D outlyingness to compute the d-dimensional signal space. Indeed, consider λn corrupted points of magnitude some (large) constant multiple of σ_A, all aligned. Using matrix concentration arguments (we develop these arguments in detail in the sequel) it is easy to see that the output of PCA can be strongly manipulated; on the other hand, since the noise magnitude is $\sqrt{p} \approx \sqrt{n}$ in a direction perpendicular to the principal components, the Mahalanobis distance of each corrupted point will be very small. Similarly, the S-D outlyingness of the corrupted points in this example is smaller than that of the authentic points, again due to the overwhelming magnitude of the noise component of each authentic point.

Subsampling and leave-one-out attempts at outlier rejection also fail to work, this time because of the large number (a constant fraction) of outliers. Other algorithms designed for robust estimation of the covariance matrix fail because there are not enough observations compared to the dimensionality. For instance, the widely used Minimum Volume Ellipsoid (MVE) estimator [19] finds the minimum volume ellipsoid that covers half the points, and uses it to define a robust covariance matrix. Finding such an ellipsoid is typically hard (combinatorial). Yet beyond this issue, in the high dimensional regime, the minimum volume ellipsoid problem is fundamentally ill posed.

The discussion above lies at the core of the failure of many popular algorithms. Indeed, in [17], several classical covariance estimators including M-estimator [20], Convex Peeling [21], [22], Ellipsoidal Peeling [23], [24], Classical Outlier Rejection [25], [26], Iterative Deletion [27] and Iterative Trimming [28], [29] are all shown to have breakdown points upper-bounded by the inverse of the dimensionality, hence not useful in the regime of interest.

Next, we turn to Algorithmic Tractability. Projection pursuit algorithms seek to find a direction (or set of directions) that maximizes some robust measure of variance in this low-dimensional setting. A common example (and one which we utilize in the sequel) is the so-called trimmed variance in a particular direction, $\mathbf{w}$. This projects all points onto $\mathbf{w}$, and computes the average squared distance from the origin for the $(1-\eta)$-fraction of the points for some $\eta \in (0,1)$. As a byproduct of our analysis, we show that this procedure has excellent robustness properties; in particular, our analysis implies that this has breakdown point 50% if η is set as 0.5. However, it is easy to see that this procedure requires the solution of a non-convex optimization problem. To the best of our knowledge, there is no tractable algorithm that can do this. (As part of our work, we implicitly provide a randomized algorithm with guaranteed approximation rate for this problem). In the classical setting, we note that the situation is different. In [30], the authors propose a fast approximate Projection-Pursuit algorithm, avoiding the non-convex optimization problem of finding the optimal direction, by only examining the directions defined by sample. In the classical regime, in most samples the signal component is larger than the noise component, and hence many samples make an acute angle with the principal components to be recovered. In contrast, in the high-dimensional setting this algorithm fails, since as discussed above, the direction of each sample is almost orthogonal to the direction of true principal components. Such an approach would therefore only be examining candidate directions nearly orthogonal to the true maximizing

Finally, we discuss works addressing robust PCA using *low-rank techniques and matrix decomposition*. Starting with the work in [31], [32] and [33], recent focus has turned to the problem of recovering a low-rank matrix from corruption. The work in [31], [32] consider matrix completion — recovering a low-rank matrix from an overwhelming number of erasures. The work initiated in [33], and subsequently continued and extended in [**?**], [34] focuses on recovering a low-rank matrix from erasures and possibly gross *but sparse* corruptions. In the noiseless case, stacking all our samples as columns of a $p \times n$ matrix, we indeed obtain a corrupted low rank matrix. But the corruption is not sparse; rather, the corruption is *column-sparse*, with the corrupted columns corresponding to the corrupted points. in addition to this, the matrix has Gaussian noise. It is easy to check via simple simulation, and not at all surprising, that the sparse-plus-low-rank matrix decomposition approaches fail to recover a low-rank matrix corrupted by a column-sparse matrix.

When this manuscript was under review, a subset of us, together with co-authors, developed a low-rank matrix de-

composition technique to handle outliers (i.e., column-wise corruption) [35], [36], see also [37] for a similar study performed independently. In [35], [36], we give conditions that guarantee the exact recovery of the principal components and the identity of the outliers in the noiseless case, up to a (small) constant fraction of outliers depending on the number of principal components. We provide parallel approximate results in the presence of Frobenius-bounded noise. Outside the realm where the guarantees hold, the performance of matrix decomposition approach is unknown. In particular, its breakdown point depends inversely on the number of principal components, and the dependence of noise is severe. Specifically, the level of noise considered here would result in only trivial bounds. In short, we do not know of performance guarantees for the matrix decomposition approach that are comparable to the results presented here (although it is clearly a topic of interest).

III. HR-PCA: Setup, Algorithm and Guarantees

In this section we describe the precise setting, then provide the HR-PCA algorithm, and finally state the main theorems of the paper, providing the performance guarantees.

A. Problem Setup

This paper is about the following problem: Given a mix of *authentic* and *corrupted* points, our goal is to find a low-dimensional subspace that captures as much variance *of the authentic points*. The corrupted points are arbitrary in every way except their number, which is controlled. We consider two settings for the authentic points: deterministic (arbitrary) model, and then a stochastic model. In the deterministic setting, we assume nothing about the authentic points; in the stochastic setting, we assume the standard generative model, namely, that authentic points are generated according to $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{v}_i$, as we explain below. In either case, we measure the quality of our solution (i.e., of the low-dimensional subspace) by comparing to how much variance of the authentic points we capture, compared to the maximum possible. The guarantees for the deterministic setting are, necessarily, presented in reference to the optimal solution which is a function of all the points. The stochastic setting allows more interpretable results, since the optimal solution is defined by the matrix A.

We now turn to the basic definitions.

- Let n denote the total number of samples, and p the ambient dimension, so that $\mathbf{y}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. Let λ denote the fraction of corrupted points; thus, there are $t = (1-\lambda)n$ "authentic samples" $\mathbf{z}_1, \ldots, \mathbf{z}_t \in \mathbb{R}^p$. We assume $\lambda < 0.5$. Hence we have $0.5n \leq t \leq n$, i.e., t and n are of the same order.
- The remaining λn points are outliers (the corrupted data) and are denoted $\mathbf{o}_1, \ldots, \mathbf{o}_{n-t} \in \mathbb{R}^p$ and as emphasized above, they are arbitrary (perhaps even maliciously chosen).
- We only observe the contaminated data set

$$\mathcal{Y} \triangleq \{\mathbf{y}_1 \ldots, \mathbf{y}_n\} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\} \bigcup \{\mathbf{o}_1, \ldots, \mathbf{o}_{n-t}\}.$$

An element of $\mathcal{Y}$ is called a "point".

Setup 1: In the deterministic setup, we make no assumptions whatsoever on the authentic points, and thus there is no implicit assumption that there is a good low-dimensional approximation of these points. The results are necessarily finite-sample, and their quality is a function of all the authentic points.

Setup 2: The stochastic setup is the familiar one: the authentic samples are generated by

$$\mathbf{z}_i = A\mathbf{x}_i + \mathbf{v}_i.$$

Here, $\mathbf{x}_i \in \mathbb{R}^d$ (the "signal") are i.i.d. samples of a random variable $\mathbf{x} \sim \mu$, and $\mathbf{v}_i$ (the "noise") are independent realizations of $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, I_p)$. The matrix $A \in \mathbb{R}^{p \times d}$ maps the low-dimensional signal $\mathbf{x}$ to $\mathbb{R}^p$. We note that the intrinsic dimension d, and the distribution of $\mathbf{x}$ (denoted by μ) are unknown. We assume μ is spherically symmetric with mean zero and variance I_d. We denote its one-dimensional marginal by $\overline{\mu}$. We assume $\overline{\mu}(\{0\}) < 0.5$ and it is sub-exponential, i.e., there exists $\alpha > 0$ such that $\overline{\mu}\left((-\infty, -x] \bigcup [x, +\infty)\right) \leq \exp(1 - \alpha x)$ for all $x > 0$.[1]

Remark 1: We briefly explain some of the assumptions made in Setup 2. While we assume the noise to be Gaussian, similar results still hold for sub-Gaussian noise. The assumption that μ has a unit co-variance matrix is made without loss of generality, due to the fact that we can normalize the variance of μ by picking an appropriate A. We assume μ to be zero-mean as this can be achieved by subtracting from every point the mean of the true samples. Notice that unlike robust PCA, robustly estimating the mean of true samples under outliers is a well-studied problem [6], and effective methods are readily available. The spherical symmetry assumption on μ is non-trivial: without it, the results appear to be somewhat weaker, depending on the skew of the distribution. We demonstrate how our results are translated to this setting in Remark 2 below.

The goal of this paper is to compute $\hat{d}$ principal components, $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_{\hat{d}}$ that approximate the authentic points in the least squared error sense. As is well-known, this is equivalent to asking that they capture as much *variance* of the projected authentic points, (i.e., they maximize the average squared distance from the origin of the authentic points projected onto the span of the $\{\overline{\mathbf{w}}_i\}$). We compare the output of our algorithm to the best possible variance captured by the optimal principal $\hat{d}$ components $\mathbf{w}_1^*, \ldots, \mathbf{w}_{\hat{d}}^*$. Note that in Setup 1 there is no intrinsic dimension d defined. In Setup 2 the number, d, of columns of A is a natural candidate. However, this may not be known, or, one may seek an approximation to a subspace of lower-yet dimension. Naturally, the results are most interesting for small values of $\hat{d}$.

High Dimensional Setting and Asymptotic Scaling: While we provide results for the deterministic setting (Setup 1) the primary focus of this paper is the stochastic case. Even our finite sample results are best understood in the context of the

[1] As we discuss below, d can go infinity. In such a statistical setup, instead of requiring the d-dimensional distribution to satisfy some properties such as sub-exponentiality (which is void as d can go infinity), the standard approach (e.g., [38]) is to require that the 1-d marginal of the distribution must satisfy these properties.

asymptotic results we provide. To this end, we must discuss the asymptotic scaling regime in force throughout. We focus on the high dimensional statistical case where $n \approx p \gg d$, and n, p, d can go infinity simultaneously. Moreover, we require that $\mathrm{trace}(A^\top A) \gg d$ or equivalently $\frac{1}{d}\sum_{j=1}^{d}(\sigma_j^*)^2 \gg 1$ where σ_j^* is the j^{th} singular vector of A, i.e., the signal strength scales to infinity. However, its rate can be arbitrary, and in particular, the signal strength can scale much more slowly than the scaling of n and p.

We are particularly interested in the asymptotic performance of HR-PCA when *the dimension and the number of observations grow together* to infinity, faster than d and much faster than the signal strength. Precisely, our asymptotic setting is as follows. Suppose there exists a sequence of sample sets $\{\mathcal{Y}(j)\} = \{\mathcal{Y}(1), \mathcal{Y}(2), \dots\}$, where for $\mathcal{Y}(j)$, $n(j)$, $p(j)$, $A(j)$, $d(j)$, etc., denote the corresponding values of the quantities defined above. Then the following must hold for some positive constants c_1, c_2:

$$\limsup_{j\to\infty} \frac{p(j)}{n(j)} < +\infty; \quad \frac{n(j)}{d(j)[\log^5 d(j)]} \uparrow \infty; \quad n(j) \uparrow +\infty;$$
$$\frac{\mathrm{trace}(A(j)^\top A(j))}{d(j)} \uparrow +\infty; \quad \limsup_{j\to\infty} \frac{\hat{d}(j)}{d(j)} < +\infty. \tag{1}$$

B. Key Idea and Main Algorithm

The key idea of our algorithm is remarkably simple. It focuses on simultaneously discovering structure and casting out *potential* corrupted points. The work-horse of the HR-PCA algorithm we present below is a tool from classical robust statistics: a robust variance estimator capable of estimating the variance in the classical (low-dimensional, with many more samples than dimensions) setting, even in the presence of a constant fraction of arbitrary outliers. While we cannot optimize it directly as it is nonconvex[2] we provide a randomized algorithm that does so. We use the so-called *trimmed variance* as our Robust Variance Estimator (RVE), defined as follows: For $\mathbf{w} \in \mathcal{S}_p$, we define the Robust Variance Estimator (RVE) as

$$\overline{V}_{\hat{t}}(\mathbf{w}) \triangleq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{y}|^2_{(i)},$$

where $\hat{t} = (1-\hat{\lambda})n$ is any *lower bound* on the number of authentic points. If we know $t = (1-\lambda)n$ exactly, we take $\hat{t} = t$. The RVE above computes the following statistics: project $\mathbf{y}_i$ onto the direction $\mathbf{w}$, remove the furthest (from original) $n - \hat{t}$ samples, and then compute the empirical variance of the remaining ones. Intuitively, the RVE provides an approximate measure of the variance (of authentic samples) captured by a candidate direction.

The main algorithm of HR-PCA is as given below. Note that as input it takes an upper bound on the number of corrupted points.

We remark that while computing the covariance matrix as well as removing points are performed over $\hat{\mathcal{Y}}$, computing RVE $\overline{V}_{\hat{t}}(\mathbf{w}_j)$ is performed over the original data-set $\mathcal{Y}$. This is to ensure that each candidate direction is measured correctly, even if some authentic points get removed in the process of the algorithm.

There are three parameters for HR-PCA, namely $\hat{d}$, $\hat{t}$ and $\overline{T}$, which we explain below.

- The parameter $\overline{T}$ does not affect the performance as long as it is large enough, namely, one can take $\overline{T} = n-1$. Interestingly, the algorithm is indeed an "any-time algorithm", i.e., one can stop the algorithm at any time, and the algorithm reports the best solution so far.
- As mentioned above, $(n - \hat{t})$ is an upper bound on the number of corrupted points, thus any value $\hat{t} \in (1/2, t]$ yields nontrivial guarantees. However, these guarantees improve the smaller we make $(t - \hat{t})$, which is to say that a better knowledge of how many corrupted points to expect, results in improved solutions. We note that tuning $\hat{t}$ is computationally simple, as it is possible to generate the solutions for multiple values of $\hat{t}$ in a single run of the algorithm.
- Tuning the parameter $\hat{d}$ is inherent to any PCA approach, with outliers or otherwise. Sometimes the choice of parameter $\hat{d}$ is known, where as others we may need to estimate, or search for it, thresholding the incremental change in variance captured. As we see from the performance guarantees of the algorithm, the success of the algorithm is not affected even if $\hat{d}$ is not perfectly tuned.

Intuition on Why The Algorithm Works: On any given iteration, we select candidate directions based on standard PCA – thus directions chosen are those with largest empirical variance. Now, given candidate directions $\mathbf{w}_1, \dots, \mathbf{w}_{\hat{d}}$, our robust variance estimator measures the variance of the $(n-\hat{t})$-smallest points projected in those directions. If this is large, it means that many of the points have a large variance in this direction – the points contributing to the robust variance estimator, and the points that led to this direction being selected by PCA. If the robust variance estimator is small, it is likely that a number of the largest variance points are corrupted, and thus removing one of them randomly, in proportion to their distance in the directions $\mathbf{w}_1, \dots, \mathbf{w}_{\hat{d}}$, results in the removal of a corrupted point.

Thus in summary, the algorithm works for the following intuitive reason. If the corrupted points have a very high variance along a direction with large angle from the span of the principal components, then with some probability, our algorithm removes them. If they have a high variance in a direction "close to" the span of the principal components, then this can only help in finding the principal components. Finally, if the corrupted points do not have a large variance, they may well survive the random removal process, but then the distortion they can cause in the output of PCA is necessarily limited.

The remainder of the paper makes this intuition precise, providing lower bounds on the probability of removing corrupted points, and subsequently upper bounds on the maximum distortion the corrupted points can cause.

Before finishing this subsection, we remark that an equally appealing idea would be to remove the largest point along

[2]Recall that maximizing this directly is the idea behind projection pursuit.

Algorithm 1 HR-PCA

Input: Contaminated sample-set $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \subset \mathbb{R}^p$, $\hat{d}$, $\overline{T}$, $\hat{t}$.
Output: $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_{\hat{d}}$.
Algorithm:

1) Let $\hat{\mathbf{y}}_i := \mathbf{y}_i$ for $i = 1, \ldots n$; $\hat{\mathcal{Y}} := \{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_n\}$; $s := 0$; $\mathrm{Opt} := 0$.
2) While $s \leq \overline{T}$, do
   a) Compute the empirical variance matrix
   $$\hat{\Sigma} := \frac{1}{n-s} \sum_{i=1}^{n-s} \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i^\top.$$
   b) Perform PCA on $\hat{\Sigma}$. Let $\mathbf{w}_1, \ldots, \mathbf{w}_{\hat{d}}$ be the $\hat{d}$ principal components of $\hat{\Sigma}$.
   c) If $\sum_{j=1}^{\hat{d}} \overline{V}_{\hat{t}}(\mathbf{w}_j) > \mathrm{Opt}$, then let $\mathrm{Opt} := \sum_{j=1}^{\hat{d}} \overline{V}_{\hat{t}}(\mathbf{w}_j)$ and let $\overline{\mathbf{w}}_j := \mathbf{w}_j$ for $j = 1, \cdots, \hat{d}$.
   d) Randomly remove a point from $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s}$ according to
   $$\Pr(\hat{\mathbf{y}}_i \text{ is removed from } \hat{\mathcal{Y}}) \propto \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^\top \hat{\mathbf{y}}_i)^2;$$
   e) Denote the remaining points by $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s-1}$;
   f) $s := s + 1$.
3) Output $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_{\hat{d}}$. End.

the project direction. However, this method may break under adversarial outliers in the sense that even the direction found in an iteration is completely wrong, the adversary can select corrupted points so that the algorithm still removes an authentic sample. Examples illustrating this are not hard to design.

C. Performance Guarantees: Fixed Design

We consider first the setting where the authentic points are arbitrary. The performance measure, as always, is the variance captured by the principal components we output. The performance is judged compared to the optimal output. As discussed above, in the fixed design setting, this optimal performance is a function of all the points. In particular, we want to give lower bounds on the quantity: $\sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\overline{\mathbf{w}}_j^\top \mathbf{z}_i)^2$. To do this, we also require a measure of the concentration of the authentic points, which essentially determines something akin to identifiability. Consider, for instance, the setting where all but a few of the authentic points are at the origin. Then the few remaining authentic points may indeed have a large variance along some direction; however, given the nature of our corruption, this direction is unidentifiable as the authentic points contributing to this variance are essentially indistinguishable from the corrupted points. The theorem below gives guarantees that are a function of just such a notion of concentration (or spread) of the authentic points. This is given by the functions φ^+ and φ^- defined in the theorem.

Theorem 1 (Fixed Design): Let $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_{\hat{d}}$ denote the output of the HR-PCA algorithm, and denote the optimal $\hat{d}$ Principal Components of $\mathbf{z}_1, \ldots, \mathbf{z}_t$ as $\mathbf{w}_1^*, \ldots, \mathbf{w}_{\hat{d}}^*$. Let $\varphi^-(\cdot)$ and $\varphi^+(\cdot)$ be any functions that satisfy the following: for any $t' \leq t$, $\mathbf{w} \in \mathbb{R}^p$ with $\|\mathbf{w}\|_2 = 1$,

$$\begin{aligned} \varphi^-(t'/t) \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{z}_i)^2 \quad &\leq \quad \sum_{i=1}^{t'} (\mathbf{w}^\top \mathbf{z}_{(i)})^2 \\ &\leq \varphi^+(t'/t) \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{z}_i)^2. \end{aligned}$$

Here, the middle term is the empirical variance of the smallest t' projections of the authentic points in the direction $\mathbf{w}$. Then, for any $\kappa > 0$, with high probability,

$$\begin{aligned} &\varphi^-\Big(\frac{t - s_0(\kappa)}{t}\Big) \varphi^-\Big(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\Big) \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\overline{\mathbf{w}}_j^\top \mathbf{z}_i)^2 \\ &\leq (1+\kappa) \varphi^+\Big(\frac{\hat{t}}{t}\Big) \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^{*\top} \mathbf{z}_i)^2, \end{aligned}$$

where there exists a universal constant C such that

$$\begin{aligned} s_0(\kappa)/t \quad &\leq \quad \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)} + \frac{C(1+\kappa)^2 \log n}{\kappa^2 n} \\ &\quad + \frac{C(1+\kappa)^{3/2} (\log n)^{1/2}}{\kappa^{3/2} n^{1/2}}. \end{aligned}$$

The parameter κ is introduced in the proof, and it is implicitly optimized by the algorithm. It controls the tradeoff between the fraction of the total variance in a particular direction captured by the authentic vs. the corrupted points, and the probability that a corrupted point is removed in the random removal (Step 2 d.) of the algorithm.

D. Performance Guarantees: Stochastic Design

In the stochastic design setup, it is possible to further simplify terms in Theorem 1, and in particular functions $\varphi^-(\cdot)$ and $\varphi^+(\cdot)$. This leads to the main contribution of this paper: performance guarantees of the stochastic design, which we

discuss in detail in this subsection. In the stochastic design case, we can compare any solution to the ideal solution, namely, the top $\hat{d}$ singular vectors of the matrix A. Note that while we allow $\hat{d} \geq d$, the most interesting case is $\hat{d} \leq d$. Thus, we seek a collection of orthogonal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{\hat{d}}$, that maximize the performance metric called the *Expressed Variance*:

$$\mathrm{EV}_{\hat{d}}(\mathbf{w}_1, \cdots, \mathbf{w}_{\hat{d}}) \triangleq \frac{\sum_{j=1}^{\hat{d}} \mathbf{w}_j^\top A A^\top \mathbf{w}_j}{\sum_{j=1}^{\hat{d}} \mathbf{w}_j^{*\top} A A^\top \mathbf{w}_j^*},$$

where $\mathbf{w}_1^*, \ldots, \mathbf{w}_{\hat{d}}^*$ are the $\hat{d}$ leading principal components of A, equivalently, the top $\hat{d}$ leading eigenvectors of $AA^\top$.[3] Note that unlike the fixed design setting, the quality of any solution is judged in terms of the ideal solution, and is not a function of the actual realization of the authentic points.

The Expressed Variance represents the portion of signal $A\mathbf{x}$ being expressed by $\mathbf{w}_1, \ldots, \mathbf{w}_{\hat{d}}$ compared to the optimal solution. The EV is always less than one, with equality achieved when the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{\hat{d}}$ have the span of the true principal components $\mathbf{w}_1^*, \ldots, \mathbf{w}_{\hat{d}}^*$. Notice that when $\hat{d} \geq d$, the denominator equals $\mathrm{trace}(AA^\top)$.

If Expressed Variance equals 1, this represents perfect recovery. Expressed variance bounded away from zero indicates a solution has a non-trivial performance bound. We show below that HR-PCA produces a solution with expressed variance bounded away from zero for all values of λ up to 50% (i.e., up to 50% corrupted points) and has expressed variance equal to one, i.e., perfect recovery, when the number of corrupted points scales more slowly than the number of points. In contrast, we do not know of any other algorithm that can guarantee a positive expressed variance for *any positive value of* λ.

The performance of HR-PCA directly depends on λ, the fraction of corrupted points. In addition, it depends on the distribution μ of $\mathbf{x}$ (more precisely, $\overline{\mu}$, as we allow d itself to go infinity). If $\overline{\mu}$ has longer tails, outliers that affect the variance (and hence are far from the origin) and authentic samples in the tail of the distribution, become more difficult to distinguish. To quantify this effect, we need the following "tail weight" function.

Definition 1: For any $\gamma \in [0,1]$, let $\delta_\gamma \triangleq \min\{\delta \geq 0 | \overline{\mu}([-\delta, \delta]) \geq \gamma\}$, $\gamma^- \triangleq \overline{\mu}((-\delta_\gamma, \delta_\gamma))$. Then the "tail weight" function $\mathcal{V} : [0,\, 1] \to [0,\, 1]$ is defined as follows

$$\mathcal{V}(\gamma) \triangleq \lim_{\epsilon \downarrow 0} \int_{-\delta_\gamma + \epsilon}^{\delta_\gamma - \epsilon} x^2 \overline{\mu}(dx) + (\gamma - \gamma^-)\delta_\gamma^2.$$

In words, $\mathcal{V}(\cdot)$ represents the contribution to its variance by the smallest γ fraction of the distribution. Hence $1 - \mathcal{V}(\cdot)$ represents how the tail of $\overline{\mu}$ contributes to its variance. Notice that $\mathcal{V}(0) = 0$, and $\mathcal{V}(1) = 1$. Furthermore $\mathcal{V}(0.5) > 0$ since $\overline{\mu}(\{0\}) < 0.5$. At a high level, controlling this is similar to the role of the φ functions in the deterministic setting.

We now provide bounds on the performance of HR-PCA for both the finite-sample and asymptotic case. Both bounds are functions of λ and the function $\mathcal{V}(\cdot)$.

[3]In case $\hat{d} > d$, $\mathbf{w}_1^*, \ldots, \mathbf{w}_{\hat{d}}^*$ are be the d Principal Components of A, and any $\hat{d} - d$ orthnormal unit vectors.

Theorem 2 (Finite Sample Performance): As we have done above, let $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_{\hat{d}}$ denote the output of HR-PCA, and $\mathbf{w}_1^*, \ldots, \mathbf{w}_{\hat{d}}^*$ denote the top $\hat{d}$ singular vectors of A. Let $\tau \triangleq \max(p/n, 1)$. Then, there exist absolute constants c and C, such that with high probability, the following holds for any κ:

$$\begin{aligned}\mathrm{EV}_{\hat{d}}(\overline{\mathbf{w}}_1, \cdots, \overline{\mathbf{w}}_{\hat{d}}) \geq\ & \frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(1 - \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ & - \frac{10}{\mathcal{V}(0.5)}\left(\frac{c\hat{d}\tau}{\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^{*\top} A\|_2^2}\right)^{\frac{1}{2}} \\ & - \frac{C\{\alpha^{\frac{1}{2}} d^{\frac{1}{4}} (\log^{\frac{5}{4}} n) n^{-\frac{1}{4}} \vee \alpha[(1+\kappa)/\kappa]^{\frac{3}{2}} (\log^{\frac{5}{2}} n) n^{-\frac{1}{2}}\}}{\mathcal{V}(0.5)}.\end{aligned} \tag{2}$$

As in the fixed design case, the parameter κ is implicitly optimized by the algorithm; here as well, it controls the trade-off between the fraction of the total variance in a particular direction captured by the authentic vs. the corrupted points, and the probability that a corrupted point is removed in the random removal (Step 2 d.) of the algorithm.

Remark 2: We briefly explain how variations of the specifics in Setup 2 may affect the results promised in Theorem 2. The following results can be obtained essentially by a similar argument as that presented in the proof of Theorem 2.

- The assumption that the noise $\mathbf{v}_i$ follows a Gaussian distribution can be relaxed; if the noise is sub-Gaussian, Theorem 2 still holds, with the only difference being the constant c, which then depends on the sub-Gaussian norm of the noise.
- The log terms in the last term of Equation 2 can be improved if $\overline{\mu}$ is assumed to be sub-Gaussian.
- As mentioned above, the assumption of spherical symmetry of μ is non-trivial. In the absence of spherical symmetry, the theorem holds with some modifications. When μ is not spherically symmetric, we may have different tail-weight functions in different directions. Thus, using $\overline{\mu}_{\mathbf{v}}$ to denote the 1-d marginal along direction $\mathbf{v} \in \mathcal{S}_d$, let $\mathcal{V}_{\mathbf{v}}(\cdot)$ denote the corresponding "tail weight" function of $\overline{\mu}_{\mathbf{v}}$. Define $\mathcal{V}^+(\gamma) \triangleq \sup_{\mathbf{v} \in \mathcal{S}_d} \mathcal{V}_{\mathbf{v}}(\gamma)$ and $\mathcal{V}^-(\gamma) \triangleq \inf_{\mathbf{v} \in \mathcal{S}_d} \mathcal{V}_{\mathbf{v}}(\gamma)$. Then, with essentially unchanged algorithm and proof, we obtain the following for the non-symmetric case:

$$\begin{aligned}\mathrm{EV}_{\hat{d}} \geq\ & \frac{\mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\mathcal{V}^-\left(1 - \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)}{(1+\kappa)\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)} \\ & - \frac{10}{\mathcal{V}^-(0.5)}\left(\frac{c\hat{d}\tau}{\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^{*\top} A\|_2^2}\right)^{\frac{1}{2}} \\ & - \frac{C\{\alpha^{\frac{1}{2}} d^{\frac{1}{4}} (\log^{\frac{5}{4}} n) n^{-\frac{1}{4}} \vee \alpha[(1+\kappa)/\kappa]^{\frac{3}{2}} (\log^{\frac{5}{2}} n) n^{-\frac{1}{2}}\}}{\mathcal{V}^-(0.5)}.\end{aligned}$$

As an essentially immediate corollary of the above theorem, we can obtain asymptotic guarantees for the performance of HR-PCA, in the scaling regime defined above. In particular, if we have τ, κ and $\overline{\mu}$ fixed, then the right-hand-side of

Equation (2) is non-trivial as long as $\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^{*\top} A\|_2^2/\hat{d} \to \infty$ and $n/(d\log^5 d) \to \infty$. In this case, the last two terms go to zero as n goes to infinity, producing the following asymptotic performance guarantees.

Theorem 3 (Asymptotic Performance): Consider a sequence of $\{\mathcal{Y}(j)\}$, where the asymptotic scaling in Expression (1) holds, $\lambda^* \triangleq \limsup \lambda(j)$, and again, $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_{\hat{d}}$ are the output of HR-PCA. Then the following holds in probability when $j \uparrow \infty$ (i.e., when $n, p \uparrow \infty$),

$$\liminf_j \mathrm{EV}_{\hat{d}}\{\overline{\mathbf{w}}_1(j), \ldots, \overline{\mathbf{w}}_{\hat{d}}(j)\} \geq \max_{\kappa} \left[\frac{\mathcal{V}\left(1 - \frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\right)}{(1+\kappa)}\right] \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right]. \quad (3)$$

Remark 3: The bounds in the two bracketed terms in the asymptotic bound may be, roughly, explained as follows. The first term is due to the fact that the removal procedure may well not remove all large-magnitude corrupted points, while at the same time, some authentic points may be removed. The second term accounts for the fact that not all the outliers may have large magnitude. These will likely not be removed, and will have some (controlled) effect on the principal component directions reported in the output. Another interesting interpretation of this is as follows: the second term is the performance bound for the (non-convex) projection pursuit algorithm using trimmed variance (our RVE), while the first bound can be regarded as the approximation factor incurred by our randomized algorithm.

We have made two claims in particular about the performance of HR-PCA: It is asymptotically optimal when the number of outliers scales sublinearly, and it is maximally robust with a breakdown point of 50%, the best possible for any algorithm. These results are implied by the next two corollaries.

For small λ, we can make use of the light tail condition on $\overline{\mu}$, to establish the following bound that simplifies (3). The proof is deferred to Appendix D.

Corollary 1: Under the settings of the above theorem, the following holds in probability when $j \uparrow \infty$ (i.e., when $n, p \uparrow \infty$),

$$\begin{aligned}\liminf_j \mathrm{EV}_{\hat{d}}\{\overline{\mathbf{w}}_1(j), \ldots, \overline{\mathbf{w}}_{\hat{d}}(j)\} &\geq \max_{\kappa}\left[1 - \kappa - \frac{C\alpha\lambda^*\log^2(1/\lambda^*)}{\kappa\mathcal{V}(0.5)}\right] \\ &\geq 1 - \frac{C'\sqrt{\alpha\lambda^*}\log(1/\lambda^*)}{\mathcal{V}(0.5)}.\end{aligned}$$

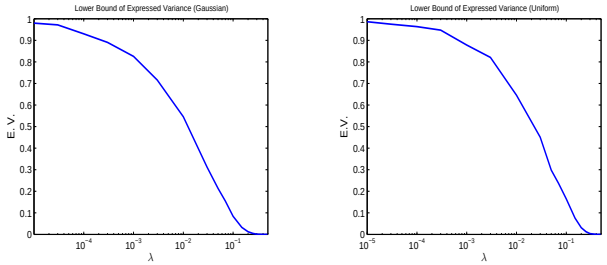
Remark 4: Thus indeed, if $(n-t) = o(n)$, i.e., the number of outliers scales sub linearly and hence f $\lambda(j) \downarrow 0$ then Corollary 1 shows that the expressed variance converges to 1, i.e., HR-PCA is asymptotically optimal. This is in contrast to PCA, where the existence of *even a single* corrupted point is sufficient to bound the output *arbitrarily* away from the optimum.

Next we show that that HR-PCA has a breakdown point of 50%. Recall that the Break-down point is defined as the fraction of (malicious) outliers required to change the output of a statistical algorithm arbitrarily. In the context of PCA, it measures the fraction of outliers required to make the output orthogonal to the desired subspace, or equivalently to make the expressed variance of the output zero. The next corollary shows that the expressed variance of HR-PCA stays strictly positive as long as $\lambda < 0.5$. Therefore, the breakdown point of HR-PCA converges to 50%, and hence HR-PCA achieves the maximal possible break-down point (a breakdown point greater than 50% is never possible.)

Corollary 2: Suppose $\overline{\mu}(\{0\}) = 0$. Then, under the same assumptions as the above theorem, as long as $\lambda^* < 0.5$, the sequence of outputs of HR-PCA, denotes $\{\overline{\mathbf{w}}_1(j), \ldots, \overline{\mathbf{w}}_{\hat{d}}(j)\}$, satisfy the following in probability:

$$\liminf_j \mathrm{EV}_{\hat{d}}\{\overline{\mathbf{w}}_1(j), \ldots, \overline{\mathbf{w}}_{\hat{d}}(j)\} > 0.$$

The graphs in Figure 1 illustrate the lower-bounds of asymptotic performance when the 1-dimensional marginal of μ is the Gaussian distribution (Figure (a)) or the Uniform distribution (Figure (b)).



(a) Gaussian distribution (b) Uniform distribution

Fig. 1. This figure shows the lower bounds on the asymptotic performance of HR-PCA, under Gaussian and Uniform distribution for $\mathbf{x}$.

IV. Kernelization

We consider kernelizing HR-PCA in this section: given a feature mapping $\Upsilon(\cdot) : \mathbb{R}^p \to \mathcal{H}$ equipped with a kernel function $k(\cdot,\cdot)$, i.e., $\langle \Upsilon(\mathbf{a}), \Upsilon(\mathbf{b})\rangle = k(\mathbf{a},\mathbf{b})$ holds for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, we perform the dimensionality reduction in the feature space $\mathcal{H}$ without knowing the explicit form of $\Upsilon(\cdot)$.

We assume that $\{\Upsilon(\mathbf{y}_1), \cdots, \Upsilon(\mathbf{y}_n)\}$ is centered at origin without loss of generality, since we can center any $\Upsilon(\cdot)$ with the following feature mapping

$$\hat{\Upsilon}(\mathbf{x}) \triangleq \Upsilon(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n}\Upsilon(\mathbf{y}_i),$$

whose kernel function is

$$\begin{aligned}\hat{k}(\mathbf{a},\mathbf{b}) &= k(\mathbf{a},\mathbf{b}) - \frac{1}{n}\sum_{j=1}^{n}k(\mathbf{a},\mathbf{y}_j) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{y}_i,\mathbf{b}) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(\mathbf{y}_i,\mathbf{y}_j).\end{aligned}$$

Notice that HR-PCA involves finding a set of PCs $\mathbf{w}_1, \ldots, \mathbf{w}_d \in \mathcal{H}$, and evaluating $\langle \mathbf{w}_q, \Upsilon(\cdot)\rangle$ (Note that RVE is a function of $\langle \mathbf{w}_q, \Upsilon(\mathbf{y}_i)\rangle$, and random removal depends on $\langle \mathbf{w}_q, \Upsilon(\hat{\mathbf{y}}_i)\rangle$). The former can be kernelized by applying

Kernel PCA introduced by [39], where each of the output PCs admits a representation

$$\mathbf{w}_q = \sum_{j=1}^{n-s} \alpha_j(q) \Upsilon(\hat{\mathbf{y}}_j).$$

Thus, $\langle \mathbf{w}_q,\, \Upsilon(\cdot) \rangle$ is easily evaluated by

$$\langle \mathbf{w}_q,\, \Upsilon(\mathbf{v}) \rangle = \sum_{j=1}^{n-s} \alpha_j(q) k(\hat{\mathbf{y}}_j, \mathbf{v}); \quad \forall \mathbf{v} \in \mathbb{R}^p$$

Therefore, HR-PCA is kernelizable since both steps are easily kernelized and we have the following Kernel HR-PCA.

Here, the kernelized RVE is defined as

$$\begin{aligned} \overline{V}_{\hat{t}}(\boldsymbol{\alpha}) &\triangleq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \left[\left| \langle \sum_{j=1}^{n-s} \alpha_j \Upsilon(\hat{\mathbf{y}}_j), \Upsilon(\mathbf{y}) \rangle \right|_{(i)} \right]^2 \\ &= \frac{1}{\hat{t}} \sum_{i=1}^{\hat{n}} \left[\left| \sum_{j=1}^{n-s} \alpha_j k(\hat{\mathbf{y}}_j, \mathbf{y}) \right|_{(i)} \right]^2 . \end{aligned}$$

V. Numerical Illustrations

In this section we illustrate the performance of HR-PCA via numerical results on synthetic data. The main purpose is twofold: to show that the performance of HR-PCA is as claimed in the theorems and corollaries above, and to compare its performance with standard PCA, and several popular robust PCA algorithms, namely, Multi-Variate iterative Trimming (MVT), ROBPCA proposed in [18], and the (approximate) Project-Pursuit (PP) algorithm proposed in [30]. Our numerical examples illustrate, in particular, how the properties of the high-dimensional regime discussed in Section II can degrade, or even completely destroy, the performance of available robust PCA algorithms.

We report the $d = 1$ case first. We randomly generate a $p \times 1$ matrix and scale it so that its leading eigenvalue has magnitude equal to a given σ. A λ fraction of outliers are generated on a line with a uniform distribution over $[-\sigma \cdot \mathrm{mag}, \sigma \cdot \mathrm{mag}]$. Thus, mag represents the ratio between the magnitude of the outliers and that of the signal $A\mathbf{x}_i$. For each parameter setup, we report the average result of 20 tests (and the 90% confidence interval of the mean) . The MVT algorithm breaks down in the $n = p$ case since it involves taking the inverse of the covariance matrix which is ill-conditioned. Hence we do not report MVT results in any of the experiments with $n = p$, as shown in Figure 2 and perform a separate test for MVT, HR-PCA and PCA under the case that $p \ll n$ reported in Figure 4.

We make the following three observations from Figure 2. First, PP and ROBPCA can break down when λ is large, while on the other hand, the performance of HR-PCA is rather robust even when λ is as large as 40%. Second, the performance of PP and ROBPCA depends strongly on σ, i.e., the signal magnitude (and hence the magnitude of the corrupted points). Indeed, when σ is very large, ROBPCA achieves effectively optimal recovery of the A subspace. However, the performance of both algorithms is not satisfactory when σ is small, and sometimes even worse than the performance of standard PCA. Finally, and perhaps most importantly, the performance of PP and ROBPCA degrades as the dimensionality increases, which makes them essentially not suitable for the high-dimensional regime we consider here. This is more explicitly shown in Figure 3 where the performance of different algorithms versus dimensionality is reported. We notice that the performance of ROBPCA (and similarly other algorithms based on Stahel-Donoho outlyingness) has a sharp decrease at a certain threshold that corresponds to the dimensionality where S-D outlyingness becomes invalid in identifying outliers.

Figure 4 shows that the performance of MVT depends on the dimensionality m. Indeed, the breakdown property of MVT is roughly $1/p$ as predicted by the theoretical analysis, which makes MVT less attractive in the high-dimensional regime.

A similar numerical study for $d = 3$ is also performed, where the outliers are generated on 3 random chosen lines. The results are reported in Figure 5. The same trends as in the $d = 1$ case are observed, although the performance gap between different strategies are smaller, because the effect of outliers are decreased since they are on 3 directions.

While this paper was under review, two new robust PCA methods based on the decomposition of a matrix into the sum of a low-rank matrix (via nuclear norm) and an "error" matrix have been proposed. In particular, in [40] the authors proposed the *RPCA* method in which the error is modeled as a sparse matrix, and in [36] the authors proposed the so-called *Outlier Pursuit* method in which the error is modeled as a column-sparse matrix. The first method (RPCA) is not designed to deal with the kind of corruption we have here, but rather considers the setting where each point is corrupted in a few coordinates. Nevertheless, we compare its performance empirically.

Under the same setup as Figure 4, we compare the proposed method with these two methods. In addition, to demonstrate that HRPCA is resilient to the parameter selection, we also report the performance of HRPCA where $\hat{t}$ is fixed to be $0.5n$ regardless of the fraction of the outliers (labeled HRPCA(0.5) in the figure). Figure 6 and 7 report the simulation results for $d = 1$ and $d = 3$ respectively. We make the following three observations: (i) The performance of HRPCA and HRPCA(0.5) is essentially the same, demonstrating that HRPCA is resilient to parameter selection; (ii) RPCA and Outlier Pursuit perform well for small λ, but break down when λ becomes larger. This is well expected, and has been observed in previous studies [36], [40]; (iii) The performance of RPCA and Outlier Pursuit degrades significantly when σ becomes small (equivalently, when the noise becomes large). This is not surprising – as we discussed in Section II, one drawback of these methods is that their performance scales unfavorably with the magnitude of the noise.

VI. Proof of the Main Result

In this section we provide the main steps of the proof of the finite-sample and asymptotic performance bounds, including the precise statements and the key ideas in the proof, but deferring some of the more standard or tedious elements to the appendix. The proof consists of four main steps.

1) We begin with the fixed-design setup, i.e., no assumptions on the authentic points $\{\mathbf{z}_i\}$ are made. The first

Algorithm 2 Kernel HR-PCA

Input: Contaminated sample-set $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \subset \mathbb{R}^p$, $\hat{d}$, T, $\hat{n}$.
Output: $\overline{\boldsymbol{\alpha}}(1), \ldots, \overline{\boldsymbol{\alpha}}(\hat{d})$.
Algorithm:

1) Let $\hat{\mathbf{y}}_i := \mathbf{y}_i$ for $i = 1, \ldots n$; $s := 0$; Opt $:= 0$.
2) While $s \leq T$, do
   a) Compute the Gram matrix of $\{\hat{\mathbf{y}}_i\}$:
   $$K_{ij} := k(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j); \quad i, j = 1, \cdots, n-s.$$
   b) Let $\hat{\sigma}_1^2, \cdots, \hat{\sigma}_{\hat{d}}^2$ and $\hat{\boldsymbol{\alpha}}(1), \cdots, \hat{\boldsymbol{\alpha}}(\hat{d})$ be the $\hat{d}$ largest eigenvalues and the corresponding eigenvectors of K.
   c) Normalize: $\boldsymbol{\alpha}(q) := \hat{\boldsymbol{\alpha}}(q)/\hat{\sigma}_q$, so that $\langle \mathbf{w}_q, \mathbf{w}_q \rangle = 1$.
   d) If $\sum_{q=1}^{\hat{d}} \overline{V}_{\hat{t}}(\boldsymbol{\alpha}(q)) > \text{Opt}$, then let $\text{Opt} := \sum_{q=1}^{\hat{d}} \overline{V}_{\hat{t}}(\boldsymbol{\alpha}(q))$ and let $\overline{\boldsymbol{\alpha}}(q) := \boldsymbol{\alpha}(q)$ for $q = 1, \cdots, \hat{d}$.
   e) Randomly remove a point from $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s}$ according to
   $$\Pr(\hat{\mathbf{y}}_i \text{ is removed}) \propto \sum_{q=1}^{\hat{d}} (\sum_{j=1}^{n-s} \alpha_j(q) k(\hat{\mathbf{y}}_j, \hat{\mathbf{y}}_i))^2;$$
   f) Denote the remaining points by $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s-1}$;
   g) $s := s + 1$.
3) Output $\overline{\boldsymbol{\alpha}}(1), \ldots, \overline{\boldsymbol{\alpha}}(\hat{d})$. End.

step shows that with high probability, the algorithm finds a "good" solution within a bounded number of steps. In particular, this involves showing that if in a given step the algorithm has not found a good solution, in the sense that the variance along a principal component is not mainly due to the authentic points, then the random removal scheme removes a corrupted point with probability bounded away from zero. We then use martingale arguments to show that as a consequence of this, there cannot be many steps with the algorithm finding at least one "good" solution, since in the absence of good solutions, most of the corrupted points are removed by the algorithm.

2) The previous step shows the existence of a "good" solution. The second step shows two things: first, that this good solution has performance that is close to that of the optimal solution, and second, that the final output of the algorithm is close to that of the "good" solution. Combining them together, we derive a performance guarantee for the fixed design case, i.e., for any $\{\mathbf{z}_i\}_{i=1}^t$.
3) From the third step onwards, we turn to the stochastic design case. When $\{\mathbf{z}_i\}_{i=1}^t$ are generated according to Setup 2, we can derive more interpretable results than the fixed design case. In order to achieve that, we prove in this step that RVE is a valid variance estimator with high probability.
4) We then combine results from previous steps, and simplify the expressions, to derive the finite-sample bound.

In what follows, letters c, C and their variants are reserved for absolute constants, whose value may change from line to line.

A. Step 1

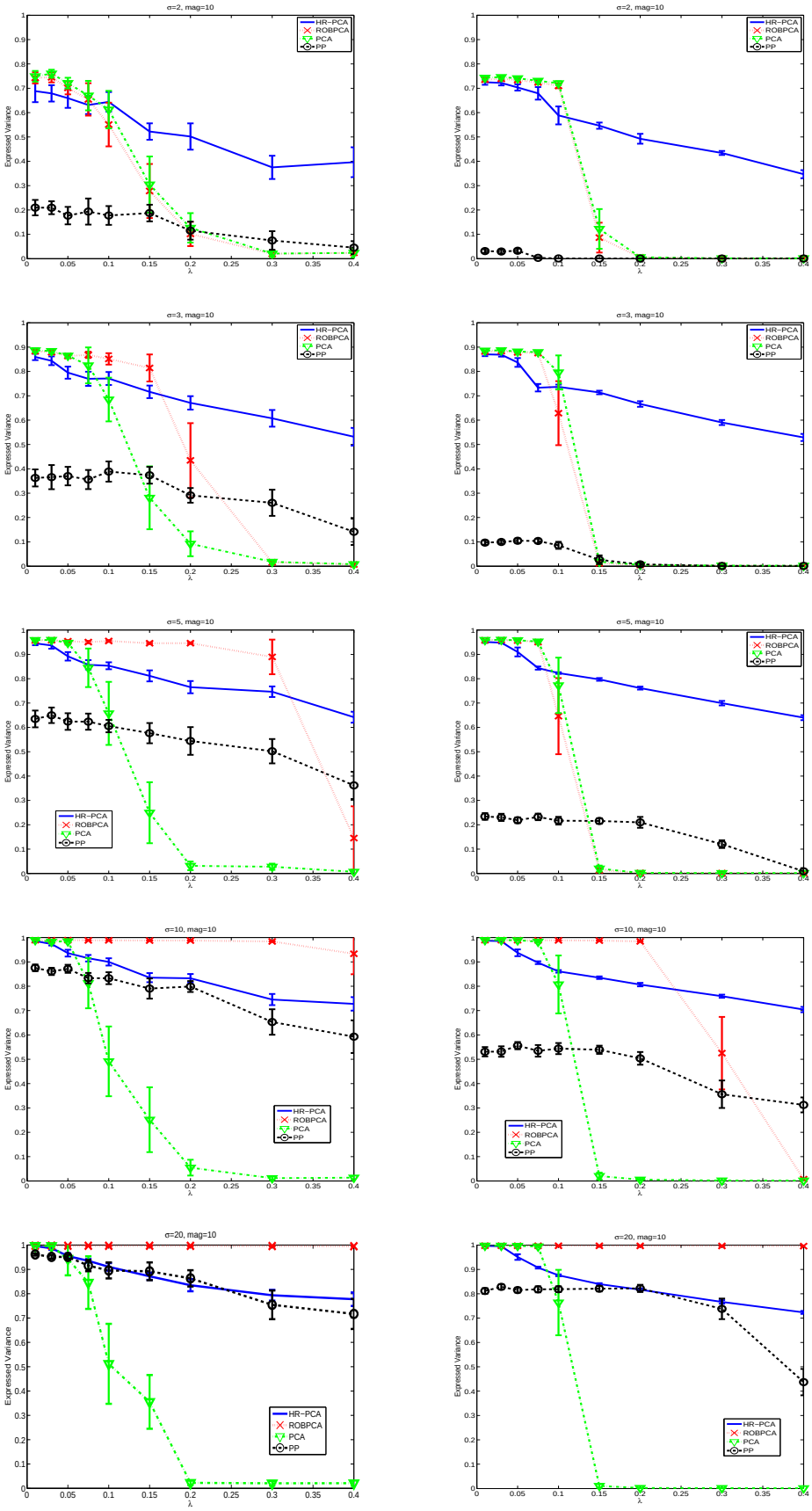
The first step shows that the algorithm finds a good solution in a small number of steps. Proving this involves showing that at any given step, either the algorithm finds a good solution, or the random removal eliminates one of the corrupted points with a guaranteed probability (i.e., probability bounded away from zero). The intuition then, is that there cannot be too many steps without finding a good solution, since too many of the corrupted points will have been removed. This section makes this intuition precise.

Let us fix a $\kappa > 0$. Let $\mathcal{Z}(s)$ and $\mathcal{O}(s)$ be the set of remaining authentic samples and the set of remaining corrupted points after the s^{th} stage, respectively. Then with this notation, the set of remaining points is $\mathcal{Y}(s) = \mathcal{Z}(s) \bigcup \mathcal{O}(s)$. Observe that $|\mathcal{Y}(s)| = n - s$. Let $\overline{r}(s) = \mathcal{Y}(s-1) \backslash \mathcal{Y}(s)$, the point removed at stage s. Let $\mathbf{w}_1(s), \ldots, \mathbf{w}_{\hat{d}}(s)$ be the $\hat{d}$ PCs found in the s^{th} stage — these points are the output of standard PCA on $\mathcal{Y}(s-1)$. These points are a good solution if the variance of the points projected onto their span is mainly due to the authentic samples rather than the corrupted points. We denote this "good output event at step s" by $\mathcal{E}(s)$, defined as follows:

$$\mathcal{E}(s) = \{\sum_{j=1}^{\hat{d}} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 \geq \frac{1}{\kappa} \sum_{j=1}^{\hat{d}} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2\}.$$

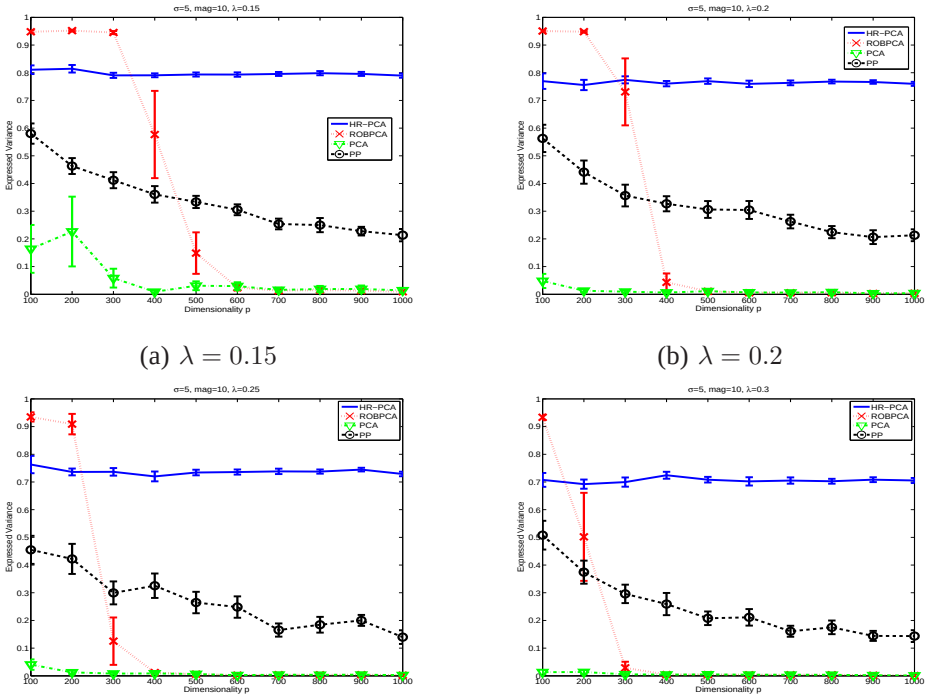
We show in the next theorem that with high probability, $\mathcal{E}(s)$ is true for at least one "small" s, by showing that at every s where it is not true, the random removal procedure removes a corrupted point with probability at least $\kappa/(1+\kappa)$.

Theorem 4: With high probability event $\mathcal{E}_\kappa(s)$ is true for



(a) $n = p = 100$ (b) $n = p = 1000$

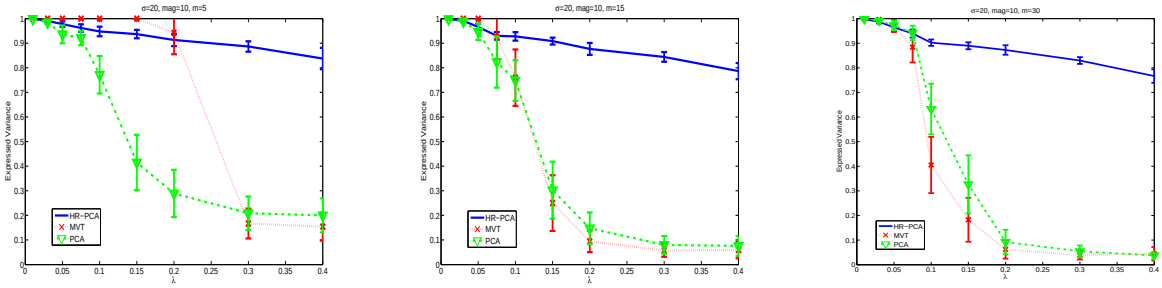
Fig. 2. Performance of HR-PCA vs ROBPCA, PP, PCA ($d = 1$).



(a) $\lambda = 0.15$ (b) $\lambda = 0.2$

(c) $\lambda = 0.25$ (d) $\lambda = 0.3$

Fig. 3. Performance vs dimensionality.



(a) $n = 100, p = 5$ (b) $n = 100, p = 15$ (c) $n = 100, p = 30$

Fig. 4. Performance of HR-PCA vs MVT for $p \ll n$.

some $1 \le s \le s_0(\kappa)$, where

$$s_0(\kappa) \triangleq n \wedge \left\{(1+\epsilon)\frac{(1+\kappa)\lambda n}{\kappa}\right\};$$
$$\epsilon = C\left\{\frac{(1+\kappa)\log n}{\kappa\lambda n} + \sqrt{\frac{(1+\kappa)\log n}{\kappa\lambda n}}\right\}.$$

In this step, the κ is fixed, hence we will simply write s_0 and $\mathcal{E}(s)$ to lighten the notation.

Remark 5: Divide s_0 by t leads to (notice $n \ge t = (1-\lambda)n \ge 0.5n$, and hence t and n are of same order)

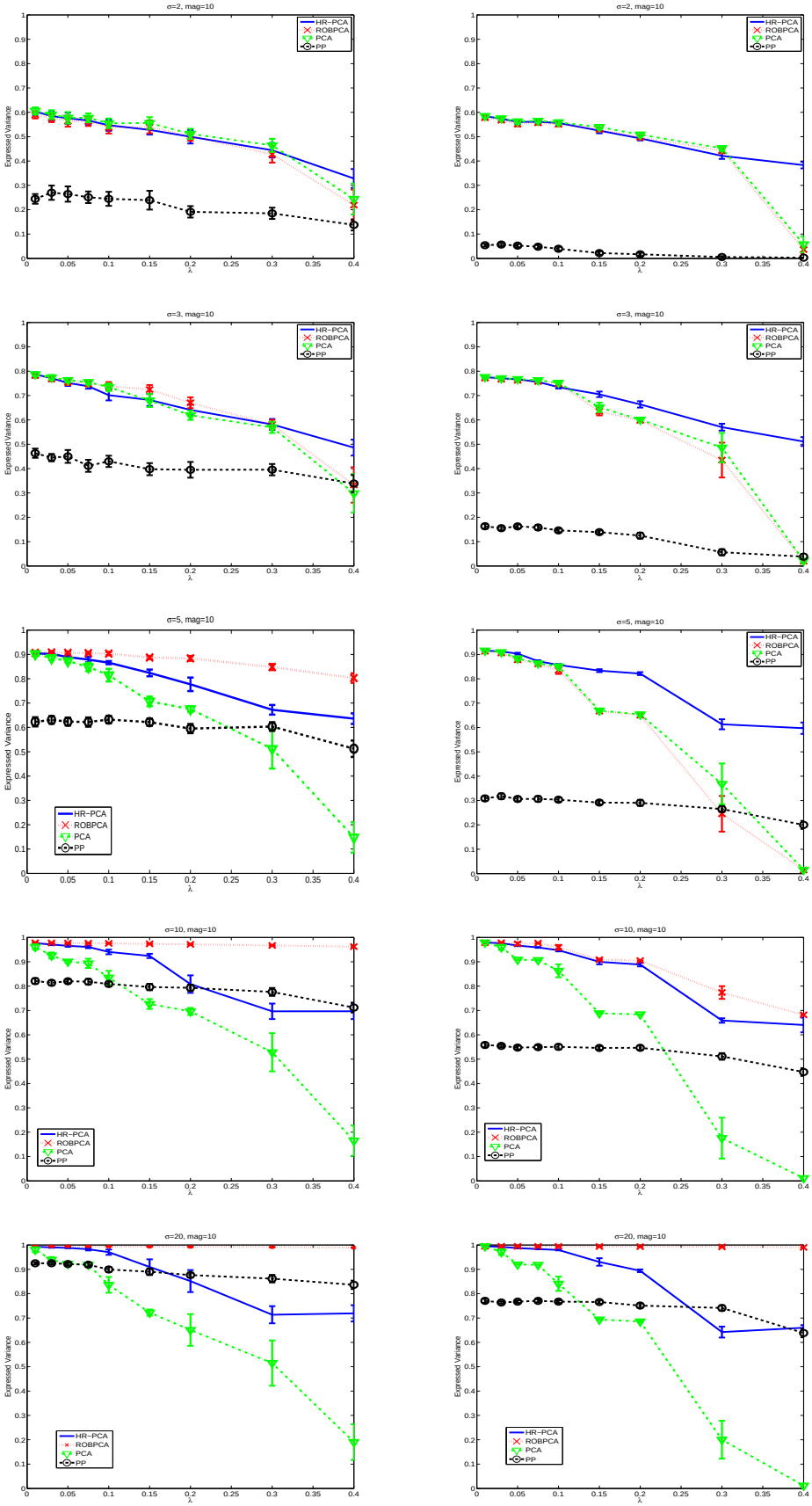
$$s_0(\kappa)/t \le \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)} + \frac{C(1+\kappa)^2\log n}{\kappa^2 n} + \frac{C(1+\kappa)^{3/2}(\log n)^{1/2}}{\kappa^{3/2}n^{1/2}}.$$

Notice that when $(1+\kappa)^3\log n/(\kappa^3 n) < 1$, then the second term is dominated by the third term; on the other hand, if $(1+\kappa)^3\log n/(\kappa^3 n) \ge 1$, then $s_0(\kappa) \le n \le 2t$ implies $s_0(\kappa)/t \le C'(1+\kappa)^{3/2}(\log n)^{1/2}/[\kappa^{3/2}n^{1/2}]$, thus we have

$$s_0(\kappa)/t \le \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)} + \frac{C'(1+\kappa)^{3/2}(\log n)^{1/2}}{\kappa^{3/2}n^{1/2}} \stackrel{\text{def}}{=} \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)} + \varepsilon_\kappa. \quad (4)$$

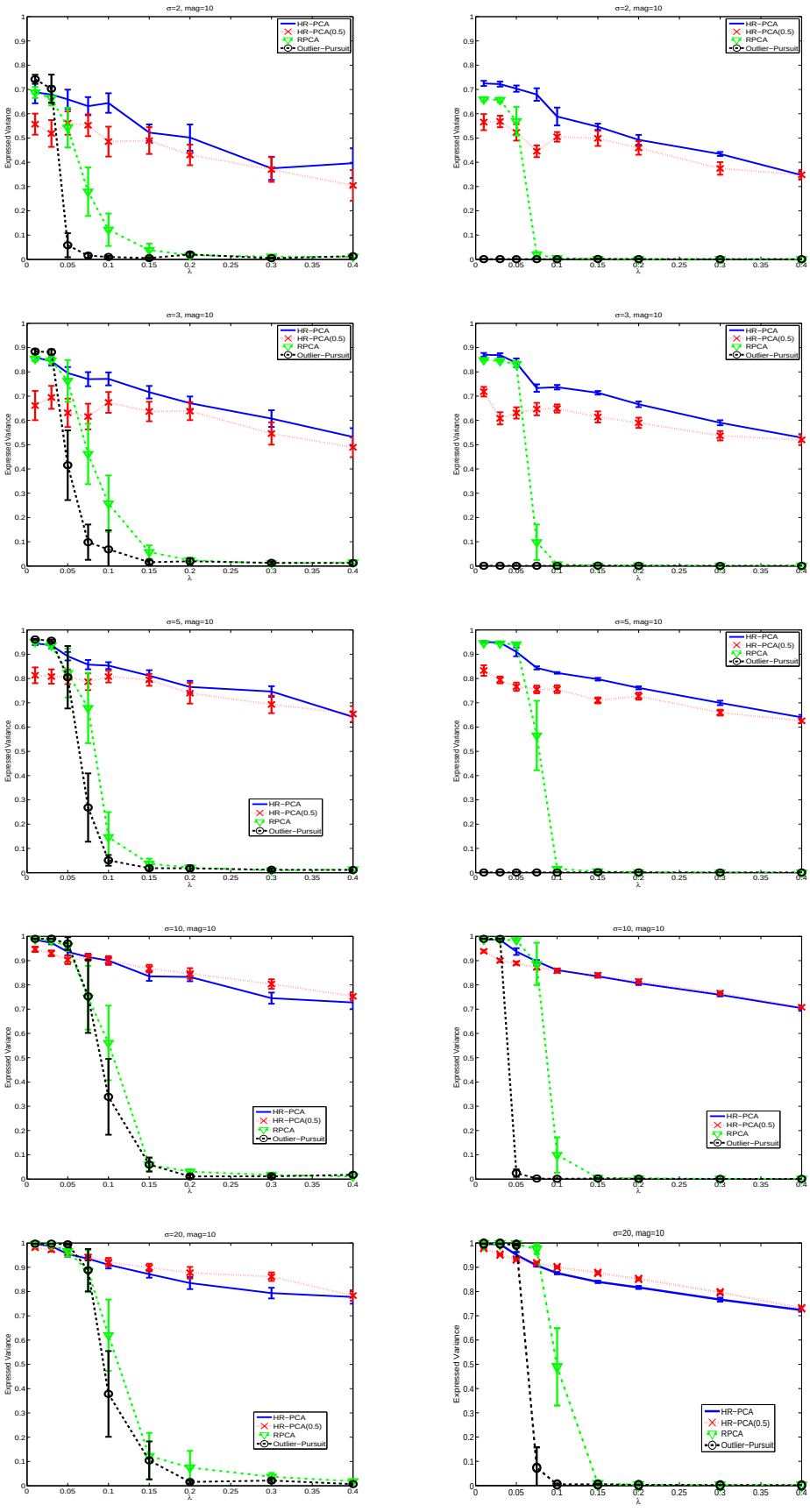
The right hand side of Equation (4) converges to $(1+\kappa)\lambda/\kappa(1-\lambda)$ for any fixed κ (indeed, for any sequence of κ_n such that $\kappa_n \sim \omega(\log n/n)^{1/3}$). Therefore, $s_0 \le t$ if $(1+\kappa)\lambda < \kappa(1-\lambda)$ and n is large.

When $s_0 = n$, Theorem 4 holds trivially. Hence we focus on the case where $s_0 < n$. En route to proving this theorem, we first prove that when $\mathcal{E}(s)$ is not true, our procedure removes a corrupted point with high probability. To this end, let $\mathcal{F}_s$ be the filtration generated by the set of events until stage s. Observe that $\mathcal{O}(s), \mathcal{Z}(s), \mathcal{Y}(s) \in \mathcal{F}_s$. Furthermore, since given



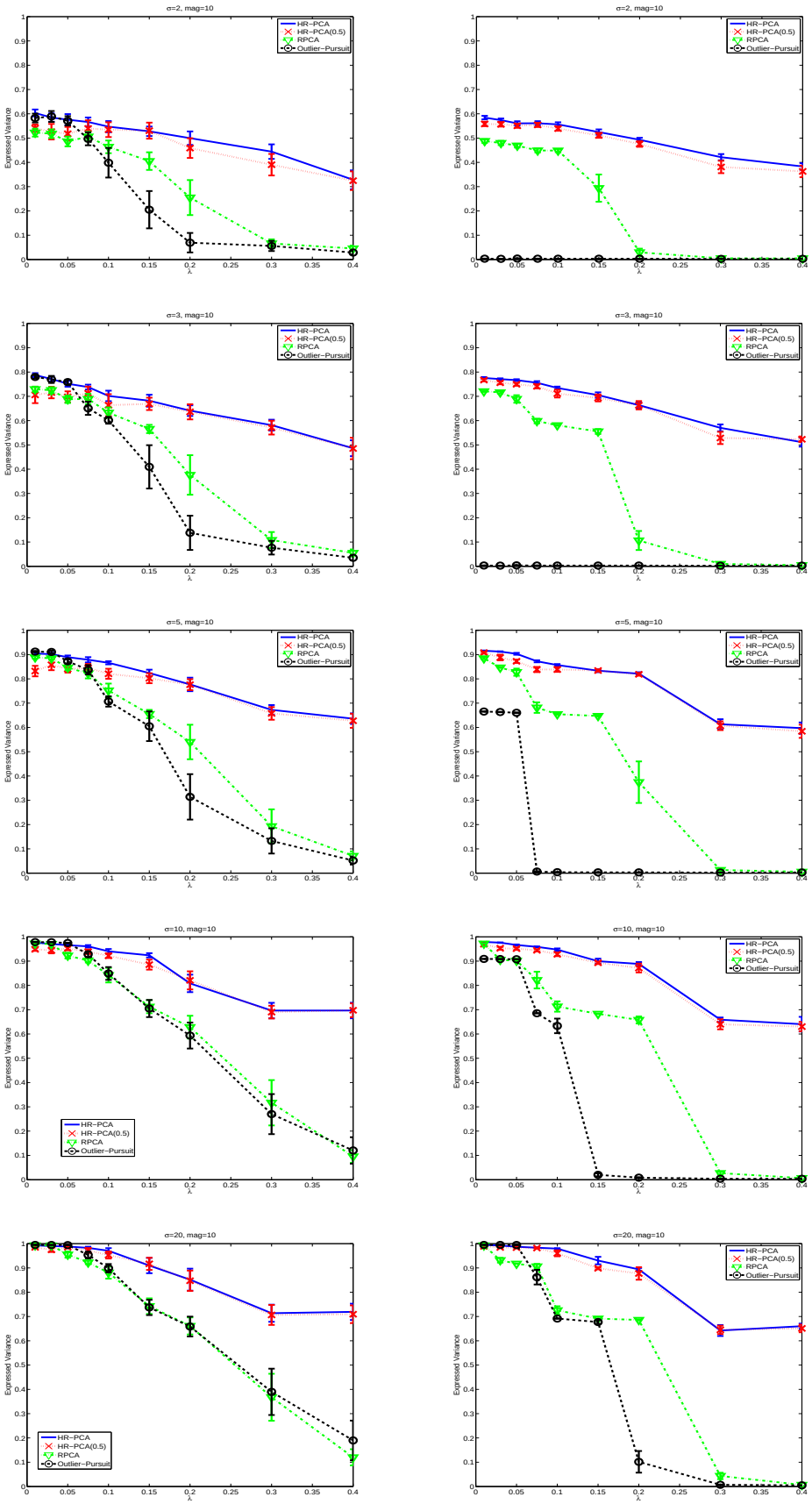
(a) $n = p = 100$ (b) $n = p = 1000$

Fig. 5. Performance of HR-PCA vs ROBPCA, PP, PCA ($d = 3$).



(a) $n = p = 100$ (b) $n = p = 1000$

Fig. 6. Performance of HR-PCA vs HR-PCA(0.5), RPCA, Outlier Pursuit ($d = 1$).



(a) $n = p = 100$ (b) $n = p = 1000$

Fig. 7. Performance of HR-PCA vs HR-PCA(0.5), RPCA, Outlier Pursuit ($d = 1$).

$\mathcal{Y}(s)$, performing a PCA is deterministic, $\mathcal{E}(s+1) \in \mathcal{F}_s$.

Theorem 5: If $\mathcal{E}^c(s)$ is true, then

$$\Pr(\{\overline{r}(s) \in \mathcal{O}(s-1)\}|\mathcal{F}_{s-1}) > \frac{\kappa}{1+\kappa}.$$

Proof: If $\mathcal{E}^c(s)$ is true, then

$$\sum_{j=1}^{\hat{d}} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 < \frac{1}{\kappa} \sum_{j=1}^{\hat{d}} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2,$$

which is equivalent to

$$\sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 + \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2 < \frac{1+\kappa}{\kappa} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2.$$

Note that

$$\begin{aligned} &\Pr(\{\overline{r}(s) \in \mathcal{O}(s-1)\}|\mathcal{F}_{s-1}) \\ =& \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} \Pr(\overline{r}(s) = \mathbf{o}_i|\mathcal{F}_{s-1}) \\ =& \frac{\sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2}{\sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 + \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2} \\ >& \frac{\kappa}{1+\kappa}. \end{aligned}$$

Here, the second equality follows from the definition of the algorithm, and in particular, that in stage s, we remove a point $\mathbf{y}$ with probability proportional to $\sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{y})^2$, and independent to other events. ∎

As a consequence of this theorem, we can now prove Theorem 4. The intuition is rather straightforward: if the events were independent from one step to the next, then since "the expected number of corrupted points removed" is at least $\kappa/(1+\kappa)$, then after $s_0 = (1+\epsilon)(1+\kappa)\lambda n/\kappa$ steps, with exponentially high probability all the outliers would be removed, and hence we would have a good event with high probability, for some $s \le s_0$. Since subsequent steps are not independent, we have to rely on martingale arguments.

Let $T = \min\{s|\mathcal{E}(s) \text{ is true}\}$. Note that since $\mathcal{E}(s) \in \mathcal{F}_{s-1}$, we have $\{T > s\} \in \mathcal{F}_{s-1}$. Define the following random variable

$$X_s = \begin{cases} |\mathcal{O}(T-1)| + \frac{\kappa(T-1)}{1+\kappa}, & \text{if } T \le s; \\ |\mathcal{O}(s)| + \frac{\kappa s}{1+\kappa}, & \text{if } T > s. \end{cases}$$

Lemma 1: $\{X_s, \mathcal{F}_s\}$ is a supermartingale.

Proof Sketch: The proof essentially follows from the definition of X_s, and the fact that if $\mathcal{E}(s)$ is true, then $|\mathcal{O}(s)|$ decreases by one with probability $\kappa/(1+\kappa)$. The full details are deferred to the appendix. ∎

From here, the proof of Theorem 4 follows fairly quickly.

Proof Sketch: Note that

$$\begin{aligned} &\Pr\left(\bigcap_{s=1}^{s_0} \mathcal{E}(s)^c\right) = \Pr\left(T > s_0\right) \\ \le& \Pr\left(X_{s_0} \ge \frac{\kappa s_0}{1+\kappa}\right) = \Pr\left(X_{s_0} \ge (1+\epsilon)\lambda n\right), \end{aligned} \tag{5}$$

where the inequality is due to $|\mathcal{O}(s)|$ being non-negative. Recall that $X_0 = \lambda n$. Thus the probability that no good events occur before step s_0 is at most the probability that a supermartingale with bounded incremements increases in value by a constant factor of $(1+\epsilon)$, from λn to $(1+\epsilon)\lambda n$. An appeal to Azuma's inequality shows that this is exponentially unlikely. The details are left to the appendix. ∎

B. Step 2

Theorem 6 (Fixed Design): The following three statements hold for the fixed design case:

1) For any $\kappa > 0$ such that $s_0(\kappa) < n$, with high probability there exists $s \le s_0(\kappa)$, such that

$$\frac{1}{1+\kappa} \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t-s_0(\kappa)} \left|\mathbf{w}_j^* \mathbf{z}\right|_{(i)}^2 \le \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2. \tag{6}$$

2) For any $s \le n$,

$$\sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t}-\frac{\lambda t}{1-\lambda}} |\mathbf{w}_j(s)^\top \mathbf{z}|_{(i)}^2 \le \sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t}} |\overline{\mathbf{w}}_j^\top \mathbf{z}|_{(i)}^2. \tag{7}$$

3) Let $\varphi^-(\cdot)$ and $\varphi^+(\cdot)$ satisfy for any $t' \le t$, $\mathbf{w} \in \mathbb{R}^p$ with $\|\mathbf{w}\|_2 = 1$,

$$\begin{aligned} \varphi^-(t'/t) \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{z}_i)^2 \le& \sum_{i=1}^{t'} |\mathbf{w}^\top \mathbf{z}_i|^2 \\ &\le \varphi^+(t'/t) \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{z}_i)^2, \end{aligned}$$

then with high probability,

$$\begin{aligned} &\varphi^-\Big(\frac{t-s_0(\kappa)}{t}\Big)\varphi^-\Big(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\Big) \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^{*\top} \mathbf{z}_i)^2 \\ \le& (1+\kappa)\varphi^+\Big(\frac{\hat{t}}{t}\Big) \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\overline{\mathbf{w}}_j^\top \mathbf{z}_i)^2. \end{aligned}$$

Proof: **Part 1:** With high probability, there exists $s \le s_0(\kappa)$ such that $\mathcal{E}_\kappa(s)$ is true. Then we have

$$\sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 \ge \frac{1}{\kappa} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2.$$

Recall that $\mathcal{Y}(s-1) = \mathcal{Z}(s-1) \bigcup \mathcal{O}(s-1)$, and that $\mathcal{Z}(s-1)$ and $\mathcal{O}(s-1)$ are disjoint. We thus have

$$\begin{aligned} &\frac{1}{1+\kappa} \sum_{\mathbf{y}_i \in \mathcal{Y}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{y}_i)^2 \\ &\le \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2. \end{aligned} \tag{8}$$

Since $\mathbf{w}_1(s), \ldots, \mathbf{w}_{\hat{d}}(s)$ are the solution of the s^{th} stage, the following holds by definition of the algorithm

$$\sum_{\mathbf{y}_i \in \mathcal{Y}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^{*\top} \mathbf{y}_i)^2 \leq \sum_{\mathbf{y}_i \in \mathcal{Y}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{y}_i)^2. \quad (9)$$

Further note that by $\mathcal{Z}(s-1) \subseteq \mathcal{Y}(s-1)$ and $\mathcal{Z}(s-1) \subseteq \mathcal{Z}$, we have

$$\sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^{*\top} \mathbf{z}_i)^2 \leq \sum_{\mathbf{y}_i \in \mathcal{Y}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^{*\top} \mathbf{y}_i)^2,$$

and

$$\begin{aligned}\sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 &\leq \sum_{\mathbf{z}_i \in \mathcal{Z}} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 \\ &= \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2.\end{aligned}$$

Substituting them into (8) and (9) we have

$$\frac{1}{1+\kappa} \sum_{j=1}^{\hat{d}} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j^{*\top} \mathbf{z}_i)^2 \leq \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2.$$

Note that $|\mathcal{Z}(s-1)| \geq t - (s-1) \geq t - s_0(\kappa)$, hence for all $j = 1, \cdots, d$,

$$\sum_{i=1}^{t-s_0} \left|\mathbf{w}_j^* \mathbf{z}\right|_{(i)}^2 \leq \sum_{i=1}^{|\mathcal{Z}(s-1)|} \left|\mathbf{w}_j^* \mathbf{z}\right|_{(i)}^2 \leq \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j^* \mathbf{z}_i)^2,$$

which in turn implies

$$\frac{1}{1+\kappa} \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t-s_0(\kappa)} \left|\mathbf{w}_j^* \mathbf{z}\right|_{(i)}^2 \leq \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2. \quad (10)$$

Part 2: The definition of algorithm implies that

$$\sum_{j=1}^{\hat{d}} \overline{V}_{\hat{t}}(\mathbf{w}_j(s)) \leq \sum_{j=1}^{\hat{d}} \overline{V}_{\hat{t}}(\overline{\mathbf{w}}_j).$$

Recall that $\overline{V}_{\hat{t}}(\mathbf{w}) = \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{y}|_{(i)}^2$, hence we have

$$\sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t}} |\mathbf{w}_j(s)^\top \mathbf{y}|_{(i)}^2 \leq \sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t}} |\overline{\mathbf{w}}_j^\top \mathbf{y}|_{(i)}^2. \quad (11)$$

Further notice that for any unit-norm $\mathbf{w} \in \mathbb{R}^p$, since $\mathcal{Z} \subset \mathcal{Y}$ and $|\mathcal{Y} \backslash \mathcal{Z}| = \lambda n = \lambda t/(1-\lambda)$, we have

$$\sum_{i=1}^{\hat{t} - \frac{\lambda t}{1-\lambda}} |\mathbf{w}^\top \mathbf{z}|_{(i)}^2 \leq \sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{y}|_{(i)}^2 \leq \sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{z}|_{(i)}^2.$$

Here, the first inequality holds because for any $\hat{t}$ elements in $\mathcal{Y}$, at least $\hat{t} - \lambda t/(1-\lambda)$ belongs to $\mathcal{Z}$; the second inequality holds because any subset of $\mathcal{Z}$ with $\hat{t}$ elements, is also a subset of $\mathcal{Y}$ with $\hat{t}$ elements, thus the inequality follows from the definition of order statistics (i.e., the smallest elements).

Substitute this into Equation (11), we have

$$\sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t} - \frac{\lambda t}{1-\lambda}} |\mathbf{w}_j(s)^\top \mathbf{z}|_{(i)}^2 \leq \sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t}} |\overline{\mathbf{w}}_j^\top \mathbf{z}|_{(i)}^2. \quad (12)$$

Part 3: By definition of $\varphi^+(\cdot)$ and $\varphi^-(\cdot)$, Equation (10) leads to

$$\begin{aligned}\varphi^-\big(\frac{t-s_0}{t}\big) \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\mathbf{w}_j^{*\top} \mathbf{z}_i)^2 \leq &\sum_{j=1}^{\hat{d}} \sum_{i=1}^{t-s_0} \left|\mathbf{w}_j^{*\top} \mathbf{z}\right|_{(i)}^2 \\ \leq &(1+\kappa) \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2.\end{aligned}$$

Similarly, Equation (12) leads to

$$\begin{aligned}&\varphi^-\big(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\big) \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 \\ &\leq \sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t} - \frac{\lambda t}{1-\lambda}} |\mathbf{w}_j(s)^\top \mathbf{z}|_{(i)}^2 \\ &\leq \sum_{j=1}^{\hat{d}} \sum_{i=1}^{\hat{t}} |\overline{\mathbf{w}}_j^\top \mathbf{z}|_{(i)}^2 \leq \varphi^+\big(\frac{\hat{t}}{t}\big) \sum_{j=1}^{\hat{d}} \sum_{i=1}^{t} (\overline{\mathbf{w}}_j^\top \mathbf{z}_i)^2.\end{aligned}$$

Combining these together, we have that

$$\begin{aligned}&\varphi^-\Big(\frac{t-s_0(\kappa)}{t}\Big) \varphi^-\Big(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\Big) \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\mathbf{w}_j^{*\top} \mathbf{z}_i)^2 \\ &\leq (1+\kappa) \varphi^+\Big(\frac{\hat{t}}{t}\Big) \sum_{i=1}^{t} \sum_{j=1}^{\hat{d}} (\overline{\mathbf{w}}_j^\top \mathbf{z}_i)^2.\end{aligned}$$

■

C. Step 3

From this step on we focus on the stochastic design case. Recall that in this case, the authentic samples $\{\mathbf{z}_i\}_{i=1}^t$ are generated according to $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{v}_i$ for i.i.d. $\mathbf{x}_i \in \mathbb{R}^d$, and Gaussian noise $\mathbf{v}_i \in \mathbb{R}^p$. Our main goal is this step is to show that for any $\mathbf{w}_1, \ldots, \mathbf{w}_{\hat{d}}$ and $t' \leq t$, $\sum_{j=1}^{\hat{d}} \sum_{i=1}^{t'} \left|\mathbf{w}_j \mathbf{z}\right|_{(i)}^2$ is a good indicator of $\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2$. Thus, combining with the result in Step 2 establishes Theorem 2. En route to this, we require following lemmas about the properties of $\mathcal{V}(\cdot)$. The proofs are deferred to Appendix B.

Lemma 2 (Monotonicity of $\mathcal{V}$): Given $0 \leq a_1 < a_2 < a_3 \leq 1$, we have

$$\frac{\mathcal{V}(a_2) - \mathcal{V}(a_1)}{a_2 - a_1} \leq \frac{\mathcal{V}(a_3) - \mathcal{V}(a_2)}{a_3 - a_2}.$$

Lemma 3: 1) For any $a \in [0, 1]$, we have

$$\mathcal{V}(a) \leq a.$$

2) For any $0 \leq a_1 < a_2 \leq 1$, we have

$$\mathcal{V}(a_2) - \mathcal{V}(a_1) \leq \frac{a_2 - a_1}{1 - a_1}.$$

Lemma 4: For any $\epsilon > 0$ and $\kappa \in [\epsilon, 1]$, we have $\mathcal{V}(\kappa) - \mathcal{V}(\kappa - \epsilon) \leq C\alpha\epsilon \log^2(1/\epsilon)$.

The rest of the this section depends the following concentration condition.

Condition 1: (I)

$$\sup_{\mathbf{w}\in\mathcal{S}_p}\frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top\mathbf{v}_i)^2\le c\tau.$$

(II)

$$\sup_{\mathbf{q}\in\mathcal{S}_d}\Big|\frac{1}{t}\sum_{i=1}^{t}(\mathbf{q}^\top\mathbf{x}_i)^2-1\Big|\le C\alpha\sqrt{\frac{d\log^3 n}{n}}\overset{\text{def}}{=}\varepsilon_0.$$

(III) Suppose $\varepsilon_0\le 1$. For all $\mathbf{q}\in\mathcal{S}_d$ and $\overline{t}\le t$,

$$\begin{aligned}&\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top\mathbf{x}]^2_{(i)}-\mathcal{V}(\overline{t}/t)\right|\\ \le\;&\frac{Ct(1+\varepsilon_0)\sqrt{d\log n/n}}{t-\overline{t}}\wedge C\alpha^{\frac{1}{2}}d^{\frac{1}{4}}(\log n)^{\frac{5}{4}}n^{-\frac{1}{4}}\\ \overset{\text{def}}{=}\;&\varepsilon_1(\overline{t}/t).\end{aligned}$$

Theorem 7: Condition 1 holds with high probability.

The proof of Theorem 7 is lengthy, and hence deferred to Appendix C. We are now ready to show the main result of this step.

Theorem 8: Suppose Condition 1 holds. Then for all $\mathbf{w}\in\mathcal{S}_p$, and $t'\le t$, the following holds:

$$\begin{aligned}&\|\mathbf{w}^\top A\|_2^2[\mathcal{V}(t'/t)-\varepsilon_1(t'/t)]-2\|\mathbf{w}^\top A\|_2\sqrt{(1+\varepsilon_0)c\tau}\\ \le\;&\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}|^2_{(i)}\\ \le\;&\|\mathbf{w}^\top A\|_2^2[\mathcal{V}(\frac{t'}{t})+\varepsilon_1(\frac{t'}{t})]+2\|\mathbf{w}^\top A\|_2\sqrt{(1+\varepsilon_0)c\tau}+c\tau.\end{aligned}$$

Proof: Recall that $\mathbf{z}_i=A\mathbf{x}_i+\mathbf{v}_i$. Fix an arbitrary $\mathbf{w}\in\mathcal{S}_p$. Let $\{\hat{\pi}_i\}_{i=1}^t$ and $\{\bar{\pi}_i\}_{i=1}^t$ be permutations of $[1,\dots,t]$ such that both $|\mathbf{w}^\top\mathbf{z}_{\hat{\pi}_i}|$ and $|\mathbf{w}^\top A\mathbf{x}_{\bar{\pi}_i}|$ are non-decreasing. Thus we have

$$\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}|^2_{(i)}=\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}_{\hat{\pi}_i}|^2\le\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}_{\bar{\pi}_i}|^2.$$

Expanding the right-hand-side yields

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}_{\bar{\pi}_i}|^2\\ =&\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top A\mathbf{x}_{\bar{\pi}_i}+\mathbf{w}^\top\mathbf{v}_{\bar{\pi}_i}|^2\\ \le&\frac{1}{t}\left\{\sum_{i=1}^{t'}(\mathbf{w}^\top A\mathbf{x}_{\bar{\pi}_i})^2+2\sum_{i=1}^{t}|\mathbf{w}^\top A\mathbf{x}_{\bar{\pi}_i}||\mathbf{w}^\top\mathbf{v}_{\bar{\pi}_i}|\right.\\ &\qquad\left.+\sum_{i=1}^{t}(\mathbf{w}^\top\mathbf{v}_{\bar{\pi}_i})^2\right\}\\ \overset{(a)}{=}&\frac{1}{t}\left\{\sum_{i=1}^{t'}|\mathbf{w}^\top A\mathbf{x}|^2_{(i)}+2\sum_{i=1}^{t}|\mathbf{w}^\top A\mathbf{x}_i||\mathbf{w}^\top\mathbf{v}_i|\right.\\ &\qquad\left.+\sum_{i=1}^{t}(\mathbf{w}^\top\mathbf{v}_i)^2\right\},\end{aligned}$$

where (a) holds due to the fact that $|\mathbf{w}^\top A\mathbf{x}_{\bar{\pi}_i}|$ are non-decreasing. We now bound three terms separately:

$$\begin{aligned}I.\;&\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top A\mathbf{x}|^2_{(i)}=\|\mathbf{w}^\top A\|_2^2\frac{1}{t}\sum_{i=1}^{t'}\Big|\frac{\mathbf{w}^\top A}{\|\mathbf{w}^\top A\|_2}\mathbf{x}\Big|^2_{(i)}\\ &\le\|\mathbf{w}^\top A\|_2^2\sup_{\mathbf{q}\in\mathcal{S}_d}\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{q}^\top\mathbf{x}|^2_{(i)}\\ &\le\|\mathbf{w}^\top A\|_2^2[\mathcal{V}(t'/t)+\varepsilon_1(t'/t)];\\ II.\;&\frac{2}{t}\sum_{i=1}^{t}|\mathbf{w}^\top A\mathbf{x}_i||\mathbf{w}^\top\mathbf{v}_i|\\ &\le 2\sqrt{\frac{1}{t}\sum_{i=1}^{t}|\mathbf{w}^\top A\mathbf{x}_i|^2}\sqrt{\frac{1}{t}\sum_{i=1}^{t}|\mathbf{w}^\top\mathbf{v}_i|^2}\\ &\le 2\|\mathbf{w}^\top A\|_2\sqrt{\sup_{\mathbf{q}\in\mathcal{S}_d,\|\mathbf{q}\|_2=1}\frac{1}{t}\sum_{i=1}^{t}|\mathbf{q}^\top\mathbf{x}_i|^2}\sqrt{\frac{1}{t}\sum_{i=1}^{t}|\mathbf{w}^\top\mathbf{v}_i|^2}\\ &\le 2\|\mathbf{w}^\top A\|_2^2\sqrt{1+\varepsilon_0}\sqrt{c\tau};\\ III.\;&\frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top\mathbf{v}_i)^2\le c\tau.\end{aligned}$$

We thus have

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}|^2_{(i)}\\ \le&\|\mathbf{w}^\top A\|_2^2[\mathcal{V}(t'/t)+\varepsilon_1(t'/t)]+2\|\mathbf{w}^\top A\|_2\sqrt{(1+\varepsilon_0)c\tau}+c\tau.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top\mathbf{z}|^2_{(i)}=\frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top A\mathbf{x}_{\hat{\pi}_i}+\mathbf{w}^\top\mathbf{v}_{\hat{j}_i}|^2\\ \ge&\frac{1}{t}\left\{\sum_{i=1}^{t'}(\mathbf{w}^\top A\mathbf{x}_{\hat{\pi}_i})^2-2\sum_{i=1}^{t}|\mathbf{w}^\top A\mathbf{x}_i||\mathbf{w}^\top\mathbf{v}_i|\right\}\\ \ge&\frac{1}{t}\left\{\sum_{i=1}^{t'}|\mathbf{w}^\top A\mathbf{x}|^2_{(i)}-2\sum_{i=1}^{t}|\mathbf{w}^\top A\mathbf{x}_i||\mathbf{w}^\top\mathbf{v}_i|\right\}\\ \ge&\|\mathbf{w}^\top A\|_2^2[\mathcal{V}(t'/t)-\varepsilon_1(t'/t)]-2\|\mathbf{w}^\top A\|_2\sqrt{(1+\varepsilon_0)c\tau}.\end{aligned}$$

■

The following corollary immediately follows from the fact that $\sum_{j=1}^{\hat{d}}|a_j|\le\sqrt{\hat{d}\sum_{j=1}^{\hat{d}}a_j^2}$ holds for any a_j.

Corollary 3: Suppose Condition 1 holds. Then for all

$\mathbf{w}_1,\dots,\mathbf{w}_{\hat{d}} \in \mathcal{S}_p$, and $t' \le t$, the following holds:

$$\begin{aligned}
&\big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big]\big[\mathcal{V}\big(\frac{t'}{t}\big) - \varepsilon_1\big(\frac{t'}{t}\big)\big] \\
&\qquad - 2\sqrt{\hat{d}\big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big](1+\varepsilon_0)c\tau} \\
&\le \frac{1}{t}\sum_{j=1}^{\hat{d}}\sum_{i=1}^{t'} |\mathbf{w}_j^\top \mathbf{z}|^2_{(i)} \\
&\le \big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big]\big[\mathcal{V}\big(\frac{t'}{t}\big) + \varepsilon_1\big(\frac{t'}{t}\big)\big] \\
&\qquad + 2\sqrt{\hat{d}\big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big](1+\varepsilon_0)c\tau} + c\hat{d}\tau.
\end{aligned}$$

In the special case where $t' = t$, we can indeed sharpen the result of Theorem 8, since in this case

$$\left|\frac{1}{t}\sum_{i=1}^{t'}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(t'/t)\right| = \big|\frac{1}{t}\sum_{i=1}^{t}(\mathbf{v}^\top \mathbf{x}_i)^2 - 1\big| \le \varepsilon_0.$$

This leads to the following corollary.

Corollary 4: Suppose Condition 1 holds. Then for all $\mathbf{w}_1,\dots,\mathbf{w}_{\hat{d}} \in \mathcal{S}_p$, the following holds:

$$\begin{aligned}
&\big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big][1-\varepsilon_0] - 2\sqrt{\hat{d}\big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big](1+\varepsilon_0)c\tau} \\
&\le \frac{1}{t}\sum_{j=1}^{\hat{d}}\sum_{i=1}^{t} |\mathbf{w}_j^\top \mathbf{z}_i|^2 \\
&\le \big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big][1+\varepsilon_0] \\
&\qquad + 2\sqrt{\hat{d}\big[\sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^\top A\|_2^2\big](1+\varepsilon_0)c\tau} + c\hat{d}\tau.
\end{aligned}$$

D. Step 4

Finally, based on all previous results, we prove the main theorem.

Theorem 2: Let the algorithm output be $\overline{\mathbf{w}}_1,\dots,\overline{\mathbf{w}}_{\hat{d}}$, and denote the optimal $\hat{d}$ Principal Components of A as $\mathbf{w}_1^*,\dots,\mathbf{w}_{\hat{d}}^*$. Denote $\tau \triangleq \max(p/n, 1)$ and

$$H^* \triangleq \sum_{j=1}^{\hat{d}} \|\mathbf{w}_j^{*\top} A\|_2^2; \qquad \overline{H} \triangleq \sum_{j=1}^{\hat{d}} \|\overline{\mathbf{w}}_j^\top A\|_2^2.$$

With high probability, the following holds for any κ,

$$\begin{aligned}
\frac{\overline{H}}{H^*} \ge\ & \frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(1 - \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)} - \frac{10}{\mathcal{V}(0.5)}\left(\frac{c\hat{d}\tau}{H^*}\right)^{1/2} \\
& - \frac{C\{\alpha^{\frac{1}{2}} d^{\frac{1}{4}}(\log^{\frac{5}{4}} n) n^{-\frac{1}{4}} \vee \alpha[(1+\kappa)/\kappa]^{\frac{3}{2}}(\log^{\frac{5}{2}} n) n^{-\frac{1}{2}}\}}{\mathcal{V}(0.5)}.
\end{aligned}$$

Here c and C are absolute constants.

Proof: Recall that with high probability Condition 1 and $\bigcup_{s=1}^{s_0(\kappa)} \mathcal{E}_\kappa(s)$ are both true. So we restrict our attention to this case. Further notice that we can assume $\varepsilon_0 \le 1$, $\varepsilon_1(\hat{t}/t) \le \mathcal{V}\left(\hat{t}/t - \lambda/(1-\lambda)\right)$ and $c\hat{d}\tau/H^* < 1$, since otherwise the theorem holds trivially as the right-hand-side of Equation (13) is negative.

Since $\bigcup_{s=1}^{s_0(\kappa)} \mathcal{E}_\kappa(s)$ is true, there exists a $s \le s_0$ such that $\mathcal{E}_\kappa(s)$ is true. To simplify notation, denote $H_s \triangleq \sum_{j=1}^{\hat{d}} \|\mathbf{w}_j(s)^\top A\|_2^2$. Theorem 6 leads to

$$\frac{1}{1+\kappa}\sum_{j=1}^{\hat{d}}\sum_{i=1}^{t-s_0(\kappa)} \big|\mathbf{w}_j^* \mathbf{z}\big|^2_{(i)} \le \sum_{j=1}^{\hat{d}}\sum_{i=1}^{t} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2.$$

Using Corollary 3 to lower bound the left-hand-side, and Corollary 4 to upper bound the right-hand-side, we have

$$\begin{aligned}
&\frac{1}{1+\kappa}\left[\Big(\mathcal{V}\big(\frac{t-s_0(\kappa)}{t}\big) - \varepsilon_1\big(\frac{\hat{t}}{t}\big)\Big)H^* - 2\sqrt{(1+\varepsilon_0)c\hat{d}H^*\tau}\right] \\
&\le (1+\varepsilon_0)H_s + 2\sqrt{(1+\varepsilon_0)c\hat{d}H_s\tau} + c\hat{d}\tau \\
&\le (1+\varepsilon_0)H_s + 2\sqrt{(1+\varepsilon_0)c\hat{d}H^*\tau} + c\hat{d}\tau,
\end{aligned}$$

where the last inequality holds because for any $\hat{d}$, $H_s \le H^*$. By re-organization, we have

$$\begin{aligned}
&\Big(\mathcal{V}\big(\frac{t-s_0(\kappa)}{t}\big) - \varepsilon_1\big(\frac{\hat{t}}{t}\big)\Big)H^* - (2\kappa+4)\sqrt{(1+\varepsilon_0)c\hat{d}H^*\tau} \\
&\qquad - (1+\kappa)c\hat{d}\tau \le (1+\kappa)(1+\varepsilon_0)H_s.
\end{aligned} \tag{13}$$

On the other hand, Theorem 6 also gives

$$\sum_{j=1}^{\hat{d}}\sum_{i=1}^{\hat{t}-\frac{\lambda t}{1-\lambda}} |\mathbf{w}_j(s)^\top \mathbf{z}|^2_{(i)} \le \sum_{j=1}^{\hat{d}}\sum_{i=1}^{\hat{t}} |\overline{\mathbf{w}}_j^\top \mathbf{z}|^2_{(i)},$$

which by applying Corollary 3 and 4 implies

$$\begin{aligned}
&\left[\mathcal{V}\big(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\big) - \varepsilon_1\big(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\big)\right] H_s - 2\sqrt{(1+\varepsilon_0)c\hat{d}\tau H_s} \\
&\le \left[\mathcal{V}\left(\frac{\hat{t}}{t}\right) + \varepsilon_1\left(\frac{\hat{t}}{t}\right)\right]\overline{H} + 2\sqrt{(1+\varepsilon_0)c\hat{d}\tau\overline{H}} + c\hat{d}\tau.
\end{aligned}$$

Notice that $\varepsilon_1(\cdot)$ is non-decreasing, and $H_s, \overline{H} \le H^*$, we can simplify the equation to the following one:

$$\begin{aligned}
&\left[\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right) - \varepsilon_1\left(\frac{\hat{t}}{t}\right)\right] H_s \\
&\le \left[\mathcal{V}\left(\frac{\hat{t}}{t}\right) + \varepsilon_1\left(\frac{\hat{t}}{t}\right)\right]\overline{H} + 4\sqrt{(1+\varepsilon_0)c\hat{d}\tau H^*} + c\hat{d}\tau.
\end{aligned} \tag{14}$$

Combining Equation (13) and (14), and notice that $\varepsilon_1(\hat{t}/t) \le$

$\mathcal{V}(\hat{t}/t-\lambda/(1-\lambda))$, we have

$$\begin{aligned}\frac{\overline{H}}{H^*} \geq & \frac{\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)\right]\left[\mathcal{V}\left(\frac{t-s_0}{t}\right)-\varepsilon_1(\hat{t}/t)\right]}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ & -\left\{\frac{(2\kappa+4)\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)\right]\sqrt{(1+\varepsilon_0)c\hat{d\tau}}}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]}\right. \\ & \left.+\frac{4(1+\kappa)(1+\varepsilon_0)\sqrt{(1+\varepsilon_0)c\hat{d\tau}}}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]}\right\}(H^*)^{-1/2} \\ & -\left\{\frac{\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)+1+\varepsilon_0\right]c\hat{d\tau}}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]}\right\}(H^*)^{-1}.\end{aligned} \tag{15}$$

Finally, we simplify the right hand side of Equation (15), by bounding the three terms separately:

$$\begin{aligned}&\frac{\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)\right]\left[\mathcal{V}\left(\frac{t-s_0}{t}\right)-\varepsilon_1(\hat{t}/t)\right]}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ \geq & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)-\varepsilon_1\left(\frac{\hat{t}}{t}\right)\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)+\mathcal{V}\left(\frac{t-s_0}{t}\right)\right]}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ \overset{(a)}{\geq} & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]}-\frac{2\varepsilon_1(\hat{t}/t)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ \overset{(b)}{\geq} & \frac{(1-\varepsilon_0)\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)}{(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]}-\frac{2\varepsilon_1(\hat{t}/t)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ \overset{(c)}{\geq} & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)}{(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]}-\frac{2\varepsilon_1(\hat{t}/t)+\varepsilon_0}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ \overset{(d)}{\geq} & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)-\varepsilon_1\left(\frac{\hat{t}}{t}\right)\right]}{(1+\kappa)\mathcal{V}^2\left(\frac{\hat{t}}{t}\right)}-\frac{2\varepsilon_1\left(\frac{\hat{t}}{t}\right)+\varepsilon_0}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ \overset{(e)}{\geq} & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)}-\frac{3\varepsilon_1(\hat{t}/t)+\varepsilon_0}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)};\end{aligned} \tag{16}$$

where (a) and (c) holds because $\mathcal{V}(\cdot)$ is upper bounded by 1; (b) and (d) follows from the fact that for any $0 \leq \epsilon < a$, $1/(a+\epsilon) \geq (a-\epsilon)/a^2$; (e) holds because $\mathcal{V}(\hat{t}/t-\lambda/(1-\lambda)) \leq \mathcal{V}(\hat{t}/t)$. Further recall from Equation (4) that

$$s_0(\kappa)/t \leq \frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}+\varepsilon_\kappa,$$

which by Lemma 4 leads to

$$\begin{aligned}\mathcal{V}\left(1-\frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)-\mathcal{V}\left(\frac{t-s_0}{t}\right) &\leq C\alpha\varepsilon_\kappa\log^2(1/\varepsilon_\kappa) \\ &\leq C\alpha\varepsilon_\kappa\log^2 n.\end{aligned}$$

Substitute into Equation (16) leads to

$$\begin{aligned}&\frac{\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)\right]\left[\mathcal{V}\left(\frac{t-s_0}{t}\right)-\varepsilon_1(\hat{t}/t)\right]}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ \geq & \frac{\mathcal{V}(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda})\mathcal{V}(1-\frac{(1+\kappa)\lambda}{\kappa(1-\lambda)})}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)}-\frac{3\varepsilon_1\left(\frac{\hat{t}}{t}\right)+\varepsilon_0+C\alpha(\log^2 n)\varepsilon_\kappa}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}.\end{aligned} \tag{17}$$

To bound the second term, we have

$$\begin{aligned}&\frac{(2\kappa+4)\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)\right]\sqrt{(1+\varepsilon_0)c\hat{d\tau}}}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ & +\frac{4(1+\kappa)(1+\varepsilon_0)\sqrt{(1+\varepsilon_0)c\hat{d\tau}}}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ \leq & \frac{(4\kappa+4)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)\right]\sqrt{1+\varepsilon_0}+4(1+\kappa)(1+\varepsilon_0)\sqrt{1+\varepsilon_0}}{(1+\varepsilon_0)(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ \leq & \frac{4\mathcal{V}\left(\frac{\hat{t}}{t}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}+\frac{4(1+\varepsilon_0)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \leq \frac{8}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}+C\frac{\varepsilon_0}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)};\end{aligned} \tag{18}$$

To bound the third term, we have

$$\begin{aligned}&\frac{\left[\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)-\varepsilon_1(\hat{t}/t)\right]+[1+\varepsilon_0]}{(1+\varepsilon_0)(1+\kappa)\left[\mathcal{V}\left(\frac{\hat{t}}{t}\right)+\varepsilon_1(\hat{t}/t)\right]} \\ \leq & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}+\frac{1}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \leq \frac{2}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}.\end{aligned} \tag{19}$$

Combining Equation (17), (18) and (19), we have

$$\begin{aligned}\frac{\overline{H}}{H^*} \geq & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(1-\frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ & -\frac{3\varepsilon_1(\hat{t}/t)+\varepsilon_0+C\alpha(\log^2 n)\varepsilon_\kappa}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ & -\frac{8}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\sqrt{\frac{c\hat{d\tau}}{H^*}}-\frac{C\varepsilon_0}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\sqrt{\frac{c\hat{d\tau}}{H^*}}-\frac{2}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\frac{c\hat{d\tau}}{H^*} \\ \geq & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(1-\frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)}-\frac{8}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\left(\frac{c\hat{d\tau}}{H^*}\right)^{\frac{1}{2}} \\ & -\frac{2}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\left(\frac{c\hat{d\tau}}{H^*}\right)-\frac{C[\varepsilon_0\vee\varepsilon_1\left(\frac{\hat{t}}{t}\right)\vee\alpha(\log^2 n)\varepsilon_\kappa]}{\mathcal{V}(0.5)} \\ \geq & \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(1-\frac{(1+\kappa)\lambda}{\kappa(1-\lambda)}\right)}{(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)}-\frac{10}{\mathcal{V}(0.5)}\left(\frac{c\hat{d\tau}}{H^*}\right)^{\frac{1}{2}} \\ & -\frac{C[\varepsilon_0\vee\varepsilon_1(\hat{t}/t)\vee\alpha(\log^2 n)\varepsilon_\kappa]}{\mathcal{V}(0.5)},\end{aligned}$$

where in the last two inequalities we use the fact that $\hat{t}/t \geq 0.5$ and $c\hat{d}\tau \leq H^*$. We can further simplify the last term by

$$\begin{aligned}
&\frac{C[\varepsilon_0 \vee \varepsilon_1(\hat{t}/t) \vee \alpha(\log^2 n)\varepsilon_\kappa]}{\mathcal{V}(0.5)} \\
\leq\ & \frac{C\alpha d^{\frac{1}{2}}(\log^{\frac{3}{2}} n)n^{-\frac{1}{2}}}{\mathcal{V}(0.5)} \vee \frac{C\alpha^{\frac{1}{2}} d^{\frac{1}{4}}(\log^{\frac{5}{4}} n)n^{-\frac{1}{4}}}{\mathcal{V}(0.5)} \\
& \vee \frac{C\alpha[(1+\kappa)/\kappa]^{\frac{3}{2}}(\log^{\frac{5}{2}} n)n^{-\frac{1}{2}}]}{\mathcal{V}(0.5)} \\
\leq\ & \frac{C[\alpha^{\frac{1}{2}} d^{\frac{1}{4}}(\log^{\frac{5}{4}} n)n^{-\frac{1}{4}} \vee \alpha[(1+\kappa)/\kappa]^{\frac{3}{2}}(\log^{\frac{5}{2}} n)n^{-\frac{1}{2}}]}{\mathcal{V}(0.5)},
\end{aligned}$$

where the last inequality holds since when $\alpha^{1/2}d^{1/4}(\log^{5/4} n)n^{-1/4} \leq 1$ (otherwise the theorem holds trivially), we have $\alpha d^{1/2}(\log^{3/2} n)n^{-1/2} \leq \alpha^{1/2}d^{1/4}(\log^{5/4} n)n^{-1/4}$. ∎

VII. Concluding Remarks

In this paper, we investigated the dimensionality-reduction problem in the case where the number and the dimensionality of samples are of the same magnitude, and a constant fraction of the points are arbitrarily corrupted (perhaps maliciously so). We proposed a High-dimensional Robust Principal Component Analysis algorithm that is tractable, robust to corrupted points, easily kernelizable and asymptotically optimal. The algorithm iteratively finds a set of PCs using standard PCA and subsequently remove a point randomly with a probability proportional to its expressed variance. We provided both theoretical guarantees and favorable simulation results about the performance of the proposed algorithm.

To the best of our knowledge, previous efforts to extend existing robust PCA algorithms into the high-dimensional case remain unsuccessful. Such algorithms are designed for low dimensional data sets where the observations significantly outnumber the variables of each dimension. When applied to high-dimensional data sets, they either lose statistical consistency due to lack of sufficient observations, or become highly intractable. This motivates our work of proposing a new robust PCA algorithm that takes into account the inherent difficulty in analyzing high-dimensional data.

Appendix

A. Proof of Theorem 4 and Lemma 1

Recall the statement of Theorem 4:

Theorem 4: With high probability $\bigcup_{s=1}^{s_0} \mathcal{E}_\kappa(s)$ is true. Here

$$\begin{aligned}
&s_0(\kappa) \triangleq (1+\epsilon)\frac{(1+\kappa)\lambda n}{\kappa}; \\
&\epsilon = C\left\{\frac{(1+\kappa)\log n}{\kappa\lambda n} + \sqrt{\frac{(1+\kappa)\log n}{\kappa\lambda n}}\right\}.
\end{aligned}$$

As κ is fixed, we will simply write $\mathcal{E}(s)$ and s_0 in the proof. Recall that we defined the random variable X_s as follows: Let $T = \min\{s|\mathcal{E}(s) \text{ is true}\}$. Note that since $\mathcal{E}(s) \in \mathcal{F}_{s-1}$, we have $\{T > s\}, \{T = s\}, \{T < s\} \in \mathcal{F}_{s-1}$. Then define:

$$X_s = \begin{cases} |\mathcal{O}(T-1)| + \frac{\kappa(T-1)}{1+\kappa}, & \text{if } T \leq s; \\ |\mathcal{O}(s)| + \frac{\kappa s}{1+\kappa}, & \text{if } T > s. \end{cases}$$

The proof of the above theorem depends on first showing that the random variable, X_s, is a supermartingale.

Lemma 1: $\{X_s, \mathcal{F}_s\}$ is a supermartingale.

Proof: Observe that $X_s \in \mathcal{F}_s$. We next show that $\mathbb{E}(X_s|\mathcal{F}_{s-1}) \leq X_{s-1}$ by enumerating the following three cases of $\mathcal{F}_{s-1}$ (recall $\{T > s\}, \{T = s\}, \{T < s\} \in \mathcal{F}_{s-1}$):

Case 1, $T > s$: Thus we have $\mathcal{E}^c(s)$ is true. By Theorem 5, under this situation,

$$\begin{aligned}
& \mathbb{E}(X_s - X_{s-1}|\mathcal{F}_{s-1}) \\
=\ & \mathbb{E}\left(\mathcal{O}(s) - \mathcal{O}(s-1) + \frac{\kappa}{1+\kappa}\Big|\mathcal{F}_{s-1}\right) \\
=\ & \frac{\kappa}{1+\kappa} - \Pr\left(\overline{r}(s) \in \mathcal{O}(s-1)\Big|\mathcal{F}_{s-1}\right) \\
=\ & \frac{\kappa}{1+\kappa} - \Pr\left(\overline{r}(s) \in \mathcal{O}(s-1)\right) \\
<\ & 0.
\end{aligned}$$

Case 2, $T = s$: By definition of X_s we have $X_s = \mathcal{O}(s-1) + \kappa(s-1)/(1+\kappa) = X_{s-1}$.

Case 3, $T < s$: Since both T and s are integer, we have $T \leq s-1$. Thus, $X_{s-1} = \mathcal{O}(T-1) + \kappa(T-1)/(1+\kappa) = X_s$.

These three cases enumerate all possible $\mathcal{F}_{s-1}$. Hence combining them together shows that $\mathbb{E}(X_s|\mathcal{F}_{s-1}) \leq X_{s-1}$, which proves the lemma. ∎

Next, we prove Theorem 4.

Proof: Note that

$$\begin{aligned}
&\Pr\left(\bigcap_{s=1}^{s_0} \mathcal{E}(s)^c\right) = \Pr\left(T > s_0\right) \\
\leq &\Pr\left(X_{s_0} \geq \frac{\kappa s_0}{1+\kappa}\right) = \Pr\left(X_{s_0} \geq (1+\epsilon)\lambda n\right),
\end{aligned} \tag{20}$$

where the inequality is due to $|\mathcal{O}(s)|$ being non-negative.

Let $y_s \triangleq X_s - X_{s-1}$, where recall that $X_0 = \lambda n$. Consider the following sequence:

$$y'_s \triangleq y_s - \mathbb{E}(y_s|y_1, \cdots, y_{s-1}).$$

Observe that $\{y'_s\}$ is a martingale difference process w.r.t. $\{\mathcal{F}_s\}$. Since $\{X_s\}$ is a supermartingale, $\mathbb{E}(y_s|y_1, \cdots, y_{s-1}) \leq 0$ a.s. Therefore, the following holds a.s.,

$$X_s - X_0 = \sum_{i=1}^{s} y_i = \sum_{i=1}^{s} y'_i + \sum_{i=1}^{s} \mathbb{E}(y_i|y_1, \cdots, y_{i-1}) \leq \sum_{i=1}^{s} y'_i. \tag{21}$$

By definition, $|y_s| \leq 1$, and hence $|y'_s| \leq 2$. Now apply Azuma's inequality,

$$\begin{aligned}
&\Pr(X_{s_0} \geq (1+\epsilon)\lambda n) \\
&\leq \Pr((\sum_{i=1}^{s_0} y'_i) \geq \epsilon\lambda n) \\
&\leq \exp(-(\epsilon\lambda n)^2/8s_0) \\
&= \exp\left(-\frac{(\epsilon\lambda n)^2\kappa}{8(1+\epsilon)(1+\kappa)\lambda n}\right) \\
&\leq \exp\left(-\frac{(\epsilon\lambda n)^2\kappa}{8(1+\epsilon)(1+\kappa)\lambda n}\right) \\
&\leq \max\left(\exp\left(-\frac{\epsilon^2\lambda n\kappa}{16(1+\kappa)}\right), \exp\left(-\frac{\epsilon\lambda n\kappa}{16(1+\kappa)}\right)\right).
\end{aligned}$$

Substituting ϵ with C large enough (e.g., $C = 160$), we have that the right hand side is upper bounded by n^{-10}. This establishes the theorem. ∎

B. Proof of Lemma 2 to 4

Lemma 2: Given $0 \le a_1 < a_2 < a_3 \le 1$, we have

$$\frac{\mathcal{V}(a_2) - \mathcal{V}(a_1)}{a_2 - a_1} \le \frac{\mathcal{V}(a_3) - \mathcal{V}(a_2)}{a_3 - a_2}.$$

Proof: By definition, $\mathcal{V}(a) = \int_{-\nu(a)}^{+\nu(a)} x^2 \, \overline{\mu}(dx)$, and notice that $\nu(\cdot)$ is increasing, we have that

$$\begin{aligned}\mathcal{V}(a_2) - \mathcal{V}(a_1) &= \int_{-\nu(a_2)}^{-\nu(a_1)} x^2 \, \overline{\mu}(dx) + \int_{+\nu(a_1)}^{+\nu(a_2)} x^2 \, \overline{\mu}(dx) \\ &\le \nu(a_2)^2 [\int_{-\nu(a_2)}^{-\nu(a_1)} \overline{\mu}(dx) + \int_{+\nu(a_1)}^{+\nu(a_2)} \overline{\mu}(dx)] \\ &= (a_2 - a_1)\nu(a_2)^2.\end{aligned}$$

On the other hand, by a similar argument, we have

$$\mathcal{V}(a_3) - \mathcal{V}(a_2) \ge (a_3 - a_2)\nu(a_2)^2.$$

The lemma thus follows. ∎

Lemma 2 immediately implies the Lemma 3. We next prove Lemma 4.

Lemma 4: For any $\epsilon > 0$ and $\kappa \in [\epsilon, 1]$, we have $\mathcal{V}(\kappa) - \mathcal{V}(\kappa - \epsilon) \le C\alpha\epsilon \log^2(1/\epsilon)$.

Proof: By monotonicity, it suffices to prove the result for $\kappa = 1$. Notice that for $K \ge 2\alpha$,

$$\begin{aligned}&\mathcal{V}(1) - \mathcal{V}(1-\epsilon) \\ &\le \epsilon K^2 + \mathbb{E}_{x\sim\overline{\mu}}\big(x^2 \cdot \mathbf{1}(x > K)\big) \\ &= \epsilon K^2 + \int_{K^2}^{\infty} \Pr_{x\sim\overline{\mu}}(x^2 > z)\, dz \\ &\le \epsilon K^2 + \int_{K^2}^{\infty} \exp(1 - \sqrt{z}/\alpha)\, dz \\ &= \epsilon K^2 + e_0 \int_{K^2/4\alpha^2}^{\infty} \exp(-2\sqrt{z})\, dz \\ &\overset{(a)}{\le} \epsilon K^2 + 2e_0 \exp(-\sqrt{z})|_{\infty}^{K^2/4\alpha^2} \\ &= \epsilon K^2 + \exp(1 + \ln 2 - K/2\alpha),\end{aligned}$$

where (a) holds because when $z \ge 1$, we have $\exp(-\sqrt{z}) \le 1/\sqrt{z}$, which implies $\exp(-2\sqrt{z}) \le \frac{d(2\exp(-\sqrt{z}))}{dz}$. Pick $K = 2\alpha\log(1/\epsilon)$, we have that

$$\mathcal{V}(1) - \mathcal{V}(1-c) \le C\alpha\epsilon\log^2(1/\epsilon).$$

∎

C. Proof of Theorem 7

This section is devoted to prove Theorem 7, i.e., to show Condition 1 holds with high probability. We establish each claims of Condition 1 separately.

Theorem 9: Let $\tau = \max(p/n, 1)$. Recall $\mathbf{v}_i$ are i.i.d. random variables following $\mathcal{N}(0, I_p)$, and $t = (1-\lambda)n$ for some $\lambda < 0.5$. Then, there exist a universal constant c such that with high probability,

$$\sup_{\mathbf{w}\in\mathcal{S}_p} \frac{1}{t}\sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{v}_i)^2 \le c\tau.$$

Proof: Theorem II.13 in [41] established that suppose Γ is an $p \times t$ matrix, whose entries are all i.i.d. $\mathcal{N}(0,1)$ Gaussian variables, then the largest singular value of Γ, denoted by $s_1(\Gamma)$, satisfies

$$\Pr\big(s_1(\Gamma) > \sqrt{p} + \sqrt{t} + \sqrt{p \vee t}\epsilon\big) \le \exp(-(p \vee t)\epsilon^2/2).$$

Our result now follows, since $\sup_{\mathbf{w}\in\mathcal{S}_p} \frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top\mathbf{v}_i)^2$ is the largest eigenvalue of $W = (1/t)\Gamma_1^\top\Gamma_1$, where Γ_1 is a $p \times t$ matrix whose entries are all i.i.d. $\mathcal{N}(0,1)$ Gaussian variables. Hence the largest eigenvalue of W is given by $\lambda_W = [s_1(\Gamma_1)]^2/t$. Thus we have

$$\begin{aligned}&\Pr\big(\lambda_W > \frac{\tau(2n + n\epsilon^2 + 2n + 4\sqrt{n^2}\epsilon)}{(1-\lambda)n}\big) \\ \le&\Pr\big(\lambda_W > \frac{p + t + (p\vee t)\epsilon^2 + 2\sqrt{pt} + 2(\sqrt{p} + \sqrt{t})\sqrt{(p\vee t)}\epsilon}{t}\big) \\ =&\Pr\big(s_1(\Gamma) > \sqrt{p} + \sqrt{t} + \sqrt{(p\vee t)}\epsilon\big) \\ \le&\exp(-(p\vee t)\epsilon^2/2) \\ \le&\exp(-(1-\lambda)n\tau\epsilon^2/2).\end{aligned}$$

Let $\epsilon = \sqrt{40(\log n)/n}$, and notice that $\lambda < 1/2$ and $\tau \ge 1$, then the right hand side is smaller than n^{-10}. Thus we conclude that with high probability

$$\sup_{\mathbf{w}\in\mathcal{S}_p} \frac{1}{t}\sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{v}_i)^2 \le c\tau.$$

∎

Notice that when $\mathbf{v}_i$ are sub-Gaussian, the theorem still holds, with c possibly depends on the sub-Gaussian moment [38].

Theorem 10: There exists an absolute constant $C > 0$, such that with high probability

$$\sup_{\mathbf{q}\in\mathcal{S}_d} \big|\frac{1}{t}\sum_{i=1}^{t}(\mathbf{q}^\top\mathbf{x}_i)^2 - 1\big| \le C\alpha\sqrt{\frac{d\log^3 n}{n}}.$$

Proof: The proof of Theorem 10 depends on the the following Lemma (adapted from Thm 5.41 of [38]).

Lemma 5: Let A be a $N \times M$ matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^M$. Let m be a number such that $\|A_i\|_2 \le \sqrt{m}$ for all i. Then for every $\beta \ge 0$, one has

$$\sqrt{N} - \beta\sqrt{m} \le \sigma_{\min}(A) \le \sigma_{\max}(A) \le \sqrt{N} + \beta\sqrt{m},$$

with probability at least $1 - 2M\exp(-c\beta^2)$, where $c > 0$ is an absolute constant.

Consider matrix X where the i^{th} row is $\mathbf{x}_i^\top$. To apply Lemma 5, we need to bound the range of each row. For any

$K > 0$:

$$\begin{aligned} & \Pr\left(\max_{i=1,\cdots,t}\|\mathbf{x}_i\|_2 \geq K\right) \\ \leq\ & t\Pr\left(\|\mathbf{x}_1\|_2 \geq K\right) \\ \leq\ & t\sum_{j=1}^{d}\Pr\left(|x_1(j)| \geq K/\sqrt{d}\right) \\ \overset{(a)}{\leq}\ & \exp\left(1 - \frac{K}{\alpha\sqrt{d}} + \log t + \log d\right). \end{aligned}$$

Here, $x_1(j)$ stands for the j-th component of $\mathbf{x}_1$. Inequality (a) holds because by sub-exponential property, we have

$$Pr(|x_1(j)| \geq K/\sqrt{d}) \leq \exp(1 - \frac{K}{\alpha\sqrt{d}})$$

Let the right-hand-side be n^{-10}, we have that with high probability, for a universal constant C,

$$\max_{i=1,\cdots,t}\|Z_i\| \leq C\alpha\log n\sqrt{d}.$$

Under this event, applying Lemma 5 on X, we have that

$$\begin{aligned} & \Pr\left(\sup_{\mathbf{q}\in\mathcal{S}_d}\left|\frac{1}{t}\sum_{i=1}^{t}(\mathbf{q}^\top\mathbf{x}_i)^2 - 1\right| \leq \frac{\beta C\alpha\log n\sqrt{d}}{\sqrt{t}}\right) \\ =\ & \Pr\Big(\sqrt{t} - \beta C\alpha\log n\sqrt{d} \leq \sigma_{\min}(X) \\ & \qquad \leq \sigma_{\max}(X) \leq \sqrt{t} + \beta C\alpha\log n\sqrt{d}\Big) \\ \geq\ & 1 - 2d\exp(-c\beta^2). \end{aligned}$$

Let the right hand side be $1 - n^{-10}$, we have $\beta = C'(\log n)^{1/2}$. Thus, with high probability,

$$\sup_{\mathbf{q}\in\mathcal{S}_d}\left|\frac{1}{t}\sum_{i=1}^{t}(\mathbf{q}^\top\mathbf{x}_i)^2 - 1\right| \leq \frac{C\alpha(\log n)^{3/2}\sqrt{d}}{\sqrt{n}}.$$

■

Theorem 11: With high probability, the following holds uniformly over $\overline{t} < t$ and $\mathbf{q} \in \mathcal{S}_d$,

$$\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top\mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t)\right| \leq \frac{Ct(1+\varepsilon_0)\sqrt{d\log n/n}}{t - \overline{t}}.$$

Proof: Consider two class of functions $\mathcal{F} = \{f_{e,\mathbf{q}} : \mathbb{R}^d \mapsto \mathbb{R} | e \in \mathbb{R}^+, \mathbf{q} \in \mathbb{R}^d\}$ and $\mathcal{G} = \{g_{e,\mathbf{q}} : \mathbb{R}^d \mapsto \mathbb{R} | e \in \mathbb{R}^+, \mathbf{q} \in \mathbb{R}^d\}$, as

$$\begin{aligned} f_{e,\mathbf{q}}(\mathbf{x}) &= [\mathbf{q}^\top\mathbf{x}]^2\mathbf{1}(|\mathbf{q}^\top\mathbf{x}| \leq e); \\ g_{e,\mathbf{q}}(\mathbf{x}) &= \mathbf{1}(|\mathbf{q}^\top\mathbf{x}| \leq e). \end{aligned}$$

Notice that the VC-dimension of $\mathcal{G}$ is at most $2d+3$, due to the fact that every $g_{e,\mathbf{v}}$ is the indicator function of the intersection of two half spaces in $\mathbb{R}^d$. Standard VC theory leads to that with high probability (i.e., at least $1 - n^{-10}$),

$$\sup_{e\geq 0,\mathbf{q}\in\mathbb{R}^d,\|\mathbf{q}\|=1}\left|\frac{1}{t}\sum_{i=1}^{t}[g_{e,\mathbf{q}}(\mathbf{x}_i)] - \mathbb{E}g_{e,\mathbf{q}}(\mathbf{x})\right| \leq C\sqrt{\frac{d\log n}{n}}. \tag{22}$$

Notice that

$$\begin{aligned} \mathbb{E}f_{e,\mathbf{q}}(\mathbf{x}) &= \mathbb{E}[\mathbf{q}^\top\mathbf{x}]^2\mathbf{1}([\mathbf{q}^\top\mathbf{x}]^2 \leq e) \\ &= \int_0^\infty \Pr\left([\mathbf{q}^\top\mathbf{x}]^2\mathbf{1}(|\mathbf{q}^\top\mathbf{x}| \leq e) > z\right)dz \\ &= \int_0^{e^2}\Pr\left([\mathbf{q}^\top\mathbf{x}]^2 > z\right)dz \\ &= \int_0^{e^2} 1 - \mathbb{E}g_{z,\mathbf{q}}(\mathbf{x})\,dz. \end{aligned}$$

Similarly, replacing μ with the empirical distribution of $\mathbf{x}_1, \ldots, \mathbf{x}_t$, we have

$$\sum_{i=1}^{t} f_{e,\mathbf{q}}(\mathbf{x}_i) = \int_0^{e^2} 1 - \frac{1}{t}\sum_{i=1}^{t} g_{z,\mathbf{q}}(\mathbf{x}_i)\,dz.$$

Due to Equation (22), we thus have with high probability, the following holds uniformly over $e > 0$, and $\mathbf{q} \in \mathbb{R}^d, \|\mathbf{q}\| = 1$,

$$\left|\frac{1}{t}\sum_{i=1}^{t} f_{e,\mathbf{q}}(\mathbf{x}_i) - \mathbb{E}f_{e,\mathbf{q}}(\mathbf{x})\right| \leq e^2 C\sqrt{\frac{d\log n}{n}}. \tag{23}$$

In the rest of the proof, we suppose Equation (22) and (23) hold, and the condition of Theorem 10 holds. Notice this requirement is satisfied with high probability.

We then have for any $\overline{t} < t$ and $\mathbf{q} \in \mathcal{S}_d$,

$$\begin{aligned} & \left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top\mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t)\right| \\ \leq & \left|\frac{1}{t}\sum_{i=1}^{t} f_{e(\overline{t}),\mathbf{q}}(\mathbf{x}_i) - \mathbb{E}f_{e(\overline{t}),\mathbf{q}}(\mathbf{x})\right| + \left|\mathbb{E}f_{e(\overline{t}),\mathbf{q}}(\mathbf{x}) - \mathcal{V}(\overline{t}/t)\right| \\ \leq & e(\overline{t})^2 C\sqrt{\frac{d\log n}{n}} + \left|\mathbb{E}f_{e(\overline{t}),\mathbf{q}}(\mathbf{x}) - \mathcal{V}(\overline{t}/t)\right|. \end{aligned} \tag{24}$$

In the first inequality, for simplicity we assume that $\mathbf{q}^\top\mathbf{x}_i \neq \mathbf{q}^\top\mathbf{x}_j$ for $i \neq j$. Such assumption can be relaxed, by considering instead $e(\overline{t}) - \epsilon$ and let $\epsilon \to 0$. Since $\mathcal{V}$ is continuous due to Lemma 4, our claim is still valid.

To bound the second term, notice that by Equation (22),

$$\begin{aligned} \left|\overline{t}/t - \overline{\mu}([-e_{\overline{t}}, e_{\overline{t}}])\right| &= \left|\frac{1}{t}\sum_{i=1}^{t} g_{e(\overline{t}),\mathbf{q}}(\mathbf{x}_i) - \mathbb{E}g_{e(\overline{t}),\mathbf{q}}(\mathbf{x})\right| \\ &\leq C\sqrt{\frac{d\log n}{n}}, \end{aligned}$$

which is equivalent to

$$\nu(\overline{t}/t - C\sqrt{\frac{d\log n}{n}}) \leq e(\overline{t}) \leq \nu(\overline{t}/t + C\sqrt{\frac{d\log n}{n}}).$$

This implies

$$\begin{aligned} & \left|\mathbb{E}f_{e(\overline{t}),\mathbf{q}}(\mathbf{x}) - \mathcal{V}(\overline{t}/t)\right| \\ \leq & \left\{\mathcal{V}\left(\overline{t}/t + C\sqrt{d\log n/n}\right) - \mathcal{V}(\overline{t}/t)\right\} \\ & \quad \vee \left\{\mathcal{V}(\overline{t}/t) - \mathcal{V}\left(\overline{t}/t - C\sqrt{d\log n/n}\right)\right\} \\ \leq & \frac{tC\sqrt{d\log n/n}}{t - \overline{t}}. \end{aligned} \tag{25}$$

where the last inequality follows from Lemma 3. To complete the proof, we bound $e(\overline{t})$. Notice that when Theorem 10 holds, we have

$$\frac{1}{t}\sum_{i=1}^{t} e(i)^2 \le 1+\varepsilon_0,$$

which combined with the fact that $e(1) \le e(2) \le \cdots \le e(t)$ leads to

$$e(\overline{t})^2 \le \frac{t(1+\varepsilon_0)}{t-\overline{t}}. \tag{26}$$

Substitute Equation (26) and (25) into Equation (24) leads to

$$\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t)\right| \le \frac{Ct(1+\varepsilon_0)\sqrt{d\log n/n}}{t-\overline{t}}.$$

■

One disadvantage of Theorem 11 is that the right-hand-side depends on $t/(t-\hat{t})$. However, this dependency can be removed, with a price of having a slower convergence rate, as the following corollary shows.

Corollary 5: Suppose $\varepsilon_0 \le C'$ for a universal constant C'. Then with high probability, the following holds uniformly over $\overline{t} < t$ and $\mathbf{q} \in \mathcal{S}_d$,

$$\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t)\right| \le C\alpha^{1/2}d^{1/4}(\log n)^{5/4}n^{-1/4}.$$

Proof: With high probability, Theorem 10 and 11 hold. Under the condition of Theorem 10 and 11, define a $t_0 \in [1:t]$ to satisfy

$$t_0 = \left[1-\Theta(\alpha^{-1/2}d^{1/4}n^{-1/4}\log^{-3/4}n)\right]t.$$

If $\overline{t} \le t_0$, then Theorem 11 leads to

$$\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t)\right| \le C\alpha^{1/2}d^{1/4}(\log n)^{5/4}n^{-1/4}.$$

If $\overline{t} > t_0$, then we have the following

$$\begin{aligned}
& \frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t) \\
\le\ & \frac{1}{t}\sum_{i=1}^{t}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t) \\
\le\ & \left|\frac{1}{t}\sum_{i=1}^{t}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - 1\right| + \left|1-\mathcal{V}(\overline{t}/t)\right| \\
\le\ & C_1\xi_0 + C_2\alpha\frac{t-t_0}{t}\log^2(t/(t-t_0)) \\
\le\ & C_1\alpha d^{\frac{1}{2}}(\log n)^{\frac{3}{2}}n^{-\frac{1}{2}} + C_2\alpha^{\frac{1}{2}}d^{\frac{1}{4}}n^{-\frac{1}{4}}(\log n)^{\frac{5}{4}} \\
\overset{(a)}{\le}\ & C\alpha^{\frac{1}{2}}d^{\frac{1}{4}}(\log n)^{\frac{5}{4}}n^{-\frac{1}{4}}.
\end{aligned}$$

where (a) holds because when $\varepsilon_0 = O(1)$, the first term is dominated by the second term. Furthermore,

$$\begin{aligned}
& \mathcal{V}(\overline{t}/t) - \frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} \\
\le\ & \mathcal{V}(\overline{t}/t) - \frac{1}{t}\sum_{i=1}^{t_0}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} \\
\le\ & \left|\frac{1}{t}\sum_{i=1}^{t_0}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(t_0/t)\right| + \left|\mathcal{V}(t_0/t) - \mathcal{V}(\overline{t}/t)\right| \\
\le\ & C_1\alpha^{\frac{1}{2}}d^{\frac{1}{4}}(\log n)^{\frac{5}{4}}n^{-\frac{1}{4}} + C_2\alpha\frac{t-t_0}{t}\log^2(t/(t-t_0)) \\
\le\ & C\alpha^{\frac{1}{2}}d^{\frac{1}{4}}(\log n)^{\frac{5}{4}}n^{-\frac{1}{4}}.
\end{aligned}$$

This implies for $\overline{t} > t_0$, we also have

$$\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}[\mathbf{q}^\top \mathbf{x}]^2_{(i)} - \mathcal{V}(\overline{t}/t)\right| \le C\alpha^{1/2}d^{1/4}(\log n)^{5/4}n^{-1/4}.$$

■

D. Proof of Corollary 1 and 2

Corollary 1: Given a sequence of $\{\mathcal{Y}(j)\}$, if the asymptotic scaling in Expression (1) holds, and denote $\lambda^* \triangleq \limsup \lambda(j)$, then the following holds in probability when $j \uparrow \infty$ (i.e., when $n, p \uparrow \infty$),

$$\begin{aligned}
& \liminf_j \mathrm{EV}_{\hat{d}}\{\overline{\mathbf{w}}_1(j), \ldots, \overline{\mathbf{w}}_{\hat{d}}(j)\} \\
& \ge \max_{\kappa}\left[1-\kappa-\frac{C\alpha\lambda^*\log^2(1/\lambda^*)}{\kappa\mathcal{V}(0.5)}\right] \\
& \ge 1-\frac{C\sqrt{\alpha\lambda^*}\log(1/\lambda^*)}{\mathcal{V}(0.5)}.
\end{aligned}$$

Proof: When $\kappa \ge 1$ the corollary holds trivially. Hence, fix $\kappa < 1$.

We bound the right-hand-side of Equation (3) to establish

the corollary. Notice that

$$\begin{aligned}
& \left[\frac{\mathcal{V}\left(1-\frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\right)}{(1+\kappa)}\right] \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] \\
\overset{(a)}{\geq} & \left[\frac{\mathcal{V}(1)-C\alpha\frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\log^2\left(\frac{(1-\lambda^*)\kappa}{\lambda^*(1+\kappa)}\right)}{(1+\kappa)}\right] \\
& \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t}\right)-C\alpha\frac{\lambda^*}{1-\lambda^*}\log^2\left(\frac{1-\lambda^*}{\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] \\
\overset{(b)}{\geq} & \left[\frac{1}{1+\kappa}-\frac{C\alpha\lambda^*}{(1-\lambda^*)\kappa}\log^2\left(\frac{(1-\lambda^*)\kappa}{\lambda^*(1+\kappa)}\right)\right] \\
& \times \left[1-\frac{C\alpha\frac{\lambda^*}{1-\lambda^*}\log^2\left(\frac{1-\lambda^*}{\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] \\
\overset{(c)}{\geq} & \left[1-\kappa-\frac{2C\alpha\lambda^*}{\kappa}\log^2\left(\frac{1}{\lambda^*}\right)\right] \\
& \times \left[1-\frac{2C\alpha\lambda^*\log^2\left(\frac{1}{\lambda^*}\right)}{\mathcal{V}(0.5)}\right] \\
\geq & 1-\kappa-\frac{C'\alpha\lambda^*}{\kappa}\log^2\left(\frac{1}{\lambda^*}\right)-\frac{C'\alpha\lambda^*\log^2\left(\frac{1}{\lambda^*}\right)}{\mathcal{V}(0.5)} \\
\overset{(d)}{\geq} & 1-\kappa-\frac{2C'\alpha\lambda^*}{\kappa\mathcal{V}(0.5)}\log^2\left(\frac{1}{\lambda^*}\right).
\end{aligned}$$

Here, (a) is due to Lemma 4; (b) is due to $\mathcal{V}(1)=1$; (c) holds because $\frac{1}{1+\kappa}\geq 1-\kappa$, $1-\lambda^*\geq 1/2$, and $\mathcal{V}(\hat{t}/t)\geq\mathcal{V}(0.5)$; (d) holds because κ and $\mathcal{V}(0.5)$ are both smaller than or equal to 1.

Taking $\kappa=\sqrt{\alpha\lambda^*}\log(1/\lambda^*)$ establishes that

$$1-\kappa-\frac{C\alpha\lambda^*\log^2(1/\lambda^*)}{\kappa\mathcal{V}(0.5)}\geq 1-\frac{C'\sqrt{\alpha\lambda^*}\log(1/\lambda^*)}{\mathcal{V}(0.5)}.$$

■

Corollary 2: Suppose $\overline{\mu}(\{0\})=0$. Given a sequence of $\{\mathcal{Y}(j)\}$, if the asymptotic scaling in Expression (1) holds, and denote $\lambda^*\triangleq\limsup\lambda(j)$, which satisfies $\lambda^*<0.5$, then the following holds in probability when $j\uparrow\infty$ (i.e., when $n,m\uparrow\infty$),

$$\liminf_j \mathrm{EV}_{\hat{d}}\{\overline{\mathbf{w}}_1(j),\ldots,\overline{\mathbf{w}}_{\hat{d}}(j)\}>0.$$

Proof: We prove the corollary by bounding the right hand side of Equation (3). To simplify notation, denote $\vartheta\triangleq 1-2\lambda^*>0$. We have the following

$$\begin{aligned}
& \frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\geq\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda^*}{1-\lambda^*}\right)\geq\mathcal{V}\left(\frac{0.5n-\lambda^*n}{(1-\lambda^*)n}\right) \\
& =\mathcal{V}\left(\frac{\vartheta}{2(1-\lambda^*)}\right)\geq\mathcal{V}(0.5\vartheta)>0.
\end{aligned}$$

Here the last inequality holds because $\vartheta>0$, and the fact that $\overline{\mu}(\{0\})=0$ implies $\mathcal{V}(c)>0$ for any positive c.

On the other hand, take $\kappa^*=1/\vartheta$, we thus have

$$\begin{aligned}
& \frac{\mathcal{V}\left(1-\frac{\lambda^*(1+\kappa^*)}{(1-\lambda^*)\kappa^*}\right)}{(1+\kappa^*)}=\frac{\mathcal{V}\left(\frac{(1-\lambda^*)\kappa^*-\lambda^*(1+\kappa^*)}{(1-\lambda^*)\kappa^*}\right)}{1+\frac{1}{\vartheta}} \\
& =\frac{\mathcal{V}\left(\frac{(1-2\lambda^*)\kappa^*-\lambda^*}{(1-\lambda^*)\kappa^*}\right)}{1+\frac{1}{\vartheta}}\overset{(a)}{=}\frac{\mathcal{V}(\vartheta)}{1+\frac{1}{\vartheta}}\overset{(b)}{>}0.
\end{aligned}$$

Here (a) follows from $\vartheta=1-2\lambda^*$ and $\kappa^*=1/\vartheta$; (b) holds since $\mathcal{V}(c)>0$ for any positive c. Thus, by Theorem 3, we have

$$\begin{aligned}
& \liminf_j \mathrm{EV}\{\overline{\mathbf{w}}_1(j),\ldots,\overline{\mathbf{w}}_{\hat{d}}(j)\} \\
\geq & \max_\kappa\left[\frac{\mathcal{V}\left(1-\frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\right)}{(1+\kappa)}\right]\times\left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] \\
\geq & \left[\frac{\mathcal{V}\left(1-\frac{\lambda^*(1+\kappa^*)}{(1-\lambda^*)\kappa^*}\right)}{(1+\kappa^*)}\right]\times\left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t}-\frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] \\
> & 0.
\end{aligned}$$

This establishes the corollary. ■

REFERENCES

[1] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. American Math. Society Lecture—Math. Challenges of the 21st Century, 2000.

[2] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[3] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[4] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[5] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, Berlin: Springer, 1986.

[6] P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

[7] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362, 1981.

[8] L. Xu and A. L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1):131–143, 1995.

[9] T. N. Yang and S. D. Wang. Robust algorithms for principal component analysis. *Pattern Recognition Letters*, 20(9):927–933, 1999.

[10] C. Croux and G. Hasebroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

[11] F. De la Torre and M. J. Black. Robust principal component analysis for computer vision. In *Proceedings of the Eighth International Conference on Computer Vision (ICCV'01)*, pages 362–369, 2001.

[12] F. De la Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1/2/3):117–142, 2003.

[13] C. Croux, P. Filzmoser, and M. Oliveira. Algorithms for Projection−Pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

[14] S. C. Brubaker. Robust PCA and clustering on noisy mixtures. In *Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 1078–1087, 2009.

[15] S. Dasgupta. Subspace detection: A robust statistics formulation. In *Proceedings of the Sixteenth Annual Conference on Learning Theory*, pages 734–734, 2003.

[16] A. Klivans, P. Long, and R. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

[17] D. L. Donoho. Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University, 1982.

[18] M. Hubert, P. J. Rousseeuw, and K. Branden. ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[19] P. J. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze, and W. Wertz, editors, *Mathematical Statistics and Applications*, pages 283–297. Reidel, Dordrecht, 1985.
[20] R. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67, 1976.
[21] V. Barnett. The ordering of multivariate data. *Journal of Royal Statistics Society Series, A*, 138:318–344, 1976.
[22] A. Bebbington. A method of bivariate trimming for robust estimation of the correlation coefficient. *Applied Statistics*, 27:221–228, 1978.
[23] D. Titterington. Estimation of correlation coefficients by ellipsoidal trimming. *Applied Statistics*, 27:227–234, 1978.
[24] J. Helbling. *Ellipsoïdes minimaux de couverture en statistique multivariée*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland, 1983.
[25] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, New York, 1978.
[26] H. David. *Order Statistics*. Wiley, New York, 1981.
[27] A. Dempster and M. Gasko-Green. New tools for residual analysis. *The Annals of Statistics*, 9(5):945–959, 1981.
[28] R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28:81–124, 1972.
[29] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62:531–545, 1975.
[30] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226, 2005.
[31] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.
[32] B. Recht. A simpler approach to matrix completion. ArXiv: 0910.0651, 2009.
[33] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. ArXiv:0906.2220, 2009.
[34] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma. Stable principal component pursuit. ArXiv:1001.2363, 2010.
[35] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, 2010.
[36] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. Forthcoming in *IEEE Transaction on Information Theory*, 2011.
[37] M. McCoy and J. Tropp. Two proposals for robust pca using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.
[38] R. Vershynin. Introductioin to the non-asymptotic analysis of random matrices. ArXiv: 1011.3027v6, 2011.
[39] B. Schölkopf, A. J. Smola, and K. R. Müller. Kernel principal component analysis. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Advances in kernel Methods – Support Vector Learning*, pages 327–352. MIT Press, Cambridge, MA, 1999.
[40] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58:1–37, 2011.
[41] K. Davidson and S. Szarek. Local operator theory, random matrices and banach spaces. In W. Johnson and J. Lindenstrauss, editors, *Handbook on the Geometry of Banach Spaces*, pages 317–366. Elsvier, 2001.

Huan Xu received the B.Eng. degree in automation from Shanghai Jiaotong University, Shanghai, China in 1997, the M.Eng. degree in electrical engineering from the National University of Singapore in 2003, and the Ph.D. degree in electrical engineering from McGill University, Canada in 2009. From 2009 to 2010, he was a postdoctoral associate at The University of Texas at Austin.

Since 2011, he has been an assistant professor at the Department of Mechanical Engineering at the National University of Singapore. His research interests include statistics, machine learning, robust optimization, and planning and control.

Constantine Caramanis (M'06) received his Ph.D. in electrical engineering and computer science from the Massachusetts Institute of Technology in 2006. Since then, he has been on the faculty in the Department of Electrical and Computer Engineering at The University of Texas at Austin. He received the NSF CAREER Award in 2011. His current research interests include robust and adaptable optimization, machine learning and high-dimensional statistics, with applications to large scale networks

Shie Mannor (S'00-M'03-SM-09') received the B.Sc. degree in electrical engineering, the B.A. degree in mathematics, and the Ph.D. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1996, 1996, and 2002, respectively. From 2002 to 2004, he was a Fulbright scholar and a postdoctoral associate at M.I.T. He was with the Department of Electrical and Computer Engineering at McGill University from 2004 to 2010 where he was the Canada Research chair in Machine Learning. He has been an associate professor at the Faculty of Electrical Engineering at the Technion since 2008. His research interests include machine learning and pattern recognition, planning and control, multi-agent systems, and communications.