Principal Component Analysis with Contaminated Data: The High Dimensional Case

Huan Xu
The University of Texas at Austin
huan.xu@mail.utexas.edu

Constantine Caramanis
The University of Texas at Austin
caramanis@mail.utexas.edu

Shie Mannor
Technion, Israel
shie@ee.technion.ac.il

Abstract

We consider the dimensionality-reduction problem (finding a subspace approximation of observed data) for contaminated data in the high dimensional regime, 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 tractable, robust to contaminated points, and easily kernelizable. The resulting subspace has a bounded deviation from the desired one, achieves maximal robustness – a breakdown point of 50% while all existing algorithms have a breakdown point of *zero*, and unlike ordinary PCA algorithms, achieves optimality in the limit case where the proportion of corrupted points goes to zero.

1 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 (Donoho, 2000; Johnstone, 2001). For example, observations on individual instances can be curves, spectra, images or even movies, where a single observation has dimensionality ranging from thousands to billions. 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" (Schölkopf & Smola, 2002) transforms virtually every data set to a high dimensional one. Efforts of extending traditional statistical tools (designed for the low dimensional case) into this high-dimensional regime are generally unsuccessful. This fact has stimulated research on formulating new data-analysis techniques able to cope with such a "dimensionality explosion."

Principal Component Analysis (PCA) is one of the most widely used statistical techniques for dimensionality reduction. Work on PCA dates back as early as Pearson (1901), and has become one of the most important techniques for data compression and feature extraction. It is widely used in statistical data analysis, communication, pattern recognition, and image processing (Jolliffe, 1986). 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. 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, results in non-convex (and intractable) 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 a 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 possibly 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 standard definitions of "outliers" are of limited use in this setting. Another important property of this setup is that the signal-to-noise ratio (SNR) can go to 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

In this paper, we give a surprisingly optimistic message: contrary to what one might expect given the brittle nature of classical PCA, and in stark contrast to previous algorithms, it is possible to recover such low SNR signals, in the high-dimensional regime, even in the face of a *constant fraction of arbitrarily corrupted data.* Moreover, we show that this can be accomplished with an efficient (polynomial time) algorithm, which we call High-Dimensional Robust PCA (HR-PCA). The algorithm we propose here is tractable, provably robust to corrupted points, and asymptotically optimal, recovering the *exact* low-dimensional subspace when the number of corrupted points scales more slowly than the number of "authentic" samples – to the best of our knowledge, the only algorithm of this kind. Moreover, it is easily kernelizable.

Organization and Notation

In Section 2 we discuss past work and the reasons that classical robust PCA algorithms fail to extend to the high dimensional regime. In Section 3 we present the setup of the problem, and the HR-PCA algorithm. We also provide finite sample and asymptotic performance guarantees. The performance guarantees are proved in Section 4. Kernelization, simulation and some technical details in the derivation of the performance guarantees are postponed to the full version (Xu et al., 2010).

Capital letters and boldface letters are used to denote matrices and vectors, respectively. A $k \times k$ unit matrix is denoted by I_k. For $c \in \mathbb{R}$, $[c]^+ \triangleq \max(0, c)$. We let $\mathcal{B}_d \triangleq \{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\| \leq 1\}$, and $\mathcal{S}_d$ be its boundary. We use a subscript $(\cdot)$ to represent order statistics of a random variable. For example, let $v_1, \cdots, v_n \in \mathbb{R}$. Then $v_{(1)}, \cdots, v_{(n)}$ is a permutation of $v_1, \cdots, v_n$, in a non-decreasing order.

2 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" (Huber, 1981), i.e., 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. We show that the algorithm we provide has breakdown point of 50%, which is the best break-down point possible for any algorithm. In addition to this, we focus on providing explicit lower 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., Devlin et al., 1981; Xu & Yuille, 1995; Yang & Wang, 1999; Croux & Hasebroeck, 2000; De la Torre & Black, 2001; De la Torre & Black, 2003; Croux et al., 2007; Brubaker, 2009).

We discuss three main pitfalls these and other existing algorithms face in high dimensions.

Diminishing Breakdown Point: If an algorithm's breakdown point has an inverse dependence on the dimensionality, then it is unsuitable in our regime. Many algorithms fall into this category. In Donoho (1982), several covariance estimators including M-estimator (Maronna, 1976), Convex Peeling (Barnett, 1976; Bebbington, 1978), Ellipsoidal Peeling (Titterington, 1978; Helbling, 1983), Classical Outlier Rejection (Barnett & Lewis, 1978; David, 1981), Iterative Deletion (Dempster & Gasko-Green, 1981) and Iterative Trimming (Gnanadesikan & Kettenring, 1972; Devlin et al., 1975) are all shown to have breakdown points upper-bounded by the inverse of the dimensionality, hence not useful in the regime of interest.

Noise Explosion: In the basic PCA model, zero mean standard Gaussian noise is added to each sample observed. Concentration results for Gaussian vectors promise that the noise magnitude will sharply concentrate around the ball of radius equal to square root of the dimension. This can be significantly larger than what we call the "signal strength," namely, the magnitude of the signal before noise was added. Thus, the ratio of the signal strength to the noise level quickly goes to zero as we scale the dimensionality up. Because of this, several perhaps counter-intuitive properties hold in this regime. First, any given authentic point is with overwhelming probability very close to orthogonal to the signal space (i.e., to the true principal components). Second, it is possible for a constant fraction of corrupted points all with a small Mahalanobis distance to significantly change the output of PCA. Indeed, by aligning the entire fraction of corrupted points magnitude some constant multiple of what we have called the signal strength, it is easy to see that the output of PCA can be strongly manipulated. On the other hand, since the noise magnitude is much larger, and in a direction perpendicular to the principal components, the Mahalanobis distance of each corrupted point will be very small. Third, the same example as above shows that it is possible for a constant fraction of corrupted points all with small Stahel-Donoho (S-D) outlyingness to significantly change the output of PCA, where recall that S-D outlyingness of a sample $\mathbf{y}_i$ is defined as:

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

Here med_k stands for taking median over all k.

The Mahalanobis distance and the 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 (Donoho, 1982) weights the contribution of observations based on their S-D outlyingness. More recently, the ROBPCA algorithm proposed in Hubert et al. (2005) selects a subset of observations with least S-D outlyingness to compute the d-dimensional signal space. Thus, in the high-dimensional case, these algorithms may run into problems since neither Mahalanobis distance nor S-D outlyingness are valid indicators of outliers. Indeed, as shown in the simulations, the empirical performance of such algorithms can be worse than standard PCA, because they remove the authentic samples.

Algorithmic Tractability: There are algorithms that do not rely on Mahalanobis distance or S-D outlyingness, and have a non-diminishing breakdown point, namely Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) (Rousseeuw, 1984) and Projection-Pursuit (Li & Chen, 1985). MVE finds the minimum volume ellipsoid that covers a certain fraction of observations. MCD finds a fraction of observations whose covariance matrix has a minimal determinant. Projection Pursuit maximizes a certain robust univariate variance estimator over all directions.

MCD and MVE are combinatorial, and hence (as far as we know) computationally intractable as the size of the problem scales. More difficult yet, MCD and MVE are ill-posed in the high-dimensional setting where the number of points (roughly) equals the dimension, since there exist infinitely many zero-volume (determinant) ellipsoids satisfying the covering requirement. Nevertheless, we note that such algorithms work well in the low-dimensional case, and hence can potentially be used as a post-processing procedure of our algorithm by projecting all observations to the output subspace to fine tune the eigenvalues and eigenvectors we produce.

Maximizing a robust univariate variance estimator as in Projection Pursuit, is also non-convex, and thus to the best of our knowledge, computationally intractable. In Croux and Ruiz-Gazen (2005), the authors propose a fast Projection-Pursuit algorithm, avoiding the non-convex optimization problem of finding the optimal direction, by only examining the directions of each sample. While this is suitable in the classical regime, 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 directions.

Low Rank Techniques: Finally, we discuss the recent (as of yet unpublished) paper (Candès et al., 2009). In this work, the authors adapt techniques from low-rank matrix approximation, and in particular, results similar to the matrix decomposition results of Chandrasekaran et al. (2009), in order to recover a low-rank matrix L_0 from highly corrupted measurements $M = L_0 + S_0$, where the noise term, S_0, is assumed to have a sparse structure. This models the scenario where we have perfect measurement of most of the entries of L_0, and a small (but constant) fraction of the entries are arbitrarily corrupted. This work is much closer in spirit, in motivation, and in terms of techniques, to the low-rank matrix completion and matrix recovery problems in Candès and Recht (2009); Recht (2009); Recht et al. (2010) than the setting we consider and the work presented herein. In particular, in our setting, even one corrupted point can change *every* element of the measurement

M.

3 HR-PCA: The Algorithm

The algorithm of HR-PCA is presented in this section. We start with the mathematical setup of the problem in Section 3.1. The HR-PCA algorithm as well as its performance guarantee are then given in Section 3.2.

3.1 Problem Setup

We now define in detail the problem described above.

- The "authentic samples" $\mathbf{z}_1, \ldots, \mathbf{z}_t \in \mathbb{R}^m$ are generated by $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$, where $\mathbf{x}_i \in \mathbb{R}^d$ (the "signal") are i.i.d. samples of a random variable $\mathbf{x}$, and $\mathbf{n}_i$ (the "noise") are independent realizations of $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, I_m)$. The matrix $A \in \mathbb{R}^{m\times d}$ and the distribution of $\mathbf{x}$ (denoted by μ) are unknown. We do assume, however, that the distribution μ is absolutely continuous with respect to the Borel measure, it is spherically symmetric (and in particular, $\mathbf{x}$ has mean zero and variance I_d) and it has light tails, specifically, there exist constants K and $C > 0$ such that $\Pr(\|\mathbf{x}\| \geq x) \leq K\exp(-Cx)$ for all $x \geq 0$. Since the distribution μ and the dimension d are both fixed, as m, n scale, the assumption that μ is spherically symmetric can be easily relaxed, and the expense of potentially significant *notational complexity*.

- The outliers (the corrupted data) are denoted $\mathbf{o}_1, \ldots, \mathbf{o}_{n-t} \in \mathbb{R}^m$ and as emphasized above, they are arbitrary (perhaps even maliciously chosen). We denote the fraction of corrupted points by $\lambda \triangleq (n-t)/n$.

- 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".

Given these contaminated observations, we want to recover the principal components of A, i.e., the top eigenvectors, $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_d$ of $AA^\top$. That is, we seek a collection of orthogonal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$, that maximize the performance metric called the *Expressed Variance* (E.V.):

$$\text{E.V.}(\mathbf{w}_1, \ldots, \mathbf{w}_d) \triangleq \frac{\sum_{j=1}^d \mathbf{w}_j^\top AA^\top \mathbf{w}_j}{\sum_{j=1}^d \overline{\mathbf{w}}_j^\top AA^\top \overline{\mathbf{w}}_j} = \frac{\sum_{j=1}^d \mathbf{w}_j^\top AA^\top \mathbf{w}_j}{\text{trace}(AA^\top)}.$$

The E.V. is always less than one, with equality achieved exactly when the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$ have the span of the true principal components $\{\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_d\}$. When $d = 1$, the Expressed Variance relates to another natural performance metric — the angle between $\mathbf{w}_1$ and $\overline{\mathbf{w}}_1$ — since by definition $E.V.(\mathbf{w}_1) = \cos^2(\angle(\mathbf{w}_1, \overline{\mathbf{w}}_1))$.[1] The Expressed Variance represents the portion of signal $A\mathbf{x}$ being expressed by $\mathbf{w}_1, \ldots, \mathbf{w}_d$. Equivalently, $1 - E.V.$ is the reconstruction error of the signal.

It is natural to expect that the ability to recover vectors with a high expressed variance depends on λ, the fraction of corrupted points — in addition, it depends on the distribution, μ generating the (low-dimensional) points $\mathbf{x}$, through its tails. If μ 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 define the following "tail weight" function $\mathcal{V} : [0, 1] \to [0, 1]$:

$$\mathcal{V}(\alpha) \triangleq \int_{-c_\alpha}^{c_\alpha} x^2 \overline{\mu}(dx);$$

where $\overline{\mu}$ is the one-dimensional margin of μ (recall that μ is spherically symmetric), and c_α is such that $\overline{\mu}([-c_\alpha, c_\alpha] = \alpha)$. Since μ has a density function, c_α is well defined. Thus, $\mathcal{V}(\cdot)$ represents how the tail of $\overline{\mu}$ contributes to its variance. Notice that $\mathcal{V}(0) = 0$, $\mathcal{V}(1) = 1$, and $\mathcal{V}(\cdot)$ is continuous in $[0, 1]$ since μ has a density function. For notational convenience, we simply let $\mathcal{V}(x) = 0$ for $x < 0$, and $\mathcal{V}(x) = \infty$ for $x > 1$.

The bounds on the quality of recovery, given in Theorems 1 and 2 below, are functions of η and the function $\mathcal{V}(\cdot)$.

[1] This geometric interpretation does not extend to the case where $d > 1$, since the angle between two subspaces is not well defined.

High Dimensional Setting and Asymptotic Scaling

In this paper, we focus on the case where $n \sim m \gg d$ and $\text{trace}(A^\top A) \gg 1$. That is, the number of observations and the dimensionality are of the same magnitude, and much larger than the dimensionality of $\mathbf{x}$; the trace of $A^\top A$ is significantly larger than 1, but may be much smaller than n and m. In our asymptotic scaling, n and m scale together to infinity, while d remains fixed. The value of $\text{trace}(A^\top A)$ also scales to infinity, but there is no lower bound on the rate at which this happens (and in particular, the scaling of $\text{trace}(A^\top A)$ can be much slower than the scaling of m and n).

While we give finite-sample results, we are particularly interested in the asymptotic performance of HR-PCA when *the dimension and the number of observations grow together* to infinity. 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 $\mathcal{Y}(j)$, $n(j)$, $m(j)$, $A(j)$, $d(j)$, etc., denote the corresponding values of the quantities defined above. Then the following must hold for some positive constants c_1, c_2:

$$\begin{aligned} &\lim_{j\to\infty} \frac{n(j)}{m(j)} = c_1; \quad d(j) \le c_2; \quad m(j) \uparrow +\infty; \\ &\text{trace}(A(j)^\top A(j)) \uparrow +\infty. \end{aligned} \tag{1}$$

While $\text{trace}(A(j)^\top A(j)) \uparrow +\infty$, if it scales more slowly than $\sqrt{m(j)}$, the SNR will asymptotically decrease to zero.

3.2 Key Idea and Main Algorithm

For $\mathbf{w} \in \mathcal{S}_m$, we define the Robust Variance Estimator (RVE) as $\overline{V}_{\hat{t}}(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{y}|^2_{(i)}$. This stands for the following statistics: project $\mathbf{y}_i$ onto the direction $\mathbf{w}$, replace the furthest (from original) $n - \hat{t}$ samples by 0, and then compute the variance. Notice that the RVE is always performed on the original observed set $\mathcal{Y}$.

The main algorithm of HR-PCA is as given below.

Algorithm 1 HR-PCA

Input: Contaminated sample-set $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\} \subset \mathbb{R}^m$, d, $\overline{T}$, $\hat{t}$.

Output: $\mathbf{w}_1^*, \dots, \mathbf{w}_d^*$.

Algorithm:

1. Let $\hat{\mathbf{y}}_i := \mathbf{y}_i$ for $i = 1, \dots n$; $s := 0$; Opt $:= 0$.
2. While $s \le \overline{T}$, 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 $\mathbf{w}_1, \dots, \mathbf{w}_d$ be the d principal components of $\hat{\Sigma}$.
   (c) If $\sum_{j=1}^d \overline{V}_{\hat{t}}(\mathbf{w}_j) >$ Opt, then let Opt $:= \sum_{j=1}^d \overline{V}_{\hat{t}}(\mathbf{w}_j)$ and let $\mathbf{w}_j^* := \mathbf{w}_j$ for $j = 1, \cdots, d$.
   (d) Randomly remove a point from $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s}$ according to
   $$\Pr(\hat{\mathbf{y}}_i \text{ is removed}) \propto \sum_{j=1}^d (\mathbf{w}_j^\top \hat{\mathbf{y}}_i)^2;$$
   (e) Denote the remaining points by $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s-1}$;
   (f) $s := s + 1$.
3. Output $\mathbf{w}_1^*, \dots, \mathbf{w}_d^*$. End.

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 a candidate direction, $\mathbf{w}$, our robust variance estimator measures the variance of the $(n - \hat{t})$-smallest points projected in that direction. 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 direction $\mathbf{w}$, will remove 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, 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, i.e., lower bounds on the expressed variance of the principal components the algorithm recovers.

There are two parameters to tune for HR-PCA, namely $\hat{t}$ and $\overline{T}$. Basically, $\hat{t}$ affects the performance of HR-PCA through Inequality 2, and as a rule of thumb we can set $\hat{t} = t$ if no *a priori* information of μ exists. (Note that our algorithm does assume knowledge of at least a lower bound on the number of authentic points, or, equivalently, an upper bound on λ, the fraction of corrupted points.) $\overline{T}$ does not affect the performance as long as it is large enough, hence we can simply set $T = n-1$, although when λ is small, a smaller T leads to the same solution with less computational cost.

The correctness of HR-PCA is shown in the following theorems for both the finite-sample bound, and the asymptotic performance.

Theorem 1 (Finite Sample Performance) *Let the algorithm above output $\{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$. Fix a $\kappa > 0$, and let $\tau = \max(m/n, 1)$. There exists a universal constant c_0 and a constant C which can possible depend on $\hat{t}/t$, λ, d, μ and κ, such that for any $\gamma < 1$, if $n/\log^4 n \geq \log^6(1/\gamma)$, then with probability $1-\gamma$ the following holds*

$$\begin{aligned}\text{E.V.}\{\mathbf{w}_1, \ldots, \mathbf{w}_\text{d}\} \quad \geq \quad & \left[\frac{\mathcal{V}\left(1 - \frac{\lambda(1+\kappa)}{(1-\lambda)\kappa}\right)}{(1+\kappa)}\right] \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] \\ & - \left[\frac{8\sqrt{c_0 \tau d}}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] (\text{trace}(AA^\top))^{-1/2} - \left[\frac{2c_0\tau}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right] (\text{trace}(AA^\top))^{-1} - C\frac{\log^2 n \log^3(1/\gamma)}{\sqrt{n}}.\end{aligned}$$

The last three terms go to zero as the dimension and number of points scale to infinity, i.e., as n and $m \to \infty$. Therefore, we immediately obtain:

Theorem 2 (Asymptotic Performance) *Given a sequence of $\{\mathcal{Y}(j)\}$, if the asymptotic scaling in Expression (1) holds, and $\limsup \lambda(j) \leq \lambda^*$, then the following holds in probability when $j \uparrow \infty$ (i.e., when n and $m \uparrow \infty$),*

$$\liminf_j \text{E.V.}\{\mathbf{w}_1(\text{j}), \ldots, \mathbf{w}_\text{d}(\text{j})\} \geq \max_\kappa \left[\frac{\mathcal{V}\left(1 - \frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\right)}{(1+\kappa)}\right] \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right]. \tag{2}$$

Remark

1. 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 (small) effect on the principal component directions reported in the output.

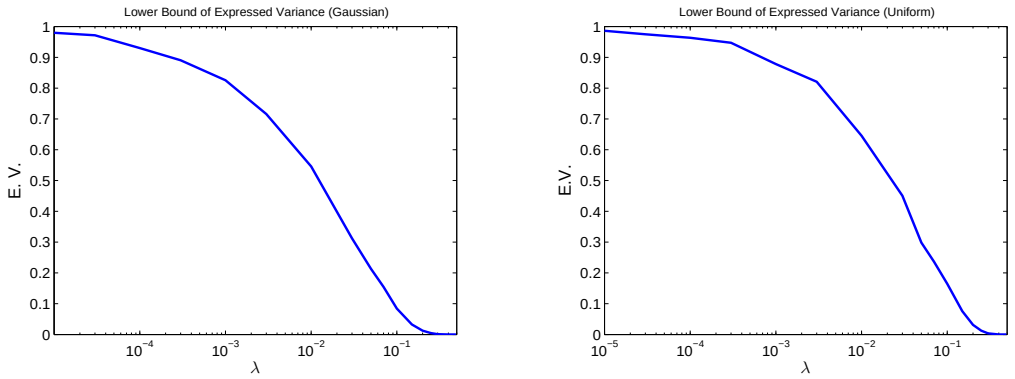
2. The terms in the second line of Theorem 1 go to zero as n and m increases, and therefore Theorem 1 immediately implies Theorem 2.

3. If $\lambda(j) \downarrow 0$, i.e., the *number* of corrupted points scales sublinearly (in particular, this holds when there are a fixed number of corrupted points), then the right-hand-side of Inequality (2) equals

1,[2] 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.

4. The breakdown point of HR-PCA converges to 50%. Note that since μ has a density function, $\mathcal{V}(\alpha) > 0$ for any $\alpha \in (0, 1]$. Therefore, for any $\lambda < 1/2$, if we set $\hat{t}$ to any value in $(\lambda n, t]$, then there exists κ large enough such that the right-hand-side is strictly positive (recall that $t = (1 - \lambda)n$). The breakdown point hence converges to 50%. Thus, HR-PCA achieves the maximal possible break-down point (note that a breakdown point greater than 50% is never possible, since then there are more outliers than samples.

The graphs in Figure 1 illustrate the lower-bounds of asymptotic performance if the 1-dimension marginal of μ is the Gaussian distribution or the Uniform distribution.



(a) Gaussian distribution (b) Uniform distribution

Figure 1: Lower Bounds of Asymptotic Performance.

We briefly discuss kernelizing HR-PCA : given a feature mapping $\Upsilon(\cdot) : \mathbb{R}^m \to \mathcal{H}$ equipped with a kernel function $k(\cdot, \cdot)$, we perform the dimensionality reduction in the feature space $\mathcal{H}$ without knowing the explicit form of $\Upsilon(\cdot)$. Notice that HR-PCA involves finding a set of PCs $\mathbf{w}_1, \ldots, \mathbf{w}_d \in \mathcal{H}$, and evaluating $\langle \mathbf{w}_q, \Upsilon(\cdot) \rangle$ (The 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 the Kernel PCA algorithm introduced by Schölkopf et al. (1999), 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 $\langle \mathbf{w}_q, \Upsilon(\mathbf{v}) \rangle = \sum_{j=1}^{n-s} \alpha_j(q) k(\hat{\mathbf{y}}_j, \mathbf{v})$, for all $\mathbf{v} \in \mathbb{R}^m$, implying that HR-PCA can be kernelized. We leave the details to the full version (Xu et al., 2010). Due to space constraints, numerical simulations are also deferred to the full version (Xu et al., 2010).

4 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 full version (Xu et al., 2010). The proof consists of three steps which we now outline. In what follows, we let d, m/n, λ, $\hat{t}/t$, and μ be fixed. We can fix a $\lambda \in (0, 0.5)$ without loss of generality, due to the fact that if a result is shown to hold for λ, then it holds for $\lambda' < \lambda$. The letter c is used to represent a constant, and ϵ is a constant that decreases to zero as n and m increase to infinity. The values of c and ϵ can change from line to line, and can possibly depend on d, m/n, λ, $\hat{t}/t$, and μ.

1. The blessing of dimensionality, and laws of large numbers: The first step involves two ideas; the first is the well-known fact (e.g., Davidson & Szarek, 2001) as n and m scale, the expectation of the covariance of the noise is bounded *independently* of m. The second involves appealing to laws of large numbers to show that sample estimates of the covariance of the noise, $\mathbf{n}$, of the signal, $\mathbf{x}$, and then of the authentic points, $\mathbf{z} = A\mathbf{x} + \mathbf{n}$, are uniformly close to their expectation, with high probability. Specifically, we prove that:

[2] We can take $\kappa(j) = \sqrt{\lambda(j)}$ and note that since μ has a density, $\mathcal{V}(\cdot)$ is continuous.

(a) With high probability, the largest eigenvalue of the variance of noise matrix is bounded. That is,

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

(b) With high probability, both the largest and the smallest eigenvalue of the signals in the original space converge to 1. That is

$$\sup_{\mathbf{w}\in\mathcal{S}_d} |\frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top \mathbf{x}_i)^2 - 1| \le \epsilon.$$

(c) Under 1b, with high probability, RVE is a valid variance estimator for the $d-$dimensional signals. That is,

$$\sup_{\mathbf{w}\in\mathcal{S}_d} |\frac{1}{t}\sum_{i=1}^{\hat{t}}|\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}\left(\frac{\hat{t}}{t}\right)| \le \epsilon.$$

(d) Under 1a and 1c, RVE is a valid estimator of the variance of the authentic samples. That is, the following holds uniformly over all $\mathbf{w} \in \mathcal{S}_m$,

$$(1-\epsilon)\|\mathbf{w}^\top A\|^2\mathcal{V}\left(\frac{t'}{t}\right) - c\|\mathbf{w}^\top A\| \le \frac{1}{t}\sum_{i=1}^{t'}|\mathbf{w}^\top \mathbf{z}|^2_{(i)} \le (1+\epsilon)\|\mathbf{w}^\top A\|^2\mathcal{V}\left(\frac{t'}{t}\right) + c\|\mathbf{w}^\top A\|.$$

2. The next 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.

3. The previous step shows the existence of a "good" solution. The final 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 these two steps, we derive the finite-sample and asymptotic performance bounds for HR-PCA.

4.1 Step 1

We state the main results for Step 1a and 1b. The proofs are deferred to the full version (Xu et al., 2010). In a nutshell, they hold by applying Theorem II.13 of Davidson and Szarek (2001), and Theorem 2.1 of Mendelson and Pajor (2006), respectively.

Theorem 3 *There exist universal constants c and c' such that for any $\gamma > 0$, with probability at least $1-\gamma$, the following holds:*

$$\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top \mathbf{n}_i)^2 \le c + \frac{c'\log\frac{1}{\gamma}}{n}.$$

Theorem 4 *There exists a constant c that only depends on μ and d, such that for any $\gamma > 0$, with probability at least $1-\gamma$,*

$$\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top \mathbf{x}_i)^2 - 1\right| \le \frac{c\log^2 n\log^3\frac{1}{\gamma}}{\sqrt{n}}.$$

The next theorem is the main result for Step 1c. Briefly speaking, since d is fixed, the result holds due to a standard uniform convergence argument. See Xu et al. (2010) for details.

Theorem 5 *Fix $\eta < 1$. There exists a constant c that depends on d, μ and η, such that for all $\gamma < 1$, t, the following holds with probability at least $1-\gamma$:*

$$\sup_{\mathbf{w}\in\mathcal{S}_d,\overline{t}\le\eta t}\left|\frac{1}{t}\sum_{i=1}^{\overline{t}}|\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}\left(\frac{\overline{t}}{t}\right)\right| \le c\sqrt{\frac{\log n + \log 1/\gamma}{n}} + c\frac{\log^{5/2} n\log^{7/2}(1/\gamma)}{n}.$$

Recall that $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$. Algebraic manipulation yields Theorem 6, which is the main result of Step 1d, and Corollary 7.

Theorem 6 *Let $t' \le t$. If there exists $\epsilon_1, \epsilon_2, \overline{c}$ such that (I) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t'} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}(\frac{t'}{t})\right| \le \epsilon_1$; (II) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{x}_i|^2 - 1\right| \le \epsilon_2$; (III) $\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{n}_i|^2 \le \overline{c}$, then for all $\mathbf{w} \in \mathcal{S}_m$ the following holds:*

$$(1-\epsilon_1)\|\mathbf{w}^\top A\|^2 \mathcal{V}\left(\frac{t'}{t}\right) - 2\|\mathbf{w}^\top A\|\sqrt{(1+\epsilon_2)\overline{c}}$$
$$\le \frac{1}{t}\sum_{i=1}^{t'} |\mathbf{w}^\top \mathbf{z}|^2_{(i)} \le (1+\epsilon_1)\|\mathbf{w}^\top A\|^2 \mathcal{V}\left(\frac{t'}{t}\right) + 2\|\mathbf{w}^\top A\|\sqrt{(1+\epsilon_2)\overline{c}} + \overline{c}.$$

Corollary 7 *Let $t' \le t$. If there exists $\epsilon_1, \epsilon_2, \overline{c}$ such that (I) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t'} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}(\frac{t'}{t})\right| \le \epsilon_1$; (II) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{x}_i|^2 - 1\right| \le \epsilon_2$; (III) $\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{n}_i|^2 \le \overline{c}$, then for any $\mathbf{w}_1, \cdots, \mathbf{w}_d \in \mathcal{S}_m$, and let $H(\mathbf{w}_1, \cdots, \mathbf{w}_d) \triangleq \sum_{j=1}^{d} \|\mathbf{w}_j^\top A\|^2$, the following holds*

$$(1-\epsilon_1)\mathcal{V}\left(\frac{t'}{t}\right) H(\mathbf{w}_1, \cdots, \mathbf{w}_d) - 2\sqrt{(1+\epsilon_2)\overline{c}dH(\mathbf{w}_1, \cdots, \mathbf{w}_d)}$$
$$\le \sum_{j=1}^{d} \frac{1}{t}\sum_{i=1}^{t'} |\mathbf{w}_j^\top \mathbf{z}|^2_{(i)} \le (1+\epsilon_1)\mathcal{V}\left(\frac{t'}{t}\right) H(\mathbf{w}_1, \cdots, \mathbf{w}_d) + 2\sqrt{(1+\epsilon_2)\overline{c}dH(\mathbf{w}_1, \cdots, \mathbf{w}_d)} + \overline{c}.$$

Letting $t' = t$ we immediately have the following corollary.

Corollary 8 *If there exists $\epsilon, \overline{c}$ such that (I) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{x}|^2 - 1\right| \le \epsilon$; and (II) $\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{n}_i|^2 \le \overline{c}$, then for any $\mathbf{w}_1, \cdots, \mathbf{w}_d \in \mathcal{S}_m$ the following holds:*

$$(1-\epsilon)H(\mathbf{w}_1, \cdots, \mathbf{w}_d) - 2\sqrt{(1+\epsilon)\overline{c}dH(\mathbf{w}_1, \cdots, \mathbf{w}_d)}$$
$$\le \sum_{j=1}^{d} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}_j^\top \mathbf{z}_i|^2 \le (1+\epsilon)H(\mathbf{w}_1, \cdots, \mathbf{w}_d) + 2\sqrt{(1+\epsilon)\overline{c}dH(\mathbf{w}_1, \cdots, \mathbf{w}_d)} + \overline{c}.$$

4.2 Step 2

The next 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 high 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, $\mathcal{Y}(s) = \mathcal{Z}(s) \bigcup \mathcal{O}(s)$. Observe that $|\mathcal{Y}(s)| = n - s$. Let $\overline{r}(s) = \mathcal{Y}(s-1) \backslash \mathcal{Y}(s)$, i.e., the point removed at stage s. Let $\mathbf{w}_1(s), \ldots, \mathbf{w}_d(s)$ be the 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:

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

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 9 *With probability at least $1-\gamma$, event $\mathcal{E}(s)$ is true for some $1 \le s \le s_0$, where*

$$s_0 \triangleq (1+\epsilon)\frac{(1+\kappa)\lambda n}{\kappa}; \quad \epsilon = \frac{16(1+\kappa)\log(1/\gamma)}{\kappa\lambda n} + 4\sqrt{\frac{(1+\kappa)\log(1/\gamma)}{\kappa\lambda n}}.$$

Remark: When κ and λ are fixed, we have $s_0/n \to (1+\kappa)\lambda/\kappa$. Therefore, $s_0 \leq t$ for $(1+\kappa)\lambda < \kappa(1-\lambda)$ and n large.

When $s_0 \geq n$, Theorem 9 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 $\mathcal{Y}(s)$, performing a PCA is deterministic, $\mathcal{E}(s+1) \in \mathcal{F}_s$.

Theorem 10 *If the complimentary of $\mathcal{E}(s)$, denoted by $\mathcal{E}^c(s)$, 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

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

which is equivalent to

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

Note that

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

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 $\sum_{j=1}^{d} (\mathbf{w}_j(s)^\top \mathbf{y})^2$, and independent of other events. ■

As a consequence of this theorem, we can now prove Theorem 9. The intuition is rather straightforward: if the events were independent from one step to the next, then since "expected 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 $s \leq s_0$. 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

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

Lemma 11 *$\{X_s, \mathcal{F}_s\}$ is a supermartingale.*

Proof: 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 full version (Xu et al., 2010). ■

From here, the proof of Theorem 9 follows straightforwardly.

Proof: Note that

$$\Pr\left(\bigcap_{s=1}^{s_0} \mathcal{E}(s)^c\right) = \Pr\left(T > s_0\right) \leq \Pr\left(X_{s_0} \geq \frac{\kappa s_0}{1+\kappa}\right) = \Pr\left(X_{s_0} \geq (1+\epsilon)\lambda n\right), \tag{3}$$

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 long version (Xu et al., 2010). ■

4.3 Step 3

Let $\overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_d$ be the eigenvectors corresponding to the d largest eigenvalues of $AA^\top$, i.e., the optimal solution. Let $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$ be the output of the algorithm. Let $\mathbf{w}_1(s), \ldots, \mathbf{w}_d(s)$ be the candidate solution at stage s. Recall that $H(\mathbf{w}_1, \cdots, \mathbf{w}_d) \triangleq \sum_{j=1}^d \|\mathbf{w}_j^\top A\|^2$, and for notational simplification, let $\overline{H} \triangleq H(\overline{\mathbf{w}}_1, \cdots, \overline{\mathbf{w}}_d)$, $H_s \triangleq H(\mathbf{w}_1(s), \ldots, \mathbf{w}_d(s))$, and $H^* \triangleq H(\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*)$.

The statement of the finite-sample and asymptotic theorems (Theorems 1 and 2, respectively) lower bound the expressed variance, E.V., which is the ratio $H^*/\overline{H}$. The final part of the proof accomplishes this in two main steps. First, Lemma 12 lower bounds H_s in terms of $\overline{H}$, where s is some step for which $\mathcal{E}(s)$ is true, i.e., the principal components found by the s^{th} step of the algorithm are "good." By Theorem 9, we know that there is a "small" such s, with high probability. The final output of the algorithm, however, is only guaranteed to have a high value of the robust variance estimator, $\overline{V}$ — that is, even if there is a "good" solution at some intermediate step s, we do not necessarily have a way of identifying it. Thus, the next step, Lemma 13, lower bounds the value of H^* in terms of the value H of *any* output $\mathbf{w}_1', \ldots, \mathbf{w}_d'$ that has a smaller value of the robust variance estimator.

We give the statement of all the intermediate results, leaving the details to the full version (Xu et al., 2010).

Lemma 12 *If $\mathcal{E}(s)$ is true for some $s \le s_0$, and there exists $\epsilon_1, \epsilon_2, \overline{c}$ such that (I) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t-s_0} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}\left(\frac{t-s_0}{t}\right)\right| \le \epsilon_1$; (II) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{x}_i|^2 - 1\right| \le \epsilon_2$; (III) $\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{n}_i|^2 \le \overline{c}$, then*

$$\frac{1}{1+\kappa}\left[(1-\epsilon_1)\mathcal{V}\left(\frac{t-s_0}{t}\right)\overline{H} - 2\sqrt{(1+\epsilon_2)\overline{c}d\overline{H}}\right] \le (1+\epsilon_2)H_s + 2\sqrt{(1+\epsilon_2)\overline{c}dH_s} + \overline{c}.$$

Lemma 13 *Fix a $\hat{t} \le t$. If $\sum_{j=1}^d \overline{V}_{\hat{t}}(\mathbf{w}_j) \ge \sum_{j=1}^d \overline{V}_{\hat{t}}(\mathbf{w}_j')$, and there exists $\epsilon_1, \epsilon_2, \overline{c}$ such that (I) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}(\frac{\hat{t}}{t})\right| \le \epsilon_1$; (II) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{\hat{t}-\frac{\lambda t}{1-\lambda}} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\right| \le \epsilon_1$; (III) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{x}_i|^2 - 1\right| \le \epsilon_2$; (IV) $\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{n}_i|^2 \le \overline{c}$, then*

$$\begin{aligned}&(1-\epsilon_1)\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right) H(\mathbf{w}_1' \cdots, \mathbf{w}_d') - 2\sqrt{(1+\epsilon_2)\overline{c}dH(\mathbf{w}_1' \cdots, \mathbf{w}_d')} \\ \le &(1+\epsilon_1)H(\mathbf{w}_1 \cdots, \mathbf{w}_d)\mathcal{V}\left(\frac{\hat{t}}{t}\right) + 2\sqrt{(1+\epsilon_2)\overline{c}dH(\mathbf{w}_1 \cdots, \mathbf{w}_d)} + \overline{c}.\end{aligned}$$

Theorem 14 *If $\bigcup_{s=1}^{s_0} \mathcal{E}(s)$ is true, and there exists $\epsilon_1 < 1, \epsilon_2, \overline{c}$ such that (I) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t-s_0} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}(\frac{t-s_0}{t})\right| \le \epsilon_1$; (II) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}(\frac{\hat{t}}{t})\right| \le \epsilon_1$; (III) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{\hat{t}-\frac{\lambda t}{1-\lambda}} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\right| \le \epsilon_1$; (IV) $\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{x}_i|^2 - 1\right| \le \epsilon_2$; (V) $\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{t}\sum_{i=1}^{t} |\mathbf{w}^\top \mathbf{n}_i|^2 \le \overline{c}$, then*

$$\begin{aligned}\frac{H^*}{\overline{H}} \ge &\frac{(1-\epsilon_1)^2\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\mathcal{V}\left(\frac{t-s_0}{t}\right)}{(1+\epsilon_1)(1+\epsilon_2)(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \\ &- \left[\frac{(2\kappa+4)(1-\epsilon_1)\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\sqrt{(1+\epsilon_2)\overline{c}d} + 4(1+\kappa)(1+\epsilon_2)\sqrt{(1+\epsilon_2)\overline{c}d}}{(1+\epsilon_1)(1+\epsilon_2)(1+\kappa)\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right](\overline{H})^{-1/2} \qquad (4) \\ &- \left[\frac{(1-\epsilon_1)\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda}{1-\lambda}\right)\overline{c} + (1+\epsilon_2)\overline{c}}{(1+\epsilon_1)(1+\epsilon_2)\mathcal{V}\left(\frac{\hat{t}}{t}\right)}\right](\overline{H})^{-1}.\end{aligned}$$

By bounding all diminishing terms in the r.h.s. of (4), it reduces to Theorem 1. Theorem 2 follows immediately.

5 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 removes 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 to the high-dimensional case were 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 intractable. This motivates our work of proposing a new robust PCA algorithm that takes into account the inherent difficulty in analyzing high-dimensional data.

Acknowledgement

We thank the anonymous reviewers for their insightful comments. Constantine Caramanis was partially supported by NSF grants EFRI-0735905, CNS-0721532, CNS-0831580, and DTRA grant HDTRA1-08-0029. Shie Mannor was supported by a Horev Fellowship, by the European Union under a Marie-Curie Reintegration Grant and by the Israel Science Foundation (contract 890015).

References

Barnett, V. (1976). The ordering of multivariate data. *Journal of Royal Statistics Society Series, A*, *138*, 318–344.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. Wiley, New York.

Bebbington, A. (1978). A method of bivariate trimming for robust estimation of the correlation coefficient. *Applied Statistics*, *27*, 221–228.

Brubaker, S. C. (2009). Robust PCA and clustering on noisy mixtures. *Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms* (pp. 1078–1087).

Candès, E., Li, X., Ma, Y., & Wright, J. (2009). Robust pricipal component analysis? ArXiv:0912.3599.

Candès, E., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*, 717–772.

Chandrasekaran, V., Sanghavi, S., Parrilo, P., & Willsky, A. (2009). Rank-sparsity incoherence for matrix decomposition. ArXiv:0906.2220.

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

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

Croux, C., & Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, *95*, 206–226.

David, H. (1981). *Order statistics*. Wiley, New York.

Davidson, K., & Szarek, S. (2001). Local operator theory, random matrices and banach spaces. *Handbook on the Geometry of Banach Spaces* (pp. 317–366). Elsvier.

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

De la Torre, F., & Black, M. J. (2003). A framework for robust subspace learning. *International Journal of Computer Vision*, *54*, 117–142.

Dempster, A., & Gasko-Green, M. (1981). New tools for residual analysis. *The Annals of Statistics*, *9*, 945–959.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*, 531–545.

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

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University.

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

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124.

Helbling, J. (1983). *Ellipsoïdes minimaux de couverture en statistique multivariée*. Doctoral dissertation, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons, New York.

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

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

Jolliffe, I. T. (1986). *Principal component analysis*. Springer Series in Statistics, Berlin: Springer.

Li, G., & Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo. *Journal of the American Statistical Association*, *80*, 759–766.

Maronna, R. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, *4*, 51–67.

Mendelson, S., & Pajor, A. (2006). On singular values of matrices with independent rows. *Bernoulli*, *12*, 761–773.

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

Recht, B. (2009). A simpler approach to matrix completion. ArXiv: 0910.0651.

Recht, B., Fazel, M., & Parrilo, P. (2010). Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. To appear in *SIAM Review*.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871–880.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. MIT Press.

Schölkopf, B., Smola, A. J., & Müller, K. R. (1999). Kernel principal component analysis. *Advances in kernel Methods – Support Vector Learning* (pp. 327–352). MIT Press, Cambridge, MA.

Titterington, D. (1978). Estimation of correlation coefficients by ellipsoidal trimming. *Applied Statistics*, *27*, 227–234.

Xu, H., Caramanis, C., & Mannor, S. (2010). Principal component analysis with contaminated data: The high dimensional case. ArXiv: 1002.4658.

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

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