Principal Component Analysis with Contaminated Data: The High Dimensional Case*

Huan Xu[†], Constantine Caramanis [‡], and Shie Mannor[§]

Abstract

We consider the dimensionality-reduction problem (finding a subspace approximation of observed data) for contaminated data in the high dimensional regime, where the 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, 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 [1, 2]. Today, it is common practice that observations on individual instances are 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" [3], a pre-processing routine which nonlinearly maps the observations into a (possibly infinite dimensional) Hilbert space, transforms virtually every data set to a high dimensional one. Efforts of extending traditional statistical tools (designed for low dimensional case) into this high-dimensional regime are generally unsuccessful. This fact has stimulated research on formulating fresh data-analysis techniques able to cope with such a "dimensionality explosion."

In this paper, we consider a high-dimensional counterpart of Principal Component Analysis (PCA) that is robust to the existence of corrupted or contaminated data. In our setup, a low dimensional Gaussian signal is 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. Then, *a constant fraction of the points are arbitrarily corrupted* in a perhaps non-probabilistic manner. We refrain from calling these "outliers" to emphasize that their distribution is entirely arbitrary, rather than from the tails of any particular distribution, e.g., the noise distribution. We call the remaining points "authentic."

Work on PCA dates back as early as [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, and image processing [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.

It is well known that such analysis is extremely sensitive to outlying, or corrupted, measurements. Indeed, one aberrant observation is sufficient to cause arbitrarily large changes in the covariance or correlation matrix, and hence the corresponding PCs.

In the low-dimensional regime where the observations significantly outnumber the variables of each observation, several robust PCA algorithms have been proposed (e.g.,[6, 7, 8, 9, 10, 11, 12]). These algorithms can be roughly divided into two classes: (i) performing a standard PCA on a robust estimation of the covariance or correlation matrix; (ii) maximizing (over all unit-norm $\mathbf{w}$) some $r(\mathbf{w})$ that is a robust estimate of the variance of univariate data obtained by projecting the observations onto direction $\mathbf{w}$. Both approaches encounter serious difficulties when applied to high-dimensional data-sets:

- There are not enough observations to robustly estimate the covariance or correlations matrix. For example, the widely-used MVE estimator [13], which treats the Minimum Volume Ellipsoid that covers half of the observations as the covariance estimation, is ill-posed in the

*Eligible for the Mark Fulk Award.

[†]Department of Electrical and Computer Engineering, McGill University, xuhuan@cim.mcgill.ca

[‡]Department of Electrical and Computer Engineering, The University of Texas at Austin, cmcaram@ece.utexas.edu

[§]Department of Electrical and Computer Engineering, McGill University and Department of Electrical Engineering, Technion shie.mannor@mcgill.ca

high-dimensional case. Indeed, to the best of our knowledge, the assumption that observations far outnumber dimensionality seems crucial for those robust variance estimators to achieve statistical consistency.

- Unlike standard PCA that has a polynomial computation time, the maximization of $r(\mathbf{w})$ is generally a non-convex problem, and becomes extremely hard to solve or approximate as the dimensionality of $\mathbf{w}$ increases. In fact, the number of the local maxima grows so fast that it is effectively impossible to find a sufficiently good solution using gradient-based algorithms with random re-initialization.

In contrast to these approaches, we propose a High-dimensional Robust PCA (HR-PCA) algorithm that takes into account the inherent difficulty in analyzing the high dimensional data. In particular, the algorithm we propose here is tractable, provably robust to corrupted points, easily kernelizable, and asymptotically optimal.

The proposed algorithm takes an "actor-critic" form: we apply standard PCA in order to find a set of candidate directions. These directions are then subjected to a hypothesis test, that uses a computationally efficient one-dimensional robust variance estimate. This hypothesis test determines if the variance is due to corrupted data, or indeed the "authentic" points. In case of the latter, the algorithm has found a true PC. In case of the former, we use a randomized point removal scheme, that guarantees quick termination of our algorithm with deviation guarantees on the PCs it ultimately reports, from the true PCs.

One notable difference between this paper and previous robust PCA work is how we measure the robustness of an algorithm. The traditional robustness measurement is the so-called "breakdown point" [14], i.e., the percentage of corrupted points that can make the output of the algorithm *arbitrarily* bad. This is an indirect measurement: except that the error is not unlimited, there is no guarantee that the output is *good enough*, even when the algorithm does not breakdown. In contrast, we directly investigate the "robust performance" of the algorithm, i.e., the performance gap between the output of the algorithm and the optimum, as a function of the fraction of corrupted points. Therefore, such direct measurement provides an explicit guarantee of the performance of the algorithm, which we regard to be of importance in practice.

The paper is organized as follows: In Section 2 we present the setup of the problem, the hypothesis test, and then the HR-PCA algorithm including the randomized point removal scheme. Based on some technical results established in Section 3, we show the validity of HR-PCA in Section 4 by providing a bound on the probability that our algorithm removes a corrupted point at any given iteration, and then using this to bound the running time of the algorithm, and finally to give finite sample and asymptotic performance guarantees. Section 5 is devoted to the kernelization of HR-PCA. We provide some numerical experiment results in Section 6.

Notation: Capital letters and boldface letters are used to denote matrices and vectors, respectively. $\Phi(\cdot)$ stands for the cumulative distribution function of $\mathcal{N}(0,1)$ and we let $\Phi^{-1}(c)$ be $-\infty$ and $+\infty$ for $c \leq 0$ and $c \geq 1$ respectively. $\Psi(\cdot)$ is the *Tracy-Widom* distribution of order one (c.f [15, 2]), implemented using a numerical lookup table. A $k \times k$ unit matrix is denoted by I_k. The largest eigenvalue of a symmetric matrix C is represented as $\lambda_{\max}(C)$. For $c \in \mathbb{R}$, $[c]^+ \triangleq \max(0, c)$.

As this paper is notationally very heavy, we introduce a number of parameters to simplify long expressions, in the hopes of highlighting the key elements. While we introduce the parameters at the appropriate parts of the text, we keep the following convention, to facilitate the reader's job. Parameters which have a physical meaning, such as the largest singular value of a matrix, or the fraction of corrupted points, etc., all have a '$^-$'. Parameters which are introduced as "slack" factors in order to deal with finite-sample estimates (and go to zero asymptotically) all have a '$\hat{}$'. Finally, parameters synthesized from the two categories above, and introduced simply to shorten and simplify expressions, all have a '$\tilde{}$'.

2 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 2.1. As discussed in the Introduction, HR-PCA follows an "actor-critic" approach in which a robust univariate variance estimator serves as a hypothesis test, to evaluate the robustness of PCs found. We call this the "Sensitivity Test" and provide its formulation in Section 2.2. The HR-PCA algorithm is then given in Section 2.3.

2.1 Problem Setup

We consider the following problem:

- 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 $\mathbf{x}_i$ (the "signal") and $\mathbf{n}_i$ (the "noise") are independent realizations of random variables $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I_d)$ and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, I_m)$ respectively. The matrix $A \in \mathbb{R}^{m \times d}$ is unknown.
- The corrupted data are denoted $\mathbf{o}_1, \ldots, \mathbf{o}_{n-t} \in \mathbb{R}^m$ and they are arbitrary (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".

We denote the fraction of corrupted points by $\overline{\eta} \triangleq n - t/n$. In this paper, we focus on the case where $n \sim m \gg d$ and $\lambda_{\max}(A^\top A) \gg 1$. That is, the number and dimensionality of observations are of the same magnitude, and much larger than the dimensionality of $\mathbf{x}$; the leading eigenvalue of $A^\top A$ is significantly larger than 1.

For a set of orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$, performance is measured by the *Expressed Variance*

$$\text{E.V.} \triangleq \frac{\sum_{i=1}^{d} \mathbf{v}_i^\top A A^\top \mathbf{v}_i}{\sum_{i=1}^{d} \mathbf{v}_i^{*\top} A A^\top \mathbf{v}_i^*},$$

where $\{\mathbf{v}_1^*, \ldots, \mathbf{v}_d^*\}$ are the largest d eigenvectors of $AA^\top$ (i.e., the desired PCs). Notice that the maximum of E.V. equals 1, and is achieved by recovering the span of the true

PCs $\{\mathbf{v}_1^*, \ldots, \mathbf{v}_d^*\}$. In addition, when $d = 1$, the Expressed Variance relates to another natural performance metric — the angle between $\mathbf{v}_1$ and $\mathbf{v}_1^*$ — since $E.V.(\mathbf{v}_1) = \cos^2(\angle(\mathbf{v}_1, \mathbf{v}_1^*))$ (see Figure 1). When $d > 1$, such geometric interpretation no longer exists since the angle between two subspaces is not well defined. The Expressed Variance represents the portion of signal $A\mathbf{x}$ being expressed by $\mathbf{v}_1, \ldots, \mathbf{v}_d$. Equivalently, $1 - E.V.$ is the reconstruction error of the signal.

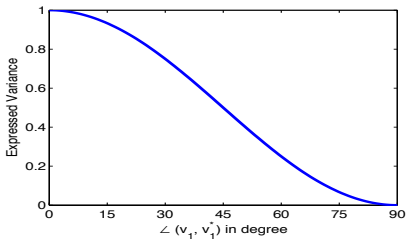


Figure 1: Expressed Variance vs Angle for d=1.

While we give finite-sample results, our main theorem gives the asymptotic performance of HR-PCA when *the dimension and the number of observations grow together* to infinity. To be more precise, our asymptotic setting is as follows. Suppose there exists a sequence of sample sets $\{\mathcal{Y}(j)\}_j = \{\mathcal{Y}(1), \mathcal{Y}(2), \ldots\}$, 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:

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

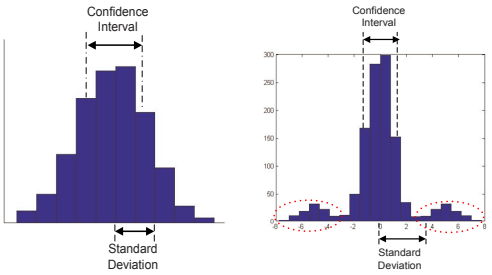
2.2 Sensitivity Test

We present the formulation of the "Sensitivity Test" in this subsection. This test is based on evaluating the "θ confidence interval" of a collection of scalar values, i.e., the shortest interval containing a θ fraction of the scalars. For unit-norm $\mathbf{w} \in \mathbb{R}^m$, let

$$\bar{l}_{\mathbf{w}} \triangleq l(0.5 + \frac{\overline{\eta}}{2}, \mathbf{w}^\top \mathbf{y}_1, \ldots, \mathbf{w}^\top \mathbf{y}_n),$$

which is the $0.5 + \overline{\eta}/2$ confidence interval for the points projected on the direction $\mathbf{w}$. This confidence interval $\bar{l}_{\mathbf{w}}$ is an estimator of standard deviation robust to the existence of corrupted points [14]. Once PCA outputs directions of largest variance, for each direction we use this estimator to determine if the confidence interval is consistent with the observed variance, and hence if the variance is a phenomenon due to the authentic points, or due to the corrupted data. We refer to Figure **??** for an illustration. This figure illustrates a sensitivity test used often in practice, whereby the variance, σ^2, of a corrupted sample set along a direction $\mathbf{w}$ is deemed consistent with the confidence interval or not, according to the rule:

$$\begin{cases} \text{consistent} & \text{if } (1+\sqrt{\overline{\eta}})(1-\overline{\eta})\mathbf{H}_{\mathbf{w}} \ge \sigma^2 \\ \text{inconsistent} & \text{if } (1+\sqrt{\overline{\eta}})(1-\overline{\eta})\mathbf{H}_{\mathbf{w}} < \sigma^2 \end{cases},$$



No Corrupted Data Some Corrupted Data

Figure 2: Illustration of the Sensitivity Test

where $\mathbf{H}_{\mathbf{w}} \triangleq \left(\frac{l_{\mathbf{w}}}{2\Phi^{-1}(0.75)}\right)^2$.

Because we are interested in a sensitivity test at each iteration, when some number s of the points have been removed in previous iterations, and because we require finite sample bounds in the sequel, we need a modification of the above sensitivity test that incorporates some slack factors. We note that asymptotically, our slack factors disappear, and our sensitivity test corresponds with the one given above. Before providing the exact form of the Sensitivity Test, we define the following terms to simplify the expressions. Again we use our convention whereby parameters with a physical meaning have a '$\bar{}$', slack parameters which go to zero asymptotically have a '$\hat{}$' and parameters synthesized from the two categories above, and introduced simply to shorten and simplify expressions, all have a '$\tilde{}$'.

In the quantities below, the subscript 't' corresponds to the number of authentic points, and therefore is a quantity that in the asymptotic analysis will go to infinity. The subscript 'δ' will be a probability parameter we use in the sequel to control the probability of finite sample deviation results, and therefore will be taken to be a very small positive number.

$$\begin{aligned} \overline{\sigma}_1 &\triangleq \sqrt{\lambda_{\max}(AA^\top)}; \\ \overline{\lambda}_{t,\delta} &\triangleq \frac{4}{1-\overline{\eta}} + \frac{2[\Psi^{-1}(1-\delta)]^+}{\sqrt{(1-\overline{\eta})t}}. \end{aligned}$$

The first quantity above implicitly has a t-index, since the size of the matrix A is fixed to t. Note further that by assumption, $\overline{\sigma}_1$ goes to infinity as $t \to \infty$. The second quantity above bounds the variance of the noise realizations, as shown in Theorem 3 below.

$$\begin{aligned} \hat{c}_{t,\delta} &\triangleq \sqrt{\frac{6d^2}{t}\left(\ln 2d^2 + \ln\frac{1}{\delta}\right)}; \\ \hat{h}_{t,\delta} &\triangleq \sqrt{\frac{8d+8}{t}\ln\frac{t}{d+1} + \frac{8}{t}\ln\frac{8}{\delta}}; \\ \hat{\phi}_{t,\delta} &\triangleq 2\Phi^{-1}(0.75) - 2\Phi^{-1}(0.75 - \hat{h}_{t,\delta}); \\ \tilde{v}_{t,\delta} &\triangleq \hat{c}_{t,\delta}\overline{\sigma}_1^2 + [1 + \hat{c}_{t,\delta}/2]\sqrt{\overline{\lambda}_{t,\delta}}\overline{\sigma}_1 + \overline{\lambda}_{t,\delta}; \\ \overline{O}_{\mathbf{w}} &\triangleq \min_{c \ge \sqrt{2\overline{\lambda}_{t,\delta}}} \left\{ \frac{l_{\mathbf{w}} + \hat{\phi}_{t,\delta}\overline{\sigma}_1 + 2c}{2\Phi^{-1}(0.75 - \frac{\overline{\lambda}_{t,\delta}}{2c^2})} \right\}; \\ \overline{H}_{\mathbf{w}} &\triangleq \overline{O}_{\mathbf{w}}^2 + \tilde{v}_{t,\delta}. \end{aligned}$$

Finally, we can define our sensitivity test.

Definition 1 *If s points have been removed, and the empirical variance in a direction $\mathbf{w}$ is σ^2, then the* Sensitivity Test $\mathbb{H}$ *is defined as*

$$\mathbb{H}(\mathbf{w},\sigma,s) \triangleq \begin{cases} \textit{consistent}, & \textit{if } \frac{(1+\sqrt{\overline{\eta}})(1-\overline{\eta})n}{n-s}\overline{H}_{\mathbf{w}} \geq \sigma^2; \\ \textit{inconsistent}, & \textit{if } \frac{(1+\sqrt{\overline{\eta}})(1-\overline{\eta})n}{n-s}\overline{H}_{\mathbf{w}} < \sigma^2. \end{cases}$$

Note that $\overline{O}$ takes the place of the term $\frac{l_{\mathbf{w}}}{2\Phi^{-1}(0.75)}$ in the expression given for the first sensitivity test above.

2.3 Main Algorithm

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

HR-PCA

Input: Contaminated sample-set $\mathcal{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_n\} \subset \mathbb{R}^m$, $\overline{\eta}$, d, δ, $\overline{\sigma}_1$.

Output: $\mathbf{v}_1,\ldots,\mathbf{v}_d$.

Algorithm:

1. Let $\sqrt{\overline{\eta}} := \sqrt{\overline{\eta}}$; $\hat{\mathbf{y}}_i := \mathbf{y}_i$ for $i = 1,\ldots n$; $s := 0$.
2. Compute the empirical variance matrix
$$\hat{\Sigma} := \frac{1}{n-s}\sum_{i=1}^{n-s} \hat{\mathbf{y}}_i\hat{\mathbf{y}}_i^\top.$$
3. Let $\hat{\sigma}_1^2,\ldots,\hat{\sigma}_d^2$ and $\mathbf{v}_1,\ldots,\mathbf{v}_d$ be the d largest eigenvalues and the corresponding eigenvectors of $\hat{\Sigma}$.
4. If there is a $j \in \{1,\ldots,d\}$ such that *Sensitivity Test* $\mathbb{H}(\mathbf{v}_j,\hat{\sigma}_j,s)$ fails, do the following:
   - randomly remove a point from $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s}$ according to
   $$\Pr(\hat{\mathbf{y}}_i \text{ is removed}) \propto (\mathbf{v}_j^\top\hat{\mathbf{y}}_i)^2;$$
   - denote the remaining points by $\{\hat{\mathbf{y}}_i\}_{i=1}^{n-s-1}$;
   - $s := s+1$, go to Step 2.
5. Output $\mathbf{v}_1,\ldots,\mathbf{v}_d$. End.

In each iteration, HR-PCA finds a set of directions maximizing the empirical variance of the points (i.e., of authentic and corrupted samples). If all directions pass the Sensitivity Test, then the variances of "authentic samples" projected on them must be *close to being the largest*, and hence the chosen directions are close to the true PCs. If the Sensitivity Test fails, then the corrupted points must have a large influence on the variance in this direction. In this case, the PC is not selected, and a point is removed in proportion to its variance. We show that this proportional removal guarantees a minimum probability that a corrupted point will be removed.

The correctness of HR-PCA is shown in the following sections. We outline here the main theorem providing an asymptotic lower bound of the performance (illustrated in Figure 3). This is based on a finite-sample result, which we state and prove in Section 4.3.

Theorem 2 *If Equation (1) holds, and $\overline{\eta}(j) \to \overline{\eta}^*$, then the following holds in probability with $j \uparrow \infty$,*

$$\frac{\sum_{q=1}^{d}\mathbf{v}_q(j)^\top(A(j)A(j)^\top)\mathbf{v}_q(j)}{\sum_{q=1}^{d}\mathbf{v}_q^*(j)^\top(A(j)A(j)^\top)\mathbf{v}_q^*(j)} \geq \frac{\int_{-\tilde{\zeta}^*}^{\tilde{\zeta}^*}\frac{x^2}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})dx}{1+\sqrt{\overline{\eta}^*}}\left(\frac{\Phi^{-1}(0.75)}{\Phi^{-1}(0.75+\frac{\overline{\eta}}{2-2\overline{\eta}^*})}\right)^2, \quad (2)$$

where $\tilde{\zeta}^* \triangleq \Phi^{-1}\left(1-\frac{\sqrt{\overline{\eta}^*}}{1-\sqrt{\overline{\eta}^*}}\right)$.

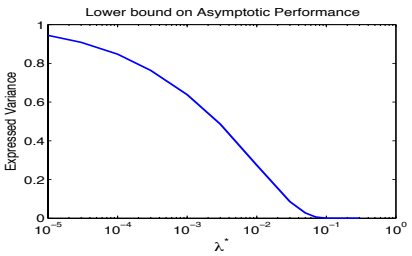


Figure 3: Lower Bound of Asymptotic Performance.

Remark 1 If $\overline{\eta}(j) \downarrow 0$ (e.g., there are a fixed *number* of corrupted points), then the right-hand-side of Inequality (2) equals 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.

3 Technical Results: Uniform Convergence

This section is devoted to establishing the uniform (w.r.t. all directions $\mathbf{w} \in \mathbb{R}^m$) convergence properties of the sample variance and the "confidence interval" for authentic samples $\{\mathbf{z}_1,\ldots,\mathbf{z}_t\}$ (Theorems 4 and 5 respectively). These results are used as technical lemmas in proving the validity of HR-PCA in Section 4. Due to space constraints, all proofs in this section are omitted and deferred to an online appendix, [16].

Consider the following events:

Condition (A): $\{\lambda_{\max}(1/t\sum_{i=1}^{t}\mathbf{n}_i\mathbf{n}_i^\top) \leq \overline{\lambda}_{t,\delta}\}$

Condition (B):

$$\left\{\sup_{\mathbf{w}\in\mathbb{R}^d,\,\|\mathbf{w}\|=1}\left|\mathbf{w}^\top(\frac{1}{t}\sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^\top - I_d)\mathbf{w}\right| \leq \hat{c}_{t,\delta}\right\};$$

Condition (C):

$$\left\{2\Phi^{-1}\left(\frac{1+\theta}{2}-\hat{h}_{t,\delta}\right) \leq l(\theta,\mathbf{w}^\top\mathbf{x}_1,\ldots,\mathbf{w}^\top\mathbf{x}_t) \leq 2\Phi^{-1}(\frac{1+\theta}{2}+\hat{h}_{t,\delta}), \quad \forall\|\mathbf{w}\|_2=1,\forall\theta\in[0,1]\right\}.$$

Theorem 3 *For sufficiently large t and m:*

1. *Condition (A) holds with probability at least $1-\delta$;*

2. *Condition (B) holds with probability at least* $1-\delta$*;*

3. *Condition (C) holds with probability at least* $1-\delta$*.*

The validity of Condition (A) follows from a lemma in [2]. Conditions (B) and (C) are finite dimensional uniform convergence results that follow from VC-dimension style arguments. (See [16] for the full proofs).

The next two theorems give analogs of Condition (B) and Condition (C) but for the high dimensional points.

Theorem 4 *Under Conditions (A) and (B), and* $n \geq 4$*, the average variance along direction* $\mathbf{w}$*, of the authentic points, has distance from the size of* $(AA^\top + I)$ *in the* $\mathbf{w}$ *direction bounded as follows:*

$$\sup_{\mathbf{w}\in\mathbb{R}^m, \|\mathbf{w}\|_2=1} \left|\mathbf{w}^\top \left(\frac{1}{t}\sum_{i=1}^{t} \mathbf{z}_i\mathbf{z}_i^\top\right)\mathbf{w} - \mathbf{w}^\top(AA^\top + I_m)\mathbf{w}\right|$$
$$\leq \hat{c}_{t,\delta}\overline{\sigma}_1^2 + \left[1+\frac{\hat{c}_{t,\delta}}{2}\right]\sqrt{\overline{\lambda}_{t,\delta}}\overline{\sigma}_1 + \overline{\lambda}_{t,\delta} - 1. \tag{3}$$

Theorem 5 *Under Conditions (A) and (C), for any* $\|\mathbf{w}\|_2 = 1$ *and* $\theta \in (0,1)$*, the* θ*-confidence interval of all the authentic points projected on direction* $\mathbf{w}$ *is sandwiched as follows:*

$$\begin{aligned}
&\sup_{c>0}\Big\{2\Phi^{-1}(\frac{1+\theta}{2} - \frac{\overline{\lambda}_{t,\delta}}{2c^2})\sqrt{\mathbf{w}^\top AA^\top\mathbf{w}} \\
&\quad - (2\Phi^{-1}(\frac{1+\theta}{2}) - 2\Phi^{-1}(\frac{1+\theta}{2} - \hat{h}_{t,\delta}))\overline{\sigma}_1 - 2c\Big\} \\
&\leq l(\theta, \mathbf{w}^\top\mathbf{z}_1, \ldots, \mathbf{w}^\top\mathbf{z}_t) \\
&\leq \inf_{c>0}\Big\{2\Phi^{-1}(\frac{1+\theta}{2} + \frac{\overline{\lambda}_{t,\delta}}{2c^2})\sqrt{\mathbf{w}^\top AA^\top\mathbf{w}} \\
&\quad + (2\Phi^{-1}(\frac{1+\theta}{2} + \hat{h}_{t,\delta}) - 2\Phi^{-1}(\frac{1+\theta}{2}))\overline{\sigma}_1 + 2c\Big\}.
\end{aligned} \tag{4}$$

Note that these bounds are, up to addition of appropriate slack factors, the θ-confidence bounds for the original low-dimensional points $\{\mathbf{x}_i\}$ as given in Condition C, elongated by $\sqrt{\mathbf{w}AA^\top\mathbf{w}}$.

4 Correctness of HR-PCA

Based on the results presented in the previous section, we show in this section the correctness of HR-PCA, i.e., with a high probability, the subspace spanned by the output $\mathbf{v}_1, \ldots, \mathbf{v}_d$ is a good approximation (in the sense of the Expressed Variance) of that spanned by $\mathbf{v}_1^*, \ldots, \mathbf{v}_d^*$.

In Section 4.1 we lower bound the probability of removing a corrupted point in each iteration. We then show that the number of iterations is small with high probability in Section 4.2. We complete the argument in Section 4.3 by showing that when HR-PCA stops within a small number of iterations, the PCs found are good approximations of the desired ones.

Throughout this section, we assume without explicitly stating it in each theorem that Conditions (A), (B) and (C) hold simultaneously. As shown in Section 3, this occurs with probability is at least $1-3\delta$. We further assume $n \geq 4$.

4.1 Probability of Removing a Corrupted Point

In this section, we lower bound the probability of removing a corrupted point in each iteration of HR-PCA.

Theorem 6 *If the Sensitivity Test fails (i.e., it returns "inconsistent") in the* s^{th} *iteration, then*

$$\Pr(\text{The next removed point is corrupted}) > \frac{\sqrt{\overline{\eta}}}{1+\sqrt{\overline{\eta}}}.$$

To prove Theorem 6, we need the following lemma.

Lemma 7 *For all* $\|\mathbf{w}\|_2 = 1$, $\frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top\mathbf{z}_i)^2 \leq \overline{H}_{\mathbf{w}}$.

Proof: By definition, an interval with length $l_{\mathbf{w}}$ covers $(0.5+\overline{\eta}/2)n$ points in $\mathcal{Y}$, which implies that it covers at least $(0.5-\overline{\eta}/2)n = 0.5t$ authentic samples. Therefore,

$$l(0.5, \mathbf{w}^\top\mathbf{z}_1, \ldots, \mathbf{w}^\top\mathbf{z}_t) \leq l_{\mathbf{w}}. \tag{5}$$

Now, from Theorem 5, when (B) and (C) hold, we have for all θ and $\|\mathbf{w}\|_2 = 1$,

$$\begin{aligned}
&2\Phi^{-1}(\frac{1+\theta}{2} - \frac{\overline{\lambda}_{t,\delta}}{2c^2})\sqrt{\mathbf{w}^\top AA^\top\mathbf{w}} \\
&\leq l(\theta, \mathbf{w}^\top\mathbf{z}_1, \ldots, \mathbf{w}^\top\mathbf{z}_t) \\
&\quad + (2\Phi^{-1}(\frac{1+\theta}{2}) - 2\Phi^{-1}(\frac{1+\theta}{2} - \hat{h}_{t,\delta}))\overline{\sigma}_1 + 2c.
\end{aligned}$$

Taking $\theta = 0.5$, by applying (5) we get for all $c \geq \sqrt{2\overline{\lambda}_t}$,

$$\sqrt{\mathbf{w}^\top AA^\top\mathbf{w}} \leq (l_{\mathbf{w}} + \hat{\phi}_{t,\delta}\overline{\sigma}_1 + 2c)/2\Phi^{-1}(0.75 - \frac{\overline{\lambda}_{t,\delta}}{2c^2}).$$

Minimizing over c implies $\mathbf{w}^\top(AA^\top + I_m)\mathbf{w} \leq \overline{O}_{\mathbf{w}}^2 + 1$. Applying Theorem 4 completes the proof. ■

Proof of Theorem 6: First recall that our point removal strategy implies

$$\begin{aligned}
&\Pr(\text{The next removed point is corrupted}) \\
=&\frac{\sum_{\mathbf{y}_i\in\hat{I}_s}(\mathbf{v}_q{}^\top\mathbf{y}_i)^2}{\sum_{\mathbf{y}_i\in I_s}(\mathbf{v}_q{}^\top\mathbf{y}_i)^2 + \sum_{\mathbf{y}_i\in\hat{I}_s}(\mathbf{v}_q{}^\top\mathbf{y}_i)^2},
\end{aligned}$$

where I_s and $\hat{I}_s$ denote the set of remaining authentic samples and remaining corrupted points, respectively. We prove that

$$\sqrt{\overline{\eta}}\sum_{\mathbf{y}_i\in I_s}(\mathbf{v}_q{}^\top\mathbf{y}_i)^2 < \sum_{\mathbf{y}_i\in\hat{I}_s}(\mathbf{v}_q{}^\top\mathbf{y}_i)^2,$$

which will conclude the proof.

By definition, if the Sensitivity Test fails it must be that for some $q \in \{1, \ldots, d\}$

$$(1+\sqrt{\overline{\eta}})(1-\overline{\eta})n\overline{H}_{\mathbf{v}_q}/(n-s) < \hat{\sigma}_q^2. \tag{6}$$

At the s^{th} iteration, there are $n-s$ remaining points. We have $|I_s| + |\hat{I}_s| = n-s$, and $I_s \subseteq \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$. Therefore,

$$\sum_{\mathbf{y}_i\in I_s}(\mathbf{v}_q{}^\top\mathbf{y}_i)^2 \leq \sum_{i=1}^{t}(\mathbf{v}_q{}^\top\mathbf{z}_i)^2 \leq (1-\overline{\eta})n\overline{H}_{\mathbf{v}_q}, \tag{7}$$

where the last inequality follows from Lemma 7. Furthermore,

$$\begin{aligned} \hat{\sigma}_q^2 &= \frac{1}{n-s} \sum_{\mathbf{y}_i \in I_s \bigcup \hat{I}_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2 \\ \implies \quad (n-s)\hat{\sigma}_q^2 &= \sum_{\mathbf{y}_i \in I_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2 + \sum_{\mathbf{y}_i \in \hat{I}_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2. \end{aligned} \tag{8}$$

Substituting (7) and (8) into Inequality (6) leads to

$$\begin{aligned} &(1+\sqrt{\overline{\eta}}) \sum_{\mathbf{y}_i \in I_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2 < \sum_{\mathbf{y}_i \in I_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2 + \sum_{\mathbf{y}_i \in \hat{I}_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2 \\ \Rightarrow \quad &\sqrt{\overline{\eta}} \sum_{\mathbf{y}_i \in I_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2 < \sum_{\mathbf{y}_i \in \hat{I}_s} (\mathbf{v}_q{}^\top \mathbf{y}_i)^2. \end{aligned}$$

■

Theorem 6 implies that if the Sensitivity Tests fails, then there is at least one corrupted point remaining.

Since the probability of removing a corrupted point at any given iteration of the algorithm is $\sqrt{\overline{\eta}}/(1+\sqrt{\overline{\eta}})$, each iteration decreases the "expected number of corrupted points" by that amount. In the next section, we bound the probability that the algorithm fails to terminate before some particular iteration. For this we use twice the number of corrupted points divided by the expected reduction at a given iteration:

$$\bar{s}_0 \triangleq 2\left(\frac{1+\sqrt{\overline{\eta}}}{\sqrt{\overline{\eta}}}\right)\overline{\eta}n.$$

4.2 Number of Iterations

In this section, we show that with high probability, HR-PCA terminates quickly: within s_0 iterations. The key to the proof is the previous theorem, that each step removes a corrupted point with probability at least $\sqrt{\overline{\eta}}/(1+\sqrt{\overline{\eta}})$. If the event of corrupted point removal at subsequent iterations were independent, then the expected number of points removed by s iterations would be $s \cdot \sqrt{\overline{\eta}}/(1+\sqrt{\overline{\eta}})$, and since there are only $\overline{\eta}n$ corrupted points in total, the result would be straightforward. Instead, we must use a Martingale argument to arrive at the desired result.

By definition, HR-PCA terminates at step s if the Sensitivity Test succeeds after $s-1$ points have been removed. Let random variable $V(s)$ denote the number of corrupted points removed up to iteration s, inclusive. Define the following stochastic process:

$$X_s \triangleq \begin{cases} V(T) - \frac{\sqrt{\overline{\eta}}(T-1)}{1+\sqrt{\overline{\eta}}}, & \text{HR-PCA stoped at } T \le s; \\ V(s) - \frac{\sqrt{\overline{\eta}}s}{1+\sqrt{\overline{\eta}}}, & \text{Otherwise.} \end{cases}$$

Let $\mathcal{F}_s$ be the filtration generated by the set of events up to iteration s. Hence, X_s is measurable w.r.t. $\mathcal{F}_s$.

Lemma 8 *$\{X_s, \mathcal{F}_s, s = 1, \ldots, n\}$ is a sub-martingale.*

Proof: Observe that $X_s \in \mathcal{F}_s$ by definition of $\mathcal{F}_s$. We show that $\mathbb{E}(X_s|\mathcal{F}_{s-1}) \ge X_{s-1}$ by enumerating the following three cases.

Case 1: the algorithm has not terminated up to step $s-1$, and the hypothesis test of the s^{th} iteration fails. Thus by Theorem 6,

$$\begin{aligned} &\mathbb{E}(X_s - X_{s-1}|\mathcal{F}_{s-1}) \\ =&\Pr(\text{The next removed point is corrupted}) - \frac{\sqrt{\overline{\eta}}}{1+\sqrt{\overline{\eta}}} \ge 0. \end{aligned}$$

Case 2: the algorithm has not terminated up to step $s-1$, and the hypothesis test of the s^{th} iteration succeeds. Thus, the algorithm terminates at step s, i.e., no extra point will be removed. Hence $V(s) = V(s-1)$. By definition of X we have $X_s = X_{s-1}$ in this case.

Case 3: the algorithm terminates at step $T \le s-1$. Observe that $X_s = X_{s-1}$ in this case.

Combining all three cases shows that $\mathbb{E}(X_s|\mathcal{F}_{s-1}) \ge X_{s-1}$, which proves the lemma. ■

Theorem 9 *For all $s \ge (1+\kappa)\lambda n/\kappa$, we have*

$$\begin{aligned} &\Pr(\text{the algorithm does not terminate up to step } s) \le \\ &\exp\Big(\frac{-(\lambda n - \frac{\kappa s}{1+\kappa})^2}{8s}\Big). \end{aligned}$$

Proof: We prove the theorem by exploiting the deviation bound of a martingale process. Let $y_s \triangleq X_s - X_{s-1}$, where recall that $X_0 = 0$. 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 sub-martingale, $\mathbb{E}(y_s|y_1, \cdots, y_{s-1}) \ge 0$ a.s. Therefore, the following holds a.s.,

$$X_s = \sum_{i=1}^{s} y_i = \sum_{i=1}^{s} y'_i + \sum_{i=1}^{s} \mathbb{E}(y_i|y_1, \cdots, y_{i-1}) \ge \sum_{i=1}^{s} y'_i. \tag{9}$$

By definition, $|y_s| \le 1$, and hence $|y'_s| \le 2$. Now for any $\theta > 0$,

$$\begin{aligned} &\mathbb{E}\big\{\exp(\theta \sum_{i=1}^{s}(-y'_i))\big\} \\ &= \mathbb{E}\big\{\exp(\theta \sum_{i=1}^{s-1}(-y'_i))\big\}\mathbb{E}(\theta - y'_s| - y'_1, \cdots, -y'_{s-1}) \\ &\le \mathbb{E}\big\{\exp(\theta \sum_{i=1}^{s-1}(-y'_i))\big\}\exp(\theta^2|-y'_s|^2/2) \\ &= \mathbb{E}\big\{\theta \exp(\sum_{i=1}^{s-1}(-y'_i))\big\}\exp(2\theta^2). \end{aligned}$$

The inequality follows from Lemma 8.1 of [17]. By iteration we have

$$\mathbb{E}\big\{\exp(\theta \sum_{i=1}^{s}(-y'_i))\big\} \le \exp(2s\theta^2).$$

Using the Markov inequality, we have that for any $\epsilon > 0$,

$$\Pr(\sum_{i=1}^{s}(-y'_i) \ge s\epsilon) \le \exp(2s\theta^2 - \theta s\epsilon).$$

Taking the minimum over θ of the right hand side and applying (9) leads to

$$\Pr(X_s \le -s\epsilon) \le \exp(-s\epsilon^2/8).$$

Now notice that, if the algorithm does not terminate up to step s, we have $X_s \leq \lambda n - \kappa s/(1+\kappa)$, because there are only λn outliers. Thus we have for all $s \geq (1+\kappa)\lambda n/\kappa$,

$$\begin{aligned}&\Pr(\text{the algorithm does not terminate up to step } s)\\ \leq&\Pr\Big(X_s \leq \lambda n - \frac{\kappa s}{1+\kappa}\Big)\\ \leq&\exp\Big(\frac{-(\lambda n - \frac{\kappa s}{1+\kappa})^2}{8s}\Big),\end{aligned}$$

which establishes the theorem. ■

The probability that the algorithm does not terminate up to $\bar{s}_0 = 2(1+\sqrt{\overline{\eta}})\overline{\eta}n/\sqrt{\overline{\eta}}$ is hence bounded by

$$e^{-\left(\frac{n\sqrt{\overline{\eta}}\overline{\eta}}{8(1+\sqrt{\overline{\eta}})}\right)},$$

which goes to zero exponentially in n.

4.3 Deviation Bound of Output PCs

In this section, we show that when HR-PCA terminates, i.e., when all $\mathbf{v}_1, \ldots, \mathbf{v}_d$ pass the Sensitivity Test, the output is close to optimal, in the sense that we bound the distance of $\sum_{i=1}^d \mathbf{v}_i^\top AA^\top \mathbf{v}_i$ to $\sum_{i=1}^d \mathbf{v}_i^{*\top} AA^\top \mathbf{v}_i^*$, where recall the $\{\mathbf{v}_i^*\}$ are the true PCs. We state some technical Lemmas, then prove a finite sample result (Theorem 13) and finally go on to prove the asymptotic result stated above in Theorem 2. The proofs are omitted due to space constraints, and deferred to [16]. To simplify the expressions, we define the following terms:

$$\begin{aligned}&\bar{\theta} \triangleq \frac{1+\overline{\eta}}{2-2\overline{\eta}};\\ &\hat{\psi}_{t,\delta} \triangleq 2\Phi^{-1}(\frac{1+\bar{\theta}}{2} + \hat{h}_{t,\delta}) - 2\Phi^{-1}(\frac{1+\bar{\theta}}{2});\\ &\tilde{\zeta}_{t,\delta} \triangleq \Phi^{-1}\left(1 - \frac{\bar{s}_0}{2t} - \sqrt{\frac{1}{8t}\log\frac{d}{\delta}}\right).\end{aligned}$$

We sometimes drop the subscript of $\tilde{\zeta}_{t,\delta}$ and simply write $\tilde{\zeta}$. It should be understood that $\tilde{\zeta}$ depends on t and δ.

Lemma 10 *For all $\|\mathbf{w}\| \leq 1$, $l_{\mathbf{w}} \leq l(\bar{\theta}, \mathbf{w}^\top \mathbf{z}_1, \ldots, \mathbf{w}^\top \mathbf{z}_t)$.*

Proof: This follows from the definition of $\overline{\theta}$. ■

Lemma 11 *For all unit-norm $\mathbf{w}$, the following holds,*

$$\begin{aligned}&\sqrt{\mathbf{w}^\top AA^\top \mathbf{w}}\\ \geq&\max_{c \geq \sqrt{2\overline{\lambda}_t}} \left\{ \frac{\Phi^{-1}(0.75 - \frac{\overline{\lambda}_{t,\delta}}{2c^2})\overline{O}_{\mathbf{w}} - 2c - (\hat{\phi}_{t,\delta} + \hat{\psi}_{t,\delta})\frac{\overline{\sigma}_1}{2}}{\Phi^{-1}(\frac{1+\bar{\theta}}{2} + \frac{\overline{\lambda}_{t,\delta}}{2c^2})} \right\}.\end{aligned}$$

The proof of this lemma follows from the lemma above, and from Theorem 5 (but see [16] for the full details).

Lemma 12 *Let $a_1, \ldots, a_t$ be i.i.d. realizations of a scalar random variable $\overline{a} \sim \mathcal{N}(0, \omega^2)$. Then, for any fixed $\gamma \in [0,1)$, the following holds with probability at least $1-2\delta$:*

$$\begin{aligned}&\min_{I \subseteq \{1,\ldots,t\}, |I| \geq (1-\gamma)t} \frac{1}{t}\sum_{i \in I} a_i^2\\ &\geq \omega^2 \Big\{ \int_{-\tilde{\tau}}^{\tilde{\tau}} \frac{x^2}{\sqrt{2\pi}} \exp(\frac{-x^2}{2}) dx - \tilde{\tau}^2 \sqrt{\frac{1}{2t}\log\frac{1}{\delta}} \Big\},\end{aligned}$$

where $\tilde{\tau} \triangleq \Phi^{-1}\left(1 - \frac{\gamma}{2} - \sqrt{\frac{1}{8t}\log\frac{1}{\delta}}\right)$.

We sketch the proof, leaving the full details to [16]. Define the function $f(a) \triangleq (\mathbf{1}_{|a| \leq \omega\tau})a^2$. Applying Hoeffding's inequality, one can show that with probability at least $(1-\delta)$,

$$\frac{1}{t}\sum_{i=1}^t f(a_i) \geq \omega^2 \int_{-\tau}^{\tau} \frac{x^2}{\sqrt{2\pi}} \exp(\frac{-x^2}{2}) dx - \omega^2\sigma^2 \sqrt{\frac{1}{2t}\log\frac{1}{\delta}}.$$

Next, defining $g(a) \triangleq \mathbf{1}_{|a| \leq \omega\tau}$, again by application of Hoeffding it is possible to show that with probability at least $(1-\delta)$,

$$\frac{1}{t}\sum_{i=1}^t g(a_i) - \mathbb{E}_{\overline{a} \sim \mathcal{N}(0,\omega^2)} g(\overline{a}) \leq \sqrt{\frac{1}{2t}\log\frac{1}{\delta}}.$$

Since $\mathbb{E}_{\overline{a} \sim \mathcal{N}(0,\omega^2)} g(\overline{a}) = \Phi(\tau) - \Phi(-\tau) = 1 - \gamma - \sqrt{\frac{1}{2t}\log\frac{1}{\delta}}$, we have that with probability $(1-\delta)$,

$$\frac{1}{t}\sum_{i=1}^t g(a_i) \leq 1 - \gamma.$$

Combining the above yields the result.

We now prove a finite-sample result, which we subsequently use to prove the main theorem of the paper.

Theorem 13 *Under (A), (B) and (C), if Algorithm 2.3 terminates at the s^{th} iteration, where $s \leq \bar{s}_0$, and $\int_{-\tilde{\zeta}}^{\tilde{\zeta}} \frac{x^2}{\sqrt{2\pi}} \exp(\frac{-x^2}{2}) dx - \tilde{\zeta}^2\sqrt{\frac{1}{2t}\log\frac{d}{\delta}} > 0$, then with probability at least $1-2\delta$ the following holds*

$$\begin{aligned}&\sum_{q=1}^d \mathbf{v}_q^{*\top}(AA^\top + I_m)\mathbf{v}_q^*\\ \leq&\frac{1+\sqrt{\overline{\eta}}}{\int_{-\tilde{\zeta}}^{\tilde{\zeta}} \frac{x^2}{\sqrt{2\pi}} \exp(\frac{-x^2}{2}) dx - \tilde{\zeta}^2\sqrt{\frac{1}{2t}\log\frac{d}{\delta}}}\\ &\times \min_{c \geq \sqrt{2\overline{\lambda}_{t,\delta}}} \left\{ \left(\frac{\Phi^{-1}(\frac{1+\bar{\theta}}{2} + \frac{\overline{\lambda}_{t,\delta}}{2c^2})}{\Phi^{-1}(0.75 - \frac{\overline{\lambda}_{t,\delta}}{2c^2})} \right)^2 (\sum_{q=1}^d \mathbf{v}_q^\top AA^\top \mathbf{v}_q)\right.\\ &+ \frac{\sqrt{d}\Phi^{-1}(\frac{1+\bar{\theta}}{2} + \frac{\overline{\lambda}_{t,\delta}}{2c^2})(4c + \hat{\phi}_{t,\delta}\overline{\sigma}_1 + \hat{\psi}_{t,\delta}\overline{\sigma}_1)}{[\Phi^{-1}(0.75 - \frac{\overline{\lambda}_{t,\delta}}{2c^2})]^2}\\ &\left.\times \sqrt{\sum_{q=1}^d \mathbf{v}_q^\top AA^\top \mathbf{v}_q + \frac{(4c + \hat{\phi}_{t,\delta}\overline{\sigma}_1 + \hat{\psi}_{t,\delta}\overline{\sigma}_1)^2}{4[\Phi^{-1}(0.75 - \frac{\overline{\lambda}_{t,\delta}}{2c^2})]^2} + \tilde{v}_{t,\delta}} \right\}.\end{aligned} \tag{10}$$

Proof: By definition, if HR-PCA terminates at the s^{th} iteration, then for all $q \in \{1, \ldots, d\}$

$$\frac{(1+\sqrt{\overline{\eta}})(1-\overline{\eta})n}{n-s}\overline{H}_{\mathbf{v}_q} \geq \hat{\sigma}_q^2.$$

Substituting the definitions of $\tilde{v}_{t,\delta}$ and $\overline{H}_{\mathbf{w}}$ into Lemma 11, with some algebra we have

$$\begin{aligned}
&\frac{n-s}{(1+\sqrt{\overline{\eta}})(1-\overline{\eta})n}\hat{\sigma}_q^2 \\
&\leq \min_{c\geq\sqrt{2\overline{\lambda}_{t,\delta}}}\left\{\left(\frac{\Phi^{-1}(\frac{1+\bar{\theta}}{2}+\frac{\overline{\lambda}_{t,\delta}}{2c^2})}{\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t,\delta}}{2c^2})}\right)^2\mathbf{v}_q^\top AA^\top\mathbf{v}_q\right. \\
&\quad+\frac{\Phi^{-1}(\frac{1+\bar{\theta}}{2}+\frac{\overline{\lambda}_{t,\delta}}{2c^2})(4c+\hat{\phi}_{t,\delta}\overline{\sigma}_1+\hat{\psi}_{t,\delta}\overline{\sigma}_1)}{[\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t,\delta}}{2c^2})]^2}\sqrt{\mathbf{v}_q^\top AA^\top\mathbf{v}_q} \\
&\quad\left.+\frac{(4c+\hat{\phi}_{t,\delta}\overline{\sigma}_1+\hat{\psi}_{t,\delta}\overline{\sigma}_1)^2}{4[\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t,\delta}}{2c^2})]^2}+\tilde{v}_{t,\delta}\right\}.
\end{aligned}$$

Summing up for $q=1,\ldots,d$, noticing that the minimal value of the sum is no less than the summation of the minimal value of each term and using the inequality $\sum_{i=1}^d a_i \leq \sqrt{d\sum_{i=1}^d a_i^2}$ for any $a_i\in\mathbb{R}$, we have

$$\begin{aligned}
&\frac{n-s}{(1+\sqrt{\overline{\eta}})(1-\overline{\eta})n}\sum_{q=1}^d\hat{\sigma}_q^2 \\
&\leq \min_{c\geq\sqrt{2\overline{\lambda}_{t,\delta}}}\left\{\left(\frac{\Phi^{-1}(\frac{1+\bar{\theta}}{2}+\frac{\overline{\lambda}_{t,\delta}}{2c^2})}{\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t,\delta}}{2c^2})}\right)^2(\sum_{q=1}^d\mathbf{v}_q^\top AA^\top\mathbf{v}_q)\right. \\
&\quad+\frac{\sqrt{d}\Phi^{-1}(\frac{1+\bar{\theta}}{2}+\frac{\overline{\lambda}_{t,\delta}}{2c^2})(4c+\hat{\phi}_{t,\delta}\overline{\sigma}_1+\hat{\psi}_{t,\delta}\overline{\sigma}_1)}{[\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t,\delta}}{2c^2})]^2} \\
&\quad\left.\times\sqrt{\sum_{q=1}^d\mathbf{v}_q^\top AA^\top\mathbf{v}_q}+\frac{(4c+\hat{\phi}_{t,\delta}\overline{\sigma}_1+\hat{\psi}_{t,\delta}\overline{\sigma}_1)^2}{4[\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t,\delta}}{2c^2})]^2}+\tilde{v}_{t,\delta}\right\}.
\end{aligned}\tag{11}$$

Since $\mathbf{v}_q$ are the PCs of the remaining points at the s^{th} iteration, then for all orthonormal $\{\hat{\mathbf{v}}_1,\ldots,\hat{\mathbf{v}}_d\}$,

$$\sum_{q=1}^d\hat{\mathbf{v}}_q^\top\hat{\Sigma}\hat{\mathbf{v}}_q\leq\sum_{q=1}^d\mathbf{v}_q^\top\hat{\Sigma}\mathbf{v}_q=\sum_{q=1}^d\hat{\sigma}_q^2,$$

where $\hat{\Sigma}$ is the covariance matrix of the remain points.

Recall that $\mathbf{v}_1^*,\ldots,\mathbf{v}_d^*$ are orthonormal and are independent to $\mathcal{Y}$. Hence for a fixed $q\in\{1,\ldots,d\}$, the projection of the authentic samples onto $\mathbf{v}_q^*$ follows a Gaussian distribution with variance $\mathbf{v}_q^{*\top}(AA^\top+I_m)\mathbf{v}_q^*$. Therefore, by Lemma 12 and since $s\leq\bar{s}_0$, we have with probability $1-2\delta/d$

$$\begin{aligned}
\mathbf{v}_q^{*\top}\hat{\Sigma}\mathbf{v}_q^*&\geq\frac{1}{n-s}\min_{I'\subseteq\{1,\ldots,n\},|I'|\geq n-s}\sum_{i\in I'}(\mathbf{v}_q^{*\top}\mathbf{y}_i)^2 \\
&\geq\frac{t}{n-s}\mathbf{v}_q^{*\top}(AA^\top+I_m)\mathbf{v}_q^* \\
&\quad\times\Big\{\int_{-\tilde{\zeta}}^{\tilde{\zeta}}\frac{x^2}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})dx-\tilde{\zeta}^2\sqrt{\frac{1}{2t}\log\frac{d}{\delta}}\Big\}.
\end{aligned}$$

Summing up over q and substituting it into Inequality (11), we establish the theorem. ■

Notice that Conditions (A), (B) and (C) hold simultaneously with probability at least $1-3\delta$, and Algorithm 2.3 terminates at $s<\bar{s}_0$ with probability at least $1-2(1+\sqrt{\overline{\eta}})/\sqrt{\overline{\eta}\eta}n$. Hence, Theorem 13 implies that this finite sample bound holds with probability at least $1-5\delta-2(1+\sqrt{\overline{\eta}})/\sqrt{\overline{\eta}\eta}n$.

Finally, we prove Theorem 2, which provides bounds on the asymptotic performance of the algorithm. To simplify the expressions, let

$$\bar{\theta}^*\triangleq\frac{1+\overline{\eta}^*}{2-2\overline{\eta}^*};$$

$$\rho(j)\triangleq\sum_{q=1}^d\mathbf{v}_q(j)^\top(A(j)A(j)^\top)\mathbf{v}_q(j);$$

$$\rho^*(j)\triangleq\sum_{q=1}^d\mathbf{v}_q^*(j)^\top(A(j)A(j)^\top)\mathbf{v}_q^*(j).$$

Proof of Theorem 2: Taking $c=\sqrt{\overline{\sigma}_1(j)}$ and dividing both sides of Theorem 13 by $\rho^*(j)$, we have:

$$\begin{aligned}
&\frac{\int_{-\tilde{\zeta}(j)}^{\tilde{\zeta}(j)}\frac{x^2}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})dx-\tilde{\zeta}(j)^2\sqrt{\frac{1}{2t(j)}\log\frac{d}{\delta}}}{1+\sqrt{\overline{\eta}}(j)} \\
&\leq\left(\frac{\Phi^{-1}(\frac{1+\bar{\theta}(j)}{2}+\frac{\overline{\lambda}_{t(j),\delta}}{2\overline{\sigma}_1(j)})}{\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t(j),\delta}}{2\overline{\sigma}_1(j)})}\sqrt{\frac{\rho(j)}{\rho^*(j)}}\right. \\
&\quad\left.+\frac{\sqrt{d}(4\sqrt{\overline{\sigma}_1(j)}+\hat{\phi}_{t(j),\delta}\overline{\sigma}_1(j)+\hat{\psi}_{t(j),\delta}\overline{\sigma}_1(j))}{2\sqrt{\rho^*(j)}}\right)^2 \\
&\quad+\frac{(4\sqrt{\overline{\sigma}_1(j)}+\hat{\phi}_{t(j),\delta}\overline{\sigma}_1(j)+\hat{\psi}_{t(j),\delta}\overline{\sigma}_1(j))^2}{4[\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t(j),\delta}}{2\overline{\sigma}_1(j)})]^2\rho^*(j)}+\frac{\tilde{v}_{t(j),\delta}}{\rho^*(j)}.
\end{aligned}\tag{12}$$

Notice that by definition $\sqrt{\rho^*(j)}\geq\overline{\sigma}_1(j)\uparrow+\infty$. Furthermore, we have

$$\hat{\phi}_{t(j),\delta}\downarrow 0;\quad\hat{\psi}_{t(j),\delta}\downarrow 0;\quad\tilde{v}_{t(j),\delta}\downarrow 0;\quad\overline{\lambda}_{t(j),\delta}\downarrow\frac{4}{1-\overline{\eta}}.$$

Thus the right-hand-side of Inequality (12) converges to

$$\left(\frac{\Phi^{-1}(\frac{1+\bar{\theta}(j)}{2}+\frac{\overline{\lambda}_{t(j),\delta}}{2\overline{\sigma}_1(j)})}{\Phi^{-1}(0.75-\frac{\overline{\lambda}_{t(j),\delta}}{2\overline{\sigma}_1(j)})}\sqrt{\frac{\rho(j)}{\rho^*(j)}}\right)^2,\tag{13}$$

since all other terms go to zero. Notice that (13) further converges to

$$\left(\frac{\Phi^{-1}(\frac{1+\bar{\theta}^*}{2})}{\Phi^{-1}(0.75)}\sqrt{\frac{\rho(j)}{\rho^*(j)}}\right)^2,$$

as $\overline{\sigma}_1(j)$ increases and $\overline{\eta}(j)\to\overline{\eta}^*$. On the other hand, noticing that $\tilde{\zeta}(j)\to\tilde{\zeta}^*$ and $\sqrt{\overline{\eta}}(j)\to\sqrt{\overline{\eta}^*}$, we have that the left-hand-side of Inequality (12) converges to

$$\frac{\int_{-\tilde{\zeta}^*}^{\tilde{\zeta}^*}\frac{x^2}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})dx}{1+\sqrt{\overline{\eta}}^*}.$$

The corollary follows by definition of $\sqrt{\overline{\eta}}^*$ and $\bar{\theta}^*$. ■

5 Kernelization

We consider kernelizing HR-PCA in this section: given a feature mapping $\Upsilon(\cdot) : \mathbb{R}^m \to \mathcal{H}$ 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 $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, 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 $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ and evaluating $l(\cdot)$ that is a function of $\{\mathbf{v}_q^\top \mathbf{y}_1, \ldots, \mathbf{v}_q^\top \mathbf{y}_n\}$ for $q = 1, \ldots, d$. The former can be kernelized by applying Kernel PCA introduced by [18], where each of the output PCs admits a representation

$$\mathbf{v}_q = \sum_{i=1}^{n-s} \alpha_i(q) \Upsilon(\hat{\mathbf{y}}_i), \quad q = 1, \ldots, d.$$

Thus, $l(\cdot)$ is easily evaluated by

$$\mathbf{v}_q^\top \Upsilon(\mathbf{y}_j) = \sum_{i=1}^{n-s} \alpha_i(q) k(\hat{\mathbf{y}}_i, \mathbf{y}_j).$$

Therefore, HR-PCA is kernelizable since both steps are easily kernelized.

6 Numerical Illustrations

We report in this section some numerical experimental results. We let $n = m = 100$, i.e., 100 points, each with 100 dimensions. Each element of A is generated according to a uniform distribution; A is then scaled so that its leading eigenvalue equals the given $\overline{\sigma}_1$. All corrupted points are generated on a randomly selected direction. We compare the performance of PCA and HR-PCA for different ratios of corrupted points, magnitudes of corrupted points and $\overline{\sigma}_1$. For each set of parameters, we report the average result of 100 tests.

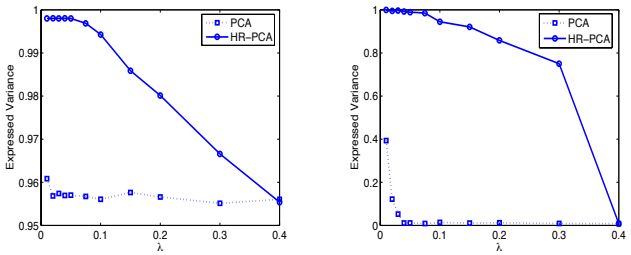


Figure 4: Performance for different ratios of corrupted points

The performance of PCA and HR-PCA of different $\overline{\eta}$ is reported in Figure 4, where both $\overline{\sigma}_1$ and the magnitude of the corrupted points are fixed as 50. As one would expect, HR-PCA outperforms PCA for all $\overline{\eta}$. The performance of HR-PCA breaks only for $\overline{\eta}$ as large as 0.4, i.e., 40% of points are corrupted. We notice that the empirical performance is much better than predicted by the theoretical lower-bound, which is to be expected since the lower bound is derived from a very pessimistic analysis. We also observe that PCA performs much better in the case $d = 5$ than for $d = 1$. This is mainly due to the fact that corrupted points are generated in only one direction. Thus even though PCA wrongly picks the corrupted point direction as a PC, for $d = 5$, the other 4 directions PCA picks are correct, and hence the total Expressed Variance seems to be acceptable. Figure 5 shows a significant performance degradation of PCA in the $d = 5$ case when the corrupted points are generated in 5 random directions.

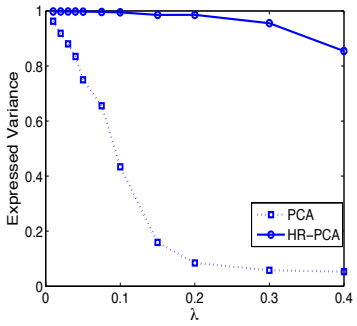


Figure 5: Performance for different ratios of corrupted points: corrupted points generated in multiple directions

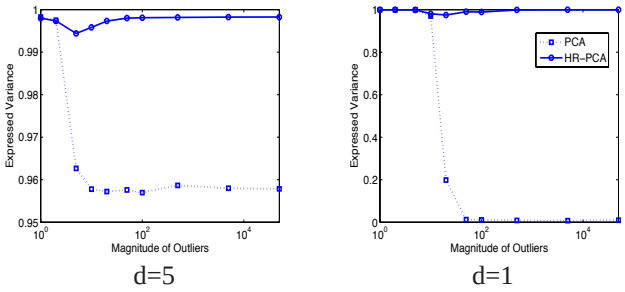


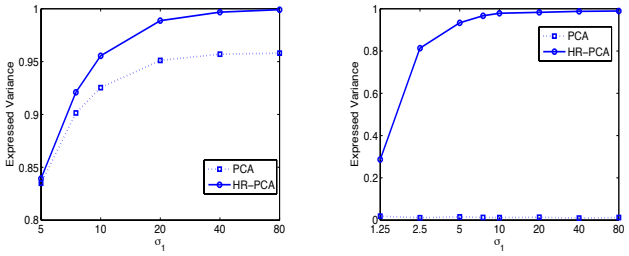
Figure 6: Performance for different magnitudes of corrupted points

Figure 6 shows the performance of HR-PCA and PCA for different magnitudes of corrupted points, with $\overline{\sigma}_1 = 50$ and $\overline{\eta} = 0.05$. One interesting observation is the performance of HR-PCA seems to be quite consistent for different magnitudes of the corrupted points. Indeed, when the corrupted points are large, the performance of HR-PCA is as good as the no-corruption case, mainly because the corrupted points become easier to remove.

Figure 7 shows that the performance of HR-PCA becomes satisfactory for reasonably large $\overline{\sigma}_1$ ($\overline{\sigma}_1 \geq 5$ for 1-d case and $\overline{\sigma}_1 \geq 20$ for 5-d case).

7 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 Princi-



d=5 d=1

Figure 7: Performance for different $\overline{\sigma}_1$

pal Component Analysis algorithm that is tractable, robust to corrupted points, easily kernelizable and asymptotically optimal. The algorithm takes an "actor-critic" form: iteratively finding a set of PCs using standard PCA and subsequently validating the robustness of those PCs using the confidence interval, using a point removal procedure in case of validation failure. 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.

References

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

[2] I. Johnstone. On the distributions of the largest eigenvalue in Principal Components Analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[3] B. Schölkopf and A. 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. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, Berlin: Springer, 1986.

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

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

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

[9] 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.

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

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

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

[13] P. 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.

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

[15] C. Tracy and H. Widom. The distribution of the largest eigenvalue in the Gaussian ensemble. In J. van Diejen and L. Vinet, editors, *Calogero-Moser-Sutherland Models*, pages 461–472. Springer, New York, 2000.

[16] H. Xu, C. Caramanis, and S. Mannor. Principal component analysis with contaminated data: The high dimensional case. Technical report, UT Austin, http://users.ece.utexas.edu/∼cmcaram/, 2009.

[17] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[18] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in kernel Methods – Support Vector Learning*, pages 327–352. MIT Press, Cambridge, MA, 1999.