Adversarial Principal Component Analysis

Daniel L. Pimentel-Alarcón, Aritra Biswas, Claudia R. Solís-Lemus
UNIVERSITY *of* WISCONSIN-MADISON

***Abstract*—This paper studies the following question: where should an adversary place an outlier of a given magnitude in order to maximize the error of the subspace estimated by PCA? We give the exact location of this *worst* possible outlier, and the exact expression of the maximum possible error. Equivalently, we determine the information-theoretic bounds on how much an outlier can *tilt* a subspace in its direction. This in turn provides universal (worst-case) error bounds for PCA under arbitrary noisy settings. Our results also have several implications on adaptive PCA, online PCA, and rank-one updates. We illustrate our results with a subspace tracking experiment.**

I. INTRODUCTION

Subspace models lie at the heart of data analysis. From biologists studying genes to astronomers studying galaxies, scientists often want to estimate the low-dimensional subspace that best explains their data. Principal Component Analysis (PCA) is arguably the most widely used technique for this purpose. However, it is well-known that PCA is sensitive to outliers. In fact, depending on its direction and magnitude, a single outlier can cause an arbitrarily inaccurate estimate.

It is easy to see that for a given direction, if the magnitude of an outlier increases, then the error of the subspace estimated by PCA will either remain constant or increase (but not decrease). In other words, larger outliers can only make things worse. On the other hand, for a fixed magnitude, some directions are worse than others. To build some intuition, consider data lying in a low-dimensional subspace X, as in Figure 1. Where would an adversary place an outlier $\mathbf{y}$ of a given magnitude such that the subspace Z estimated by PCA were as far as possible from X?

At first glance, this might look deceivingly simple. For example, if X and Z are 1-dimensional, as in the left of Figure 1, one could think that the angle φ between X and Z can only grow as θ asymptotically approaches $\pi/2$. However, this is not the case. In fact, φ will initially increase as θ grows from 0, but at some point φ will start decreasing as θ approaches $\pi/2$ (see Figure 2). In the 1-dimensional case, the question reduces to finding the angle θ (which determines the direction of $\mathbf{y}$) that maximizes φ. In general, if X and Z have dimensions larger than 1, as in the right of Figure 1, the goal is to determine the direction of $\mathbf{y}$ that maximizes the distance between X and Z. That is precisely what we do in this paper.

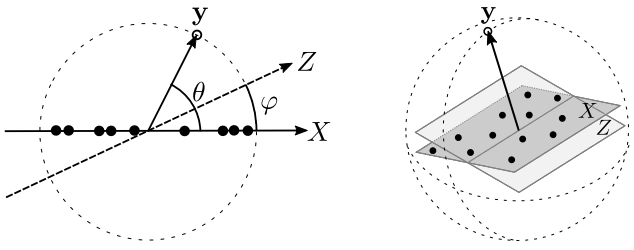

Fig. 1. Each black point represents a datum lying in a low-dimensional subspace X. $\mathbf{y}$ is an outlier and Z is the subspace that PCA estimates from all the data, including $\mathbf{y}$. Fixing $\|\mathbf{y}\|$, this paper determines where an adversary would place $\mathbf{y}$ such that Z were as far as possible from X.

In this paper we determine the exact location of the worst possible outlier. More precisely, for a given magnitude, we determine the direction of the outlier $\mathbf{y}^\star$ that maximizes the error of the subspace estimated by PCA. Our main result shows that $\mathbf{y}^\star$ must have *the right* components (which we determine exactly) in two directions: the direction of smallest variance in X, and a direction orthogonal to X. In addition, we give the exact expression of the worst possible error.

Equivalently, our results determine the information-theoretic bounds on how much an outlier can *tilt* a subspace in its direction. This in turn provides universal (worst-case) error bounds for PCA under arbitrary noisy settings. Since we also determine the position of the outlier that achieves this maximum tilting, our results also have several implications on adaptive PCA, online PCA, and rank-one updates. We illustrate our results with a subspace tracking experiment.

Organization of the paper

In Section II we formally state the problem and our main results. In Section III we discuss related work. Section IV illustrates our results with an experiment that is tightly related to online PCA, subspace tracking and rank-one updates. We give the proof of our main theorem in Section V.

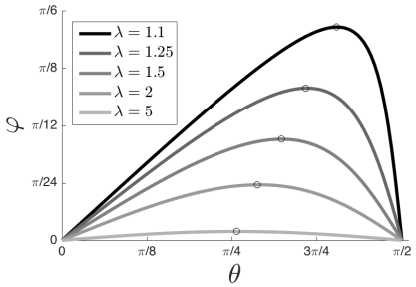

Fig. 2. If X and Z are 1-dimensional, as in the left of Figure 1, the angle φ between X and Z first increases as θ grows from 0, and then decreases as θ approaches $\pi/2$. Here $\|\mathbf{y}\| = 1$, and λ is the energy in X, which summarizes the number of inliers and their magnitudes. In this 1-dimensional case, the direction of $\mathbf{y}$ that maximizes the distance between X and Z is given by the angle θ that maximizes φ.

II. Model and Main Results

Let X be an r-dimensional subspace in $\mathbb{R}^d$, $r < d$. Let $\mathbf{X}$ be a data matrix with columns lying in X. Let $\mathbf{y}$ be an outlier, that is, a column vector in $\mathbb{R}^d \backslash X$. Let Z be r-dimensional subspace in $\mathbb{R}^d$ spanned by the r leading left singular vectors of $\mathbf{Z} := [\mathbf{X} \;\; \mathbf{y}]$.

Depending on $\mathbf{X}$ and $\mathbf{y}$, Z can be close or far from X. Given $\mathbf{X}$ and the magnitude of $\mathbf{y}$, we want to determine the direction of $\mathbf{y}$ that maximizes the distance between X and Z. To measure such distance we will use the largest principal angle between X and Z, defined as follows [1].

Definition 1 (Principal angles)**.** *Let X, Z be two r-dimensional subspaces in $\mathbb{R}^d$. The principal angles $\varphi_1, \varphi_2, \ldots, \varphi_r \in [0, \pi/2]$ between X and Z are defined recursively by*

$$\begin{aligned} \cos\varphi_i \;:=\; \max_{\mathbf{x}\in X, \mathbf{z}\in Z} \mathbf{x}^\top \mathbf{z} \;=:\; \mathbf{x}_i^\top \mathbf{z}_i \quad & \textit{s.t.} \\ \|\mathbf{x}\| = \|\mathbf{z}\| = 1 \quad & \textit{and} \\ \mathbf{x}^\top \mathbf{x}_j = \mathbf{z}^\top \mathbf{z}_j = 0 \;\; \forall \;\; & j = 1, 2, \ldots, i-1. \end{aligned}$$

Intuitively, the principal angles are the collection of smallest angles in orthogonal directions between vectors in X and Z. Notice that they satisfy $0 \le \varphi_1 \le \varphi_2 \le \cdots \le \varphi_r \le \pi/2$. To measure the distance between X and Z we will use their *largest* principal angle φ_r, which we will denote simply as φ. Given orthonormal bases $\mathbf{U}_X$ and $\mathbf{U}_Z$, $\cos(\varphi)$ is given by the smallest singular value of $\mathbf{U}_X^\top \mathbf{U}_Z$. See Chapter 12, Section 12.4.3 in [1] to know more about principal angles and their computation.

First notice that the singular vectors of $\mathbf{Z}$ only depend on $\mathbf{y}$ and the left singular vectors and values of $\mathbf{X}$. So we can assume without loss of generality (w.l.o.g.) that $\mathbf{X}$ only has r columns, whose directions and magnitudes are given by its left singular vectors and values. Furthermore, we can rewrite $\mathbf{X}$ and $\mathbf{y}$ with respect to an orthonormal basis whose first r vectors span X, and whose first $r+1$ vectors span Z. Hence we can **assume w.l.o.g.** that $\mathbf{X}$ and $\mathbf{y}$ have the following forms:

$$\mathbf{X} \;=\; \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \\ & \mathbf{0} & & \end{bmatrix}, \qquad \mathbf{y} \;=\; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \\ y_{r+1} \\ \mathbf{0} \end{bmatrix}, \qquad (1)$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r$, and blank spaces represent zeros. See Section IV for an example of how to handle arbitrary data and transform it to the form in (1).

Next observe that the leading left singular vectors of $\mathbf{Z}$ and $\mathbf{Z}/\alpha$ are the same for every $\alpha > 0$, including $\alpha = \|\mathbf{y}\|$. So we can **assume w.l.o.g.** that $\|\mathbf{y}\| = 1$ with the understanding that otherwise we can simply rescale. This way the information of the magnitude of $\mathbf{y}$ is encoded in λ_r. Large values of λ_r correspond to outliers of small magnitude compared to the energy of the inliers, and vice versa. Intuitively, we can think of λ_r as the signal-to-noise ratio.

Finally, let $\mathbf{e}_i$ denote the i^{th} canonical vector in $\mathbb{R}^d$. Notice that if $\lambda_r < 1$, i.e., if the energy of the inliers is lower than the magnitude of the outlier, then $\mathbf{y} = \mathbf{e}_{r+1}$ trivially maximizes φ. To see this, observe that X is spanned by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r$. So if $\mathbf{y} = \mathbf{e}_{r+1}$, then the $r+1$ left singular vectors of $\mathbf{Z}$ will be $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r$ and $\mathbf{e}_{r+1}$. Similarly, the singular values of $\mathbf{Z}$ will be $\lambda_1, \lambda_2, \ldots, \lambda_r$ and 1. If $\lambda_r < 1$, then $\mathbf{e}_{r+1}$ will be among the *leading* singular vectors, and the $\mathbf{e}_r$ will become the $(r+1)^{\text{th}}$ singular vector. Since X and Z share $r-1$ spanning vectors, namely $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{r-1}$, their largest principal angle φ is the angle between $\mathbf{e}_r$ and $\mathbf{e}_{r+1}$, i.e, $\pi/2$, which is maximal. Hence, we will **assume w.l.o.g.** that $\lambda_r \ge 1$ with the understanding that if this is not the case, then $\mathbf{y} = \mathbf{e}_{r+1}$ trivially maximizes φ.

With this, we are ready to present our main theorem. It determines the exact location of the outlier $\mathbf{y}$ that maximizes the angle φ between X and Z, and gives the exact expression of the maximum possible φ, both as a function of the ratio between the energy of the inliers and the magnitude of the outlier, given by λ_r. The proof is given in Section V.

Theorem 1. *Under the setup above, the unitary vector $\mathbf{y}^\star \in \mathbb{R}^d$ that maximizes the largest principal angle between X and Z has components:*

$$y_i^\star \;=\; \begin{cases} \cos\theta^\star & i = r, \\ \sin\theta^\star & i = r+1, \\ 0 & \textit{otherwise}, \end{cases}$$

where

$$\theta^\star \;:=\; \frac{1}{2}\arccos\left(-\frac{1}{\lambda_r^2}\right) \in [\pi/4, \pi/2], \qquad (2)$$

whence the largest principal angle between X and Z is

$$\varphi^\star = \arccos\left(\frac{\sin^2\theta^\star - \sigma_\star^2}{\sqrt{(\sin^2\theta^\star - \sigma_\star^2)^2 + (\sin\theta^\star\cos\theta^\star)^2}}\right),$$

with

$$\sigma_\star^2 \;=\; \frac{(\lambda_r^2+1) + \sqrt{(\lambda_r^2+1)^2 - 4\lambda_r^2\sin^2\theta^\star}}{2}.$$

In words, Theorem 1 states that the outlier $\mathbf{y}^\star$ that maximizes φ has a component $\cos\theta^\star$ in the direction of the smallest singular vector of $\mathbf{X}$, and a component $\sin\theta^\star$ in an orthogonal direction of X. The intuition is that it is easier to *tilt* smaller vectors. One interpretation is that $\mathbf{y}^\star$ is tilting X as much as possible ($\varphi^\star$) using a *lever* of size $\cos\theta^\star$ to pull the smallest singular vector of $\mathbf{X}$ (of size λ_r) in an orthogonal direction ($\mathbf{e}_{r+1}$) with strength $\sin\theta^\star$. See Figure 3 to build some intuition.

Remark 1. *Due to the symmetry of the problem, there are in fact four vectors $\mathbf{y}^\star$ that maximize φ, given by the four sign combinations in the two nonzero values in $\mathbf{y}^\star$.*

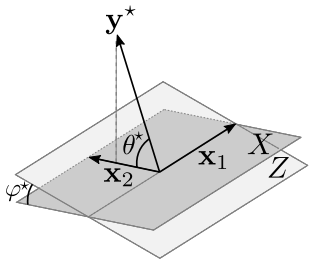

Fig. 3. Theorem 1 shows that the outlier $\mathbf{y}^\star$ that maximizes φ has a component $\cos\theta^\star$ in the direction of the smallest singular vector of $\mathbf{X}$, and a component $\sin\theta^\star$ in an orthogonal direction of X, with $\theta^\star$ as in (2).

III. Related Work

It is well known that a single outlier of large magnitude can severely compromise the performance of PCA. Hence there is a broad interest in **Robust PCA**. Common approaches include M-estimators [2], random sampling [3, 4], influence function techniques [5], alternating minimization [6] and convex relaxations [7–14]. In principle, these approaches aim to distinguish between inliers and outliers, and only estimate the subspace corresponding to the inliers. However, to the best of our knowledge, there is no analysis of PCA's performance as a function of the location of the outliers.

Our work determines the exact location of the outlier that would maximally *tilt* a subspace. In this sense, our work is tightly related to **online PCA**, **subspace tracking** and **rank-one updates**, where one aims to efficiently estimate or track subspaces as new data points arrive, without computing the full singular value decomposition. Some representative approaches include exact rank-one updates [15, 16], robust incremental algorithms [17], standard and exponentiated gradient descent [18], approximation algorithms [19, 20] and recursive algorithms [21], among others [22–24]. All these papers study how to update a subspace given a new sample. In contrast, we quantify the subspace update as a function of the location of a new sample (outlier). We take this one step further, and determine the location of the sample that would make the subspace update as large as possible. In the next section we show an experiment that illustrates our results in the framework of online PCA and rank-one updates.

IV. Experiments

In this section we illustrate our results with an experiment that is tightly related to online PCA, subspace tracking and rank-one updates. This experiment compares the error of PCA when data is contaminated with two types of outliers:

(i) Outliers located adversarially to maximize error and
(ii) Isotropic outliers in random locations.

These experiments also illustrate how to transform arbitrary data into the setup in Section II.

Suppose we observe a data matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{d\times n}$ given by

$$\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

where both $\mathbf{X} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times n}$ have $\mathcal{N}(0,1)$ i.i.d. entries, and $\mathbf{E} \in \mathbb{R}^{d\times n}$ has $\mathcal{N}(0,\epsilon)$ i.i.d. entries. This way $\mathbf{X}$ is a basis of the subspace X that we aim to identify and $\epsilon \in \mathbb{R}$ represents the noise level. Notice that $\widetilde{\mathbf{X}}$ has more than r columns and is not in the form of (1).

As in online PCA and subspace tracking, we will start with an initial estimate of X, and then update it iteratively as new data comes in. More precisely, let $\mathbf{U}_0 \in \mathbb{R}^{d\times r}$ and $\mathbf{\Lambda}_0 \in \mathbb{R}^{r\times r}$ be the matrices with the r leading left singular vectors and values of $\widetilde{\mathbf{X}}$ and let $\mathbf{X}_0 \in \mathbb{R}^{d\times r}$ be a scaled copy of $\mathbf{U}_0\mathbf{\Lambda}_0$ such that its smallest singular value is equal to λ_r. This is done so that each new datum is given the same importance. There are many alternatives to this, for example, each new datum can be given diminishing, adaptive or even adversarial importance. However, the simplest strategy of giving each datum the same importance will do for our illustration purposes. $X_0 := \operatorname{span}[\mathbf{X}_0]$ will be our initial estimate of X.

Next, at each time $t > 0$ we will generate a new datum $\mathbf{y}_t$, and update our subspace estimate. Given $\mathbf{y}_t$, let $\mathbf{U}_t \in \mathbb{R}^{d\times r}$ and $\mathbf{\Lambda}_t \in \mathbb{R}^{r\times r}$ be the matrices with the r leading left singular vectors and values of $[\mathbf{X}_{t-1}\ \mathbf{y}_t] \in \mathbb{R}^{d\times(r+1)}$, and let $\mathbf{X}_t \in \mathbb{R}^{d\times r}$ be a scaled copy of $\mathbf{U}_t\mathbf{\Lambda}_t$ such that its smallest singular value is equal to λ_r. Again, this is done so that each new datum is given the same importance. $X_t := \operatorname{span}[\mathbf{X}_t]$ will be our estimate of X at time t. Notice that $\mathbf{X}_t$ can be computed explicitly or using a rank-one update of $\mathbf{X}_{t-1}$ [15–24].

Now, each $\mathbf{y}_t$ will be a unit-norm outlier with probability p and a unit-norm inlier with probability $1-p$. Inliers are generated as $\mathbf{y}_t = \mathbf{X}\mathbf{b}_t + \boldsymbol{\epsilon}_t$, where $\mathbf{b}_t \in \mathbb{R}^r \sim \mathcal{N}(0,\mathbf{I})$ and $\boldsymbol{\epsilon}_t \in \mathbb{R}^d \sim \mathcal{N}(0,\epsilon\mathbf{I})$. Here $\mathbf{I}$ denotes the identity matrix. For the setup in (ii), we generate outliers $\mathbf{y}_t$ with $\mathcal{N}(0,1)$ entries. For the setup in (i), we generate outliers $\mathbf{y}_t$ using Theorem 1 in order to *tilt* X_{t-1} as much as possible. Notice, however, that $\mathbf{X}_{t-1}$ is not in the diagonal form of (1). Nonetheless, we can *rotate* $\mathbf{X}_{t-1}$ into that form, and then *rotate* it back. To this end, let $\bar{\mathbf{U}}_{t-1}$ be an orthonormal basis of $\mathbb{R}^d$ whose first columns are $\mathbf{U}_{t-1}$. Left-multiplying by $\bar{\mathbf{U}}_{t-1}^\mathsf{T}$ we can change our coordinate system. The representation of $\mathbf{X}_{t-1}$ with respect to this new coordinate system is given by $\mathbf{X}'_{t-1} := \bar{\mathbf{U}}_{t-1}^\mathsf{T}\mathbf{X}_{t-1}$. Now $\mathbf{X}'_{t-1} = \mathbf{\Lambda}_{t-1}$ has the diagonal form in (1). Hence the outlier that *tilts* $\operatorname{span}[\mathbf{X}'_{t-1}]$ the most is given by $\mathbf{y}^\star$ as in Theorem 1 with λ_r being the smallest singular value of $\mathbf{X}'_{t-1}$ (which is also the smallest singular value of $\mathbf{X}_{t-1}$). To recover the outlier $\mathbf{y}_t$ that tilts X_{t-1} the most, all that remains is to write $\mathbf{y}^\star$ in the initial coordinate system, which can be done by left-multiplying by $\bar{\mathbf{U}}_{t-1}$. To summarize, the outlier $\mathbf{y}_t$ that *tilts* X_{t-1} the most is given by $\mathbf{y}_t = \bar{\mathbf{U}}_{t-1}\mathbf{y}^\star$, with $\mathbf{y}^\star$ as in Theorem 1, where λ_r is the smallest singular value of $\mathbf{X}_{t-1}$. The results with $d = 5, r = 4, n = 100$ and $\epsilon = 10^{-3}$ are summarized in Figure 4.

V. Proof

In this section we give the proof of Theorem 1. The proof is divided in two main parts. First we show that the $\mathbf{y}^\star \in \mathbb{R}^d$

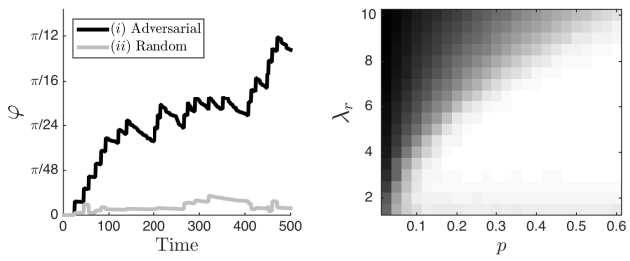

Fig. 4. **Left:** One trial of the evolution of φ over time with $\lambda_r = 5$ and a fraction of $p = 0.05$ outliers. The steady decreases in φ correspond to new inliers coming in. The dramatic increases in φ correspond to the arrival of adversarial outliers. Notice that random outliers may slightly increase or decrease φ. **Right:** Angle φ after 500 updates (average over 100 trials) as a function of λ_r, which essentially determines the influence of each new datum, and the fraction of adversarial outliers p. Black represents perfect recovery ($\varphi = 0$) and white represents maximal error ($\varphi = \pi/2$). This shows that even a small fraction of adversarial outliers of small magnitude can seriously perturb PCA's performance, even under a low-noise regime.

that maximizes φ must satisfy

$$\begin{cases} y_i^\star \neq 0 & i = r, r+1, \\ y_i^\star = 0 & \text{otherwise}, \end{cases} \tag{3}$$

and then we show the specific values of $y_r^\star$ and $y_{r+1}^\star$ that maximize φ.

Recall that we are assuming w.l.o.g. that our outlier has the form in (1). Hence $y_i^\star = 0$ for every $i > r+1$. Also, since $\mathbf{y}^\star \notin X$ by definition, $y_{r+1}^\star \neq 0$. It remains to show that $y_r^\star \neq 0$ and $y_i^\star = 0$ for every $i < r$.

To see this, notice that because of the zero blocks in (1), X and Z are r-dimensional subspaces that only have energy in the first $r+1$ coordinates. This implies that their intersection is an $(r-1)$-dimensional subspace. Let $\mathbf{W} \in \mathbb{R}^{d\times(r-1)}$ be an orthogonal basis of $W := X \cap Z$. Let $\mathbf{x} \in X$ and $\mathbf{z} \in Z$ be vectors orthogonal to W, such that $[\mathbf{W}\ \ \mathbf{x}]$ and $[\mathbf{W}\ \ \mathbf{z}]$ are orthogonal bases of X and Z. By the definition of principal angles, φ is the angle between $\mathbf{x}$ and $\mathbf{z}$ (see Figure 5 to build intuition).

It is easy to see that $\mathbf{z}$ must have a nonzero component in the direction of $\mathbf{e}_{r+1}$. Otherwise $\mathbf{z} \in X$, implying $[\mathbf{W}\ \ \mathbf{z}] \subset X$, implying $Z = X$, implying $\varphi = 0$. Similarly, $\mathbf{z}$ must have a nonzero component in the direction of $\mathbf{x}$. Otherwise, $\mathbf{z}$ is orthogonal to $\mathbf{x}$, implying $\mathbf{z}$ is orthogonal to $[\mathbf{W}\ \ \mathbf{x}] = X$, implying $\mathbf{z} = \mathbf{e}_{r+1}$, implying $\mathbf{y} = \mathbf{e}_{r+1}$ and $\|\mathbf{y}\| = 1 > \lambda_r$, contradicting our assumption that $\lambda_r > 1$.

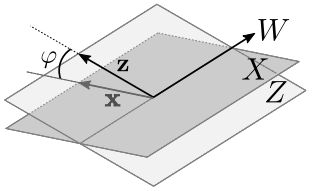

Fig. 5. $W := X \cap Z$ is an $(r-1)$-dimensional subspace in $\mathbb{R}^d$. Here $r = 2$ and $d = 3$. $\mathbf{x} \in X$ and $\mathbf{z} \in Z$ are vectors orthogonal to W. The largest principal angle φ between X and Z is the angle between $\mathbf{x}$ and $\mathbf{z}$.

Clearly, $\mathbf{z} \in \operatorname{span}[\mathbf{Z}] = \operatorname{span}[\mathbf{W}\ \ \mathbf{x}\ \ \mathbf{e}_{r+1}]$, and since $\mathbf{z}$ is orthogonal to W, it follows that $\mathbf{z}$ lies in the plane spanned by $\mathbf{x}$ and $\mathbf{e}_{r+1}$. It is easy to see that $\mathbf{z}$ will be closer to $\mathbf{x}$ (whence φ will be smaller) if the energy of $\mathbf{X}$ in the direction of $\mathbf{x}$ is large, and vice versa. It follows that φ is maximized when $\mathbf{x}$ is in the direction of lowest energy in $\mathbf{X}$, namely $\mathbf{e}_r$, because we are assuming $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$. In conclusion, φ is maximized when $\mathbf{z}$ lies in $\operatorname{span}[\mathbf{e}_r\ \ \mathbf{e}_{r+1}]$.

Similarly, fix the direction of $\mathbf{y}$ and let $\hat{\mathbf{y}}$ be the projection of $\mathbf{y}$ onto $\operatorname{span}[\mathbf{e}_r\ \ \mathbf{e}_{r+1}]$. It is easy to see that $\mathbf{z}$ will be closer to $\mathbf{y}$ (whence φ will be larger) as $\|\hat{\mathbf{y}}\|$ grows. It follows that φ is maximized when $\|\hat{\mathbf{y}}\| = \|\mathbf{y}\|$, i.e., when the only nonzero components in $\mathbf{y}$ are $\mathbf{y}_r$ and $\mathbf{y}_{r+1}$. We thus conclude that $\mathbf{y}^\star$ must satisfy (3).

It remains to determine the specific values of $y_r^\star$ and $y_{r+1}^\star$ that maximize φ, which we will do next. Since $\mathbf{y}^\star$ satisfies (3), φ is maximized when $\mathbf{Z}$ has the following block-diagonal form, where blank spaces represent zeros:

$$\mathbf{Z} = \left[\begin{array}{cccc|cc} \lambda_1 & & & & & \\ & \lambda_2 & & & & \\ & & \ddots & & & \\ & & & \lambda_{r-1} & & \\ \hline & & & & \lambda_r & y_r \\ & & & & & y_{r+1} \\ & & & & \multicolumn{2}{c}{\mathbf{0}} \end{array}\right].$$

Let $\mathbf{x}_r$ be the r^{th} column in $\mathbf{Z}$ and let $\mathbf{z}$ be the leading left singular vector of $\mathbf{Z}_r := [\mathbf{x}_r\ \ \mathbf{y}]$. Then the singular values of $\mathbf{Z}$ are $\lambda_1, \lambda_2, \ldots, \lambda_{r-1}$ and the two singular values of $\mathbf{Z}_r$, one larger and one smaller than λ_r. Since $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$, it follows that Z will be the r-dimensional subspace spanned by the first $r-1$ canonical vectors and $\mathbf{z}$.

Furthermore, since X and Z share $r-1$ spanning vectors, namely $\mathbf{W} = [\mathbf{e}_1\ \mathbf{e}_2\ \cdots\ \mathbf{e}_{r-1}]$, it follows that the largest principal angle φ between X and Z is the angle between $\mathbf{x}_r$ and $\mathbf{z}$. Our goal is to determine the location of the outlier $\mathbf{y}^\star$ that maximizes this angle. Since $\|\mathbf{y}^\star\| = 1$ by assumption and $\mathbf{y}^\star$ satisfies (3), we can write it as

$$y_i^\star = \begin{cases} \cos\theta & i = r, \\ \sin\theta & i = r+1, \\ 0 & \text{otherwise}, \end{cases}$$

for some θ. By symmetry, we can assume without loss of generality that $\theta \in [0, \pi/2]$. We thus want to determine the angle $\theta^\star \in [0, \pi/2]$ that maximizes the angle φ between $\mathbf{x}_r$ and $\mathbf{z}$. To this end we will use standard techniques: determine φ in closed form, take its derivative with respect to θ, set it to zero, and solve for θ.

To do this, let σ be the leading left singular value of $\mathbf{Z}_r$. From textbook linear algebra, we know σ^2 is the leading eigenvalue of $\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T}$, which is the largest solution to $|\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T} - \sigma^2\mathbf{I}| = 0$. Writing

$$\begin{aligned} |\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T} - \sigma^2\mathbf{I}| &= \begin{vmatrix} -\sigma^2\mathbf{I} & & \\ & \lambda_r^2 + \cos^2\theta - \sigma^2 & \sin\theta\cos\theta \\ & \sin\theta\cos\theta & \sin^2\theta - \sigma^2 \\ & & & -\sigma^2\mathbf{I} \end{vmatrix} \\ &= \left(\sigma^4 - \sigma^2(\lambda_r^2 + 1) + \lambda_r^2\sin^2\theta\right)\sigma^{2(d-2)}, \end{aligned}$$

we can see that the leading eigenvalue of $\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T}$ is

$$\sigma^2 \;=\; \frac{(\lambda_r^2+1)+\sqrt{(\lambda_r^2+1)^2-4\lambda_r^2\sin^2\theta}}{2}. \qquad (4)$$

Similarly, $\mathbf{z}$ is the leading eigenvector of $\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T}$, i.e., the solution to $(\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T}-\sigma^2\mathbf{I})\mathbf{z} = \mathbf{0}$, with σ^2 as in (4). By construction, $(\mathbf{Z}_r\mathbf{Z}_r^\mathsf{T}-\sigma^2\mathbf{I})$ is rank-deficient. In particular, from its block diagonal structure it is easy to see that only rows r and $r+1$ can be linearly dependent. This implies that

$$\begin{bmatrix}\mathbf{0} & \lambda_r^2+\cos^2\theta-\sigma^2 & \sin\theta\cos\theta & \mathbf{0}\end{bmatrix}\mathbf{z} \;=\; 0$$

if and only if

$$\begin{bmatrix}\mathbf{0} & \sin\theta\cos\theta & \sin^2\theta-\sigma^2 & \mathbf{0}\end{bmatrix}\mathbf{z} \;=\; 0.$$

We can use either equation to solve for $\mathbf{z}$. Rewriting the later as

$$(\sin\theta\cos\theta)z_r \;+\; (\sin^2\theta-\sigma^2)z_{r+1} \;=\; 0,$$

it is easy to see that one solution is given by

$$z_i \;=\; \begin{cases}\sin^2\theta-\sigma^2 & i=r,\\ -\sin\theta\cos\theta & i=r+1,\\ 0 & \text{otherwise.}\end{cases}$$

Recall that we are interested in the angle φ between $\mathbf{x}_r$ and $\mathbf{z}$. This angle is given by

$$\begin{aligned}\varphi &= \arccos\left(\frac{\langle\mathbf{x}_r,\mathbf{z}\rangle}{\|\mathbf{x}_r\|\|\mathbf{z}\|}\right)\\ &= \arccos\left(\frac{\sin^2\theta-\sigma^2}{\sqrt{(\sin^2\theta-\sigma^2)^2+(\sin\theta\cos\theta)^2}}\right), \qquad (5)\end{aligned}$$

with σ^2 as in (4). Figure 2 depicts φ as a function of θ and λ_r. We want to find the angle θ that maximizes φ, so we will take the derivative of φ with respect to θ, set it to zero, and solve for θ. Notice that σ^2 in (4) also depends on θ. The desired derivative is given by

$$\varphi' \;=\; \frac{ABCD}{EF},$$

where

$$\begin{aligned}A &= \csc(2\theta),\\ B &= |\sin(2\theta)|,\\ C &= \lambda_r^2+\cos(2\theta)+\sqrt{1+\lambda_r^4+2\lambda_r^2\cos(2\theta)},\\ D &= 1+\lambda_r^2\cos(2\theta),\\ E &= \sqrt{1+\lambda_r^4+2\lambda_r^2\cos(2\theta)},\\ F &= 1+\lambda_r^4+\lambda_r^2\sqrt{1+\lambda_r^4+2\lambda_r^2\cos(2\theta)}\\ &\quad+\cos(2\theta)(2\lambda_r^2+\sqrt{1+\lambda_r^4+2\lambda_r^2\cos(2\theta)}).\end{aligned}$$

First that $A = \csc(2\theta) \neq 0$ for every θ. Recall that we are assuming w.l.o.g. that $\theta \in [0,\pi/2]$. This implies that $B = |\sin(2\theta)|$ and $C = \lambda_r^2+\cos(2\theta)+\sqrt{1+\lambda_r^4+2\lambda_r^2\cos(2\theta)}$ are zero if and only if $\theta \in \{0,\pi/2\}$. If $\theta = 0$, then $\mathbf{y}$ and $\mathbf{z}$ are parallel to $\mathbf{x}_r$, whence $\varphi = 0$. See Figure 1 to build some intuition. On the other hand, if $\theta = \pi/2$, then $\mathbf{y}$ is the $(r+1)^{\text{th}}$ canonical vector. Since $\lambda_r > 1 = \|\mathbf{y}\|$, $\mathbf{y}$ will be the $(r+1)^{\text{th}}$ left singular vector, which implies $\varphi = 0$ as well. It follows that $\theta \in \{0,\pi/2\}$ are minimizers of φ.

Finally, $D = 1+\lambda_r^2\cos(2\theta) = 0$ if $\theta = \frac{1}{2}\arccos\left(-1/\lambda_r^2\right)$. It follows that this is the angle that maximizes φ, as claimed. Notice that since $\lambda_r > 1$ by assumption, then $-1 < -1/\lambda_r^2 < 0$, which implies $\theta^\star \in [\pi/4,\pi/2]$. □

REFERENCES

[1] G. Golub and C. Van Loan, *Matrix Computations*, 3rd edition, The Johns Hopkins University Press, 1996.

[2] R. Maronna, *Robust M-estimators of multivariate location and scatter*, The Annals of Statistics, 1976.

[3] M. Fischler and R. Bolles, *Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography*, Communications of the ACM, 1981.

[4] D. Pimentel-Alarcón and R. Nowak, *Random Consensus Robust PCA*, submitted to International Conference on Artificial Intelligence and Statistics, 2017.

[5] F. De La Torre and M. Black, *A framework for robust subspace learning*, International Journal of Computer Vision, 2003.

[6] Q. Ke and T. Kanade, *Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming*, IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[7] E. Candès and J. Romberg, *Sparsity and incoherence in compressive sampling*, Inverse Problems, 2007.

[8] J. Wright, A. Ganesh, S. Rao, Y. Peng and Y. Ma, *Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization*, Advances in Neural Information Processing Systems, 2009.

[9] J. Cai, E. Candès and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 2010.

[10] H. Xu, C. Caramanis and S. Sanghavi, *Robust PCA via outlier pursuit*, Advances in Neural Information Processing Systems, 2010.

[11] E. Candès, X. Li , Y. Ma and J. Wright, *Robust principal component analysis?*, Journal of the ACM, 2011.

[12] V. Chandrasekaran, S. Sanghavi, P. Parrilo and A. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM Journal on Optimization, 2011.

[13] L. Mackey, A. Talwalkar and M. Jordan, *Divide-and-conquer matrix factorization*, Advances in Neural Information Processing Systems, 2011.

[14] M. Rahmani and G. Atia, *A subspace learning approach for high dimensional matrix decomposition with efficient column/row sampling*, International Conference on Machine Learning, 2016.

[15] M. Gu and S. Eisenstat, *A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem*, SIAM journal on Matrix Analysis and Applications, 1994.

[16] M. Brand, *Fast low-rank modifications of the thin singular value decomposition*, Linear Algebra and its Applications, 2006.

[17] J. Feng, H. Xu and S. Yan, *Online robust PCA via stochastic optimization*, Advances in Neural Information Processing Systems, 2013.

[18] J. Nie, W. Kotlowski and M. Warmuth, *Online PCA with optimal regrets*, International Conference on Algorithmic Learning Theory, 2013.

[19] C. Boutsidis, *Online principal components analysis*, ACM-SIAM Symposium on Discrete Algorithms, 2015.

[20] Z. Karnin and E. Liberty, *Online PCA with spectral bounds*, Annual Conference on Computational Learning Theory, 2015.

[21] W. Li, H. Yue, S. Valle-Cervantes and S. Qin, *Recursive PCA for adaptive process monitoring*, Journal of process control, 2000.

[22] G. Golub, *Some modified matrix eigenvalue problems*, SIAM, 1973.

[23] J. Bunch, C. Nielsen and D. Sorensen *Rank-one modification of the symmetric eigenproblem*, Numerische Mathematik, 1978.

[24] A. Hegde, J. Principe, D. Erdogmus, U. Ozertem, Y. Rao and H. Peddaneni, *Perturbation-based eigenvector updates for on-line principal components analysis and canonical correlation analysis*, Journal of VLSI signal processing systems for signal, image and video technology, 2006.