Electronic Journal of Statistics
Vol. 9 (2015) 204–218
ISSN: 1935-7524
DOI: 10.1214/15-EJS992

High dimension low sample size asymptotics of robust PCA

Yi-Hui Zhou

Department of Biological Sciences
North Carolina State University
e-mail: yzhou19@ncsu.edu

and

J. S. Marron

Department of Statistics and Operations Research
University of North Carolina Chapel Hill
e-mail: marron@unc.edu

Abstract: Conventional principal component analysis is highly susceptible to outliers. In particular, a sufficiently outlying single data point, can draw the leading principal component toward itself. In this paper, we study the effects of outliers for high dimension and low sample size data, using asymptotics. The non-robust nature of conventional principal component analysis is verified through inconsistency under multivariate Gaussian assumptions with a single spike in the covariance structure, in the presence of a contaminating outlier. In the same setting, the robust method of spherical principal components is consistent with the population eigenvector for the spike model, even in the presence of contamination.

Keywords and phrases: Outlier, robustness, spherical PCA, spike model.

Received April 2014.

Contents

1. Introduction

Principal components analysis (PCA) is widely used for high dimensional data (Jolliffe [1]), including high dimension, low sample size (HDLSS) data. Classical PCA represents the data using orthogonal components that are ordered according to maximum successive explained variability in the data. For mean-centered data, the principal components can be derived from a spectral decomposition of the sample covariance matrix. Both the sample mean and covariance are sensitive to outlying observations, and so classical PCA tends to be unreliable in the presence of outliers.

There are two viewpoints which give close understanding of conventional PCA. A simple view is eigen analysis of the covariance matrix; a second view is finding directions of maximal variation. For conventional PCA, those two approaches give the same solution. To develop approaches to robust PCA, each viewpoins has led to useful methods which are quite different in nature.

Using the first idea, Devlin and Gnanadesikan [2] did an eigen analysis of a robust estimate of the covariance matrix to develop a robust version of PCA. Their proposed robust covariance was based on robust location and scale estimators, which replaced the usual sample means and covariances. A problem with eigen analysis of such a robust covariance matrix is it is very challenging to get non-negative definite covariance matrix estimates, especially in HDLSS contexts. An interesting solution to that problem is called "minimum volume ellipsoid" (Rousseeuw [3, 4]) that is a multivariate extension of the least median of squares. However this method required $d < n$, again rendering it useless for HDLSS data.

Li and Chen [5] used the second viewpoint of PCA focussing on the notion of direction of maximal variation. They proposed searching directly for the optimal direction to maximize an M-estimator of scale which is called *projection pursuit*. To find more than the first component, they subtract the projected residuals from the data, and apply projection pursuit again. While their method is well defined, i.e. exists, it is computationally intractable in high dimensions.

Locantore et al. [6] proposed a simple robust alternative to PCA. The idea is to first project the data onto a sphere, which will reduce the influence of the outliers. Then classical PCA is performed on the transformed data, resulting in Spherical PCA (SPCA). When the data follows a Gaussian distribution, with a single large eigenvalue, the many data points in the stretched ellipsoid will project to ice caps on the sphere, so SPCA will find essentially the same direction of maximal variation. SPCA has a close relationship to the idea of "multivariate signs", see Oja [7] for a good introduction to this area. In particular, the good robustness properties of SPCA are not surprising, because it is just the PCA of the sign representations of the data.

The asymptotic behavior of classical PCA for HDLSS data has been established by Jung and Marron [8] under various versions of the spike eigenvalue model, with one or only a few large eigenvalues (Johnstone and Silverman [9]). They explored conditions under which the conventional PCA was consistent in terms of the spike parameter α.

The major contributions of this manuscript are as follows. SPCA is shown to be consistent under the HDLSS asymptotic regime, under the same conditions as PCA. In the presence of a Gaussian outlier, conventional PCA is shown to be inconsistent. However, SPCA is shown to still be consistent in the presence of the outlier. The implications of these results are important, as they establish SPCA as an important and robust tool and an attractive alternative to PCA. Here, robustness with respect to outliers and SPCA are for the first time studied rigorously in the HDLSS asymptotic context.

1.1. Notation

Let $A = [X_1, X_2, \ldots, X_n]$ be a $d \times n$ data matrix, with fixed sample size n and large dimension $d \to \infty$, where the samples $X_j = (X_{1j}, \ldots, X_{dj})^T$, $j = 1, 2, \ldots, n$ are independent identically distributed (i.i.d.) random vectors with mean zero and unknown population covariance matrix Σ, i.e. $X_j \sim N(0, \Sigma)$. PCA is essentially equivalent to singular value decomposition (SVD) on the mean centered data matrix. The left population eigenvector matrix U is the population eigenvector matrix of Σ. The first column of U is μ_1. The SVD of A is

$$A = \widehat{U}\widehat{S}\widehat{V}^T,$$

where $\widehat{U}$ is the left sample eigenvector matrix, $\widehat{V}$ is the right sample eigenvector matrix. The diagonal entries of $\widehat{S}$ are the square roots of the non-zero eigenvalues of both AA^T and A^TA. The sample covariance matrix is $\widehat{\Sigma}$. We use "$f(d) \sim g(d)$" to denote $\lim_{d\to\infty} \frac{f(d)}{g(d)} = 1$ and "$f(d) \to constant$" to denote $\lim_{d\to\infty} f(d) = constant$. We also use "$=^L$" to mean equal in law. The symbol "$\gg$" is used for an approximation of the much greater than sign.

1.2. Spiked covariance model

One challenge of HDLSS data is that conventional principal component analysis may give inaccurate estimation of the population eigenvalues and eigenvectors. For example, all but $\mathbf{n}$ of the eignvalues of $\widehat{\Sigma}$ must be 0. HDLSS asymptotics have provided many useful insights through studying the case where a small subset of the eigenvalues are much larger than the rest (Jung and Marron [8]). This is called the spiked covariance model. The spiked covariance model assumes a covariance matrix of the type $\Sigma = U\Lambda U^T$, $\Lambda = diag(\tau_1, \tau_2, \ldots, \tau_p, \sigma, \ldots, \sigma)$, $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p > \sigma > 0$, for some $1 \leq p < d$, where U is a $d \times d$ orthogonal matrix. In this manuscript, we consider the informative simple case where $\Sigma = diag(d^\alpha, 1, \ldots, 1)$, for $\alpha > 0$. After understanding this simple scenario, it is easier to extend the result to the full multi-spike model (Jung and Marron [8]).

1.3. Spherical PCA

The Spherical Principal Components Analysis (SPCA) procedure was derived by Locantore et al. [6] as a robust functional data analysis method. The idea

is to perform classical PCA on the data, projected onto a unit sphere. Let c be the L_1 median and $Y_i = (X_i - c)/||X_i - c||$ be the projected data. Locantore's procedure consists of using the eigenvectors of the covariance matrix of Y_i.

Let $\widetilde{A}$ be A's projection on the sphere. Correspondingly, the SVD of $\widetilde{A}$ is

$$\widetilde{A} \quad = \quad \widehat{\widetilde{U}}\widehat{\widetilde{S}}\widehat{\widetilde{V}}^T, \tag{1.1}$$

where we use $\widehat{\widetilde{*}}$ to denote the estimation on the sphere. For example, let $\widehat{\widetilde{\mu}}_1$ be the first column of $\widehat{\widetilde{U}}$ which is the left sample eigenvector matrix of $\widetilde{A}$.

A simple example to illustrate how SPCA, i.e. projection onto a sphere, leads to robust PCA. A two dimensional dataset is simulated from the multivariate normal distribution with $\Sigma = \left(\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 1 \end{smallmatrix}\right)$. Among the 100 samples, 2 are outliers (big solid dots in Figure 1). The direction of sample PC1 (dashed line) is influenced by the extreme outliers, so it has a large angle with the population PC1 (thin solid line). After projecting the data on the sphere, the outliers (triangle symbol) did not show much influence of the sample eigenvectors which overlay with population PC1. This example motivates handling outliers by using SPCA for HDLSS data (Locantore et al. [6]). In this manuscript, we assume the data has been centered, i.e. $c = 0$.

1.4. Consistency and strong inconsistency

In our HDLSS study of the impact of outliers on PCA and the usefulness of SPCA in countering that, we use the definition of HDLSS consistency and strong inconsistency (Jung and Marron [8]).

- Consistency: The direction $\widehat{\widetilde{\mu}}_1$ is *consistent* with its population counter part μ_1 if $Angle(\widehat{\widetilde{\mu}}_1, \mu_1) \to^p 0°$ or $180°$ as $d \to \infty$. Note that $180°$ is included because the sign of the eigen direction μ_1 is arbitrary.
- Strong inconsistency: The direction $\widehat{\widetilde{\mu}}_1$ is said to be *strongly inconsistent* with its population counter part μ_1 if $Angle(\widehat{\widetilde{\mu}}_1, \mu_1) \to^p 90°$ as $d \to \infty$.

2. Underlying Gaussian model

In this section, we investigate the behavior of the first SPCA component computed from Gaussian distributed data when $d \to \infty$ and n is fixed.

Theorem 2.1. *Given a Gaussian HDLSS data set A, where the jth sample* $X_j \sim N_d(0, \Sigma)$ *with* $\Sigma = diag(d^\alpha, 1, \ldots, 1)$*, where* $\alpha \in R^+$ *with n fixed and* $d \to \infty$*, there are two important cases*

- *for* $\alpha > 1$*,* $Angle(\widehat{\widetilde{\mu}_1}, \mu_1) \to^p 0°$ *or* $180°$*, i.e. the spherical PC1 direction,* $\widehat{\widetilde{\mu_1}}$*, is consistent to* μ_1*;*
- *for* $0 < \alpha < 1$*,* $Angle(\widehat{\widetilde{\mu}_1}, \mu_1) \to^p 90°$*, i.e. the spherical PC1 direction,* $\widehat{\widetilde{\mu_1}}$ *is strongly inconsistent.*

This shows that SPCA provide us the same consistency properties as conventional PCA.

Proof. In the limit as $d \to \infty$, $\chi^2_{d-1} \sim d-1+\sqrt{2(d-1)}N(0,1) \sim d+O_p(d^{1/2})$ For samples $j = 1, 2, \ldots, n$, let $Z_{j1}, Z_{j2}, \ldots, Z_{jn}$ have the standard normal distribution. It follows that

$$||X_j|| =^L \sqrt{d^\alpha Z_{j1}^2 + \sum_{l=2}^{d} Z_{jl}^2} \tag{2.1}$$

$$= \sqrt{d^\alpha Z_{j1}^2 + O_p(d)}, \tag{2.2}$$

$$X_j^T X_k =^L d^\alpha Z_{j1} Z_{k1} + \sum_{l=2}^{d} Z_{jl} Z_{kl}. \tag{2.3}$$

For the assumed form of Σ, the population eigenvector with respect to the largest eigenvalue is $\mu_1 = (1, 0, \ldots, 0)^T$.

For, the $n = 1$, i.e. one sample case, the spherical sample eigenvector $\widehat{\widetilde{\mu_1}}$ is $\frac{X_1}{||X_1||}$. Therefore the inner product of the spherical sample and population eigenvectors, $\langle \widehat{\widetilde{\mu_1}}, \mu_1 \rangle = \frac{(X_{11}, X_{12}, \ldots, X_{1m})}{||X_1||}(1, 0, \ldots, 0)^T = \frac{X_{11}}{||X_1||}$.

When $\alpha > 1$, $O_p(d^\alpha) \gg d$ in the sense that $\frac{O_p(d^\alpha)}{d} \to \infty$, so by (2.2)

$$\frac{X_{11}}{||X_1||} =^L \frac{d^{\alpha/2} Z_{11}}{\sqrt{d^\alpha Z_{11}^2 + O_p(d)}} \to^d B,$$

where

$$B = \begin{cases} 1 \ w.p.1/2 \\ -1 \ w.p.1/2 \end{cases}$$

i.e. Angle$(\widehat{\widetilde{\mu_1}}, \mu_1) \longrightarrow^d 90° + B90°$, i.e. 0° or 180° w.p. 1/2, and this is consistent.

When $0 < \alpha < 1$, $d \gg O_p(d^\alpha)$, for large d, so we have

$$\frac{X_{11}}{||X_1||} = \frac{O_p(d^{\alpha/2})}{\sqrt{O_p(d)}} = O_p(d^{\frac{\alpha-1}{2}}) \longrightarrow 0.$$

i.e. Angle$(\widehat{\widetilde{\mu_1}}, \mu_1) \longrightarrow^p 90°$.

For $n \geq 2$, in the case of $\alpha > 1$, we have $d^\alpha \gg d$, and by (2.2) and (2.3) the (j, k)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ is

$$\frac{X_j^T X_k}{||X_j||||X_k||} = \frac{d^\alpha Z_{j1} Z_{k1} + \sum_{l=2}^{d} Z_{jl} Z_{kl}}{\sqrt{d^\alpha Z_{j1}^2 + O_p(d)}\sqrt{d^\alpha Z_{k1}^2 + O_p(d)}} = \frac{Z_{j1} Z_{k1}}{|Z_{j1}||Z_{k1}|} + O_p(d^{1-\alpha}). \tag{2.4}$$

So in the limit $\widetilde{A}^T \widetilde{A}$ is the rank 1 matrix which is the outer product of the vector $[\frac{Z_{11}}{|Z_{11}|}, \frac{Z_{21}}{|Z_{21}|}, \ldots, \frac{Z_{n1}}{|Z_{n1}|}]$. The maximum eigenvalue for $\widetilde{A}^T \widetilde{A}$ can be derived

as $(\frac{Z_{11}}{|Z_{11}|})^2 + (\frac{Z_{21}}{|Z_{21}|})^2 + \cdots + (\frac{Z_{n1}}{|Z_{n1}|})^2 = n$ and its corresponding right eigenvector is

$$\widetilde{V_1} = [sign(Z_{11})/\sqrt{n}, sign(Z_{21})/\sqrt{n}, \ldots, sign(Z_{n1})/\sqrt{n}]^T.$$

Using the notation (1.1), $\widehat{\widetilde{U}} = \widehat{\widetilde{A}}\widehat{\widetilde{V}}\widehat{\widetilde{S}}^{-1}$. In particular, the first element of $\widehat{\widetilde{\mu_1}}$ is

$$\left[\frac{Z_{11}}{|Z_{11}|}, \frac{Z_{21}}{|Z_{21}|}, \ldots, \frac{Z_{n1}}{|Z_{n1}|}\right] \begin{pmatrix} sign(Z_{11})/\sqrt{n} \\ sign(Z_{21})/\sqrt{n} \\ \vdots \\ sign(Z_{n1})/\sqrt{n} \end{pmatrix} \frac{1}{\sqrt{n}} = 1.$$

Hence $\langle \widehat{\widetilde{\mu_1}}, \mu_1 \rangle = \langle (1, \ldots), (1, 0, \ldots, 0) \rangle = 1$. We conclude that the dominant sample eigenvector points in the same direction as the corresponding population eigenvector when $\alpha > 1$, i.e. Spherical PCA is consistent.

In the case of $0 < \alpha < 1$, for $j \neq k$, by the Law of Large Numbers $\frac{\sum_{l=2}^{d} Z_{jl} Z_{kl}}{d} \to^p 0$, and by (4.4), it follows that the (j,k)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ is

$$\begin{aligned} \frac{X_j^T X_k}{||X_j|| ||X_k||} &= \frac{\frac{d^\alpha}{d} Z_{j1} Z_{k1} + \frac{\sum_{l=2}^{d} Z_{jl} Z_{kl}}{d}}{\frac{||X_j||}{\sqrt{d}} \frac{||X_k||}{\sqrt{d}}} \\ &\to \frac{0+0}{1 \times 1} \to^p 0. \end{aligned} \tag{2.5}$$

i.e. $\widetilde{A}^T \widetilde{A} \sim I_{n \times n}$, with the largest sample eigenvalue 1 and an arbitrary set of eigenvectors. Therefore the first element of $\widetilde{U}$ is a random direction. Thus, using HDLSS results from Hall, Marron and Neeman [10], the angle between the SPCA eigenvector and the dominant population eigenvector tends to 90°. □

3. Impact of outliers

All the properties mentioned before are based on the assumption that the data follow a Gaussian distribution. In practice, real data often violate that assumption. This happens when there are large outliers, which may not be easily distinguishable, which can severely impact conventional PCA as shown in Figure 1.

When we encounter a potential outlier, one natural viewpoint is that the observation resulted from a mistake or other extraneous effect, and should be discarded (Hampel et al. [11]). In other situations (Huber and Ronchetti [12]), outliers can have useful information, and should be only downweighted, not deleted. In the HDLSS study of Locantore et al. [6], the second view point was particularly relevant. Outliers conveyed important information about the underlying population, which would have been lost by just dropping the observations.

To model a scenario with outliers, we assume the data come from a contaminated normal distribution in which the majority of samples are from a specified

multivariate Gaussian distribution, but a small proportion are from a multivariate Gaussian distribution with much higher variance. In particular, we assume the first $n-1$ samples come from the spiked normal model of section 1.2 and the last sample $X_n \sim N(0, \Sigma_2)$, where $\Sigma_2 = d^\beta I_{n\times n}$. i.e. For β large, there will be a distinct outlier coming from a random direction.

Theorem 3.1. *Given a Gaussian HDLSS data set A, where the first $n-1$ samples $X_j \sim N_d(0, \Sigma)$ with $\Sigma = diag(d^\alpha, 1, \ldots, 1)$, the nth sample $X_n \sim N(0, \Sigma_2)$, with $\Sigma_2 = d^\beta I_{n\times n}$, where α, $\beta \in R^+$ with n fixed and $d \to \infty$, it follows that*

- *when $1 < \alpha$ and $\beta < \alpha$, $Angle(\widehat{\mu_1}, \mu_1) \to^p 0^\circ$ or 180°, i.e. the direction of $\widehat{\mu_1}$ is consistent to μ_1;*
- *for $\alpha < 1$ or $\alpha < \beta$, $Angle(\widehat{\mu_1}, \mu_1) \to^p 90^\circ$, i.e. the direction of $\widehat{\mu_1}$ is asymptotically perpendicular to μ_1, i.e. is strongly inconsistent.*

Theorem 3.1 mathematically quantifies the extent to which $\widehat{\mu_1}$ is severely influenced by outliers.

Proof. For the nth sample,

$$||X_n|| \quad =^L \quad \sqrt{d^{\beta/2}\chi^2_d} \tag{3.1}$$

In the conventional principal component analysis, the sample covariance matrix of A is

$$\frac{1}{n}AA^T = \frac{1}{n}\sum_{i=1}^{n-1} X_i X_i^T + \frac{1}{n}X_n X_n^T$$

- For the case of $1 < \alpha$ and $\beta < \alpha$, by (2.2) and (3.1),

$$\frac{AA^T}{d^\alpha} \sim \begin{pmatrix} \frac{(n-1)d^\alpha + d^\beta}{d^\alpha} & . & . & 0 \\ 0 & \frac{(n-1)+d^\beta}{d^\alpha} & . & 0 \\ . & . & . & . \\ . & . & . & \frac{(n-1)+d^\beta}{d^\alpha} \end{pmatrix}.$$

  i.e. $AA^T \sim diag((n-1)d^\alpha, 0, \ldots, 0)$, where the symbol $\sim$ means elementwise approximation asymptotically with $d \to \infty$. By eigen analysis, there is only one dominant eigenvalue of $\frac{AA^T}{n}$ and $\widehat{\mu_1}$ is consistent with μ_1.
- For the case of $\alpha < \beta$ with fixed n and $d \to \infty$, since $(n-1)d^\alpha + d^\beta \sim d^\beta$, in the sense that $\frac{(n-1)d^\alpha + d^\beta}{d^\beta} \to 1$

$$\frac{AA^T}{d^\beta} = \begin{pmatrix} \frac{(n-1)d^\alpha + d^\beta}{d^\beta} & . & . & 0 \\ 0 & \frac{(n-1)+d^\beta}{d^\beta} & . & 0 \\ . & . & . & . \\ . & . & . & \frac{(n-1)+d^\beta}{d^\beta} \end{pmatrix} \sim I_d.$$

  Therefore the sample covariance matrix of A approximates Σ_2 as $d \to \infty$. Thus the first sample eigen vector has a random direction.

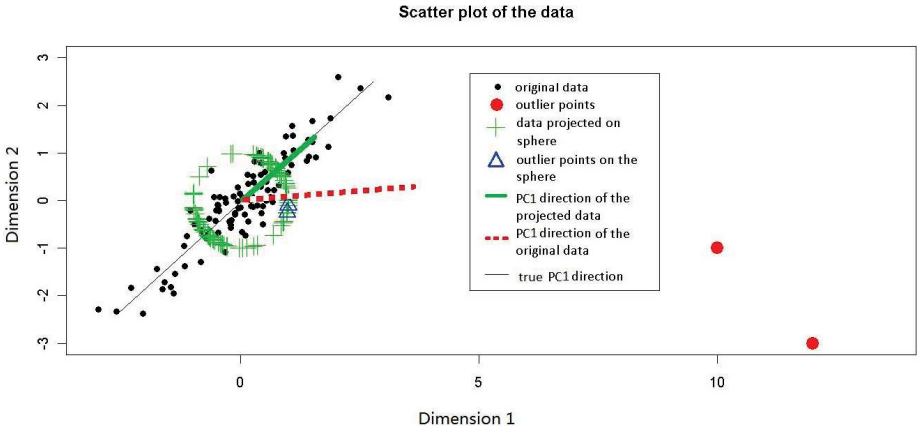

FIG 1. *Two dimensional example showing how outliers (red dots) can strongly impact the conventional PC1 direction (red dashed vector) and how spherical PCA downweights the influence of outliers (blue triangles), giving an SPC1 direction (green vector) which is a much better estimate of the true first population eigen direction (black line).*

- In the case of $\alpha < 1$, if $\alpha < 1 < \beta$, then it belongs to the case with $\alpha < \beta$; if $\alpha, \beta < 1$, by (2.2), for any $j < n$, we have $\frac{||X_j||^2}{d} \to 1$.

$$\frac{||X_n||^2}{d} \quad \to^p \quad 0.$$

For any $j < n$,

$$\frac{X_j^T X_n}{d} \quad \sim \quad \frac{d^{\alpha/2+\beta/2} Z_{j1} Z_{n1} + d^{\beta/2} \sum_{l=2}^{d} Z_{jl} Z_{nl}}{d} \quad \to^p \quad 0.$$

Similarly, for any $j < n$, $k < n$, $j \neq k$,

$$\frac{X_j^T X_k}{d} \quad = \quad \frac{d^{\alpha} Z_{j1} Z_{k1} + \sum_{l=2}^{d} Z_{jl} Z_{kl}}{d} \quad \to^p \quad 0.$$

Hence the dual sample covariate matrix satisfies $\frac{A^T A}{d} \sim \begin{pmatrix} I_{n-1} & 0 \\ 0 & 0 \end{pmatrix}$. Based on proposition 1 (Jung and Marron [8]), $\frac{\lambda_1}{d} \to$ constant which is greater than 0 and $Angle(\widehat{\mu}_1, \mu_1) \to^p 90°$ as $d \to \infty$. □

From the above, we conclude that conventional PCA is very sensitive to an outlier. However, as shown in Figure 1, if we project the data on the sphere, SPCA can be very robust to outliers. This direction is asymptotically studied in the next theorem.

Theorem 3.2. *Given a Gaussian HDLSS data set A (the same as the data set A in Theorem 3.1), we have*

- *for $\alpha > 1$, $Angle(\widehat{\widetilde{\mu}_1}, \mu_1) \to^p 0^\circ$ or 180°, i.e. the direction of $\widehat{\widetilde{\mu}_1}$ is consistent to μ_1;*
- *for $0 < \alpha < 1$, $Angle(\widehat{\widetilde{\mu}_1}, \mu_1) \to^p 90^\circ$, i.e. the direction of $\widehat{\widetilde{\mu}_1}$ is strongly inconsistent.*

Note that this result is independent of β which can be arbitrarily large and still not affect the consistency of SPCA.

Proof. In the case of $\alpha > 1$, the (j, n)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ $(j \neq n)$ is

$$\begin{aligned} \frac{X_j^T X_n}{||X_j|| ||X_n||} &= \frac{d^{\frac{\alpha+\beta}{2}} Z_{j1} Z_{n1} + d^{\frac{\beta}{2}} \sum_{l=2}^d Z_{jl} Z_{nl}}{||X_j|| ||X_n||} \\ &= \frac{d^{\alpha/2} Z_{j1} Z_{n1} + \sum_{l=2}^d Z_{jl} Z_{nl}}{||X_j|| \frac{||X_n||}{d^{\beta/2}}} \\ &\to^p \frac{d^{\alpha/2} Z_{j1} Z_{n1} + \sum_{l=2}^d Z_{jl} Z_{nl}}{||X_j||}. \end{aligned} \tag{3.2}$$

Using (2.2)

$$\begin{aligned} \frac{X_j^T X_n}{||X_j|| ||X_n||} &= \frac{d^{\alpha/2} Z_{j1} Z_{n1} + \sum_{l=2}^d Z_{jl} Z_{nl}}{\sqrt{d^\alpha Z_{j1}^2 + O_p(d)}} \\ &= \frac{Z_{j1} Z_{n1} + \frac{\sum_{l=2}^d Z_{jl} Z_{nl}}{d^{\alpha/2}}}{\frac{\sqrt{d^\alpha Z_{j1}^2 + O_p(d)}}{d^{\alpha/2}}} \\ &\to^p \frac{Z_{j1} Z_{n1}}{|Z_{j1}|} \text{ in probability.} \end{aligned}$$

The (j, k)th $(j \neq n, k \neq n, j \neq k)$ entry of the matrix $\widetilde{A}^T \widetilde{A}$ is

$$\frac{X_j^T X_k}{||X_j|| ||X_k||} \sim \frac{Z_{j1} Z_{k1}}{|Z_{j1}||Z_{k1}|}, j \neq k.$$

Thus in the case of $\alpha > 1$, where $d^\alpha \gg d$, the (j, k)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ is

$$\frac{X_j^T X_k}{||X_j|| ||X_k||} = \begin{pmatrix} \frac{X_j^T X_k}{||X_j|| ||X_k||}, j \neq k, j, k \neq n \\ \frac{Z_{j1} Z_{n1}}{|Z_{j1}|}, j \neq n \\ 1, j = k, j \neq n \end{pmatrix}.$$

In another words, $\widetilde{A}^T \widetilde{A}$ is the rank 1 matrix which is the outer product of the vector $[\frac{Z_{11}}{|Z_{11}|}, \frac{Z_{21}}{|Z_{21}|}, \ldots, \frac{Z_{n-11}}{|Z_{n-11}|}, Z_{n1}]$. The maximum eigenvalue for $\widetilde{A}^T \widetilde{A}$ can be derived as $(\frac{Z_{11}}{|Z_{11}|})^2 + (\frac{Z_{21}}{|Z_{21}|})^2 + \cdots + (\frac{Z_{n-11}}{|Z_{n-11}|})^2 + Z_{n1}^2 = n - 1 + Z_{n1}^2$ and its

corresponding eigenvector $\widetilde{V_1}$ is

$$\begin{pmatrix} \frac{Z_{11}}{|Z_{11}|}/\sqrt{n-1+Z_{n1}^2} \\ \vdots \\ \frac{Z_{n-11}}{|Z_{n-11}|}/\sqrt{n-1+Z_{n1}^2} \\ Z_{n1}/\sqrt{n-1+Z_{n1}^2} \end{pmatrix}.$$

Following the same argument as in the proof of Theorem 2.1, the first element of $\widehat{\widetilde{\mu_1}}$ is

$$\left[\frac{Z_{11}}{|Z_{11}|}, \frac{Z_{21}}{|Z_{21}|}, \ldots, \frac{Z_{n-11}}{|Z_{n-11}|}, Z_{n1}\right] \times \begin{pmatrix} sign(Z_{11})/\sqrt{n-1+Z_{n1}^2} \\ \vdots \\ sign(Z_{n-11})/\sqrt{n-1+Z_{n1}^2} \\ Z_{n1}/\sqrt{n-1+Z_{n1}^2} \end{pmatrix} \frac{1}{\sqrt{n-1+Z_{n1}^2}} = 1.$$

Hence $\langle \widehat{\widetilde{\mu_1}}, \mu_1 \rangle = 1$. We conclude that SPCA is consistent when $\alpha > 1$ with an outlier sample.

In the case of $0 < \alpha < 1$, $0 < \beta < 1$, for $j \neq n$, by (4.7) and the same logic as in (4.6), it follows that the (j,k)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ is

$$\frac{X_j^T X_n}{||X_j||||X_n||} \quad \to^p \quad \frac{d^{\alpha/2} Z_{j1} Z_{n1} + \sum_{l=2}^{d} Z_{jl} Z_{nl}}{||X_j||} \to 0 \text{ } \textit{in probability.} \quad (3.3)$$

For $j \neq k, j, k \neq n$, the (j,k)th entry of $\widetilde{A}^T \widetilde{A}$ is the same as (4.8). Thus $\widetilde{A}^T \widetilde{A} \sim I_{n \times n}$, with the largest sample eigenvalue 1 and no-fixed eigenvector corresponding to it. Therefore the first element of $\widetilde{U}$ is random. On the high dimensional sphere, the angle between the sample eigenvector and the dominant population eigenvector tends to 90°. □

In robust statistics, distributions containing outliers are commonly modeled by the contaminated normal model:

$$(1-\epsilon)N(0,\Sigma) + \epsilon N(0,\Sigma_2).$$

Good overview of these ideas and their historical background can be found in Huber [13]. The model with a single outlier in Theorems 3.1, 3.2, 4.2, 4.3 is slightly different from this because the number of outliers are not random. This was done to give simpler and more revealing insights. We conjecture that entirely parallel results can be derivied for the contaminated normal model.

4. Muti-spike model

The following three theorems extend the results of Section 2 and 3 to the multi-spike case. Here the single spike large eigenvalue from Section 3 is replaced by

several large eigenvalues of the form $\lambda_1 = d^{\alpha_1}, \ldots, \lambda_m = d^{\alpha_m}$, where $\alpha_1 > \alpha_2 > \cdots > \alpha_m > 0$. The remaining eigenvalues are assumed to be $\lambda_{m+1} = 1, \ldots, \lambda_d = 1$. From Theorem 2.1, it is not surprising that an important threshold is t where $\alpha_t > 1 > \alpha_{t+1}$.

As in section 2, first the performance of SPCA in the non-outlier case is considered. As there SPCA has asymptotic properties that are very similar to PCA.

Theorem 4.1. *Given a Gaussian HDLSS data set A, where the jth sample $X_j \sim N_d(0, \Sigma)$ with $\Sigma = diag(d^{\alpha_1}, d^{\alpha_2}, \ldots, d^{\alpha_m}, 1, \ldots, 1)$, with $n > t$ fixed in the limit as $d \to \infty$, there are two groups of eigenvalues*

- *$Angle(\widehat{\widetilde{\mu}}_l, \mu_l) \to^p 0^\circ$ or 180°, for $l = 1, 2, \ldots, t$, i.e. the spherical PC l direction, $\widehat{\widetilde{\mu}}_l$, is consistent to μ_l;*
- *$Angle(\widehat{\widetilde{\mu}}_l, \mu_l) \to^p 90^\circ$, i.e. the spherical PC l direction, $\widehat{\widetilde{\mu}}_l$ is strongly inconsistent for $l = t+1, \ldots, m$.*

Proof. As in the proof of Theorem 2.1

$$||X_j|| =^L \sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl}^2 + \sum_{l=m+1}^{d} Z_{jl}^2} \tag{4.1}$$

$$= \sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl}^2 + O_p(d)}, \tag{4.2}$$

$$X_j^T X_k =^L \sum_{l=1}^{m} d^{\alpha_l} Z_{jl} Z_{kl} + \sum_{l=m+1}^{d} Z_{jl} Z_{kl}. \tag{4.3}$$

For $l = 1, \ldots, t$, we have $d^{\alpha_1} \gg d^{\alpha_2} \gg \cdots \gg d^{\alpha_t} \gg d$, and by (4.2) and (4.3) the (j,k)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ is

$$\frac{X_j^T X_k}{||X_j|| ||X_k||} = \frac{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl} Z_{kl} + \sum_{l=m+1}^{d} Z_{jl} Z_{kl}}{\sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl}^2 + O_p(d)} \sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{kl}^2 + O_p(d)}} \tag{4.4}$$

Now we focus on the case $l = 1$,

$$\frac{X_j^T X_k}{||X_j|| ||X_k||} = \frac{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl} Z_{kl} / d^{\alpha_1} + \sum_{l=m+1}^{d} Z_{jl} Z_{kl} / d^{\alpha_1}}{\sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl}^2 / d^{\alpha_1} + O_p(d)} \sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{kl}^2 / d^{\alpha_1} + O_p(d)}} \tag{4.5}$$

$$\sim \frac{Z_{j1} Z_{k1}}{|Z_{j1}||Z_{k1}|}.$$

So in the limit $\widetilde{A}^T \widetilde{A}$ is the rank 1 matrix and the rest proof is very similar to Theorem 2.1. In particular $\langle \widehat{\widetilde{\mu}_1}, \mu_1 \rangle = \langle (1, \ldots), (1, 0, \ldots, 0) \rangle = 1$, which shows that $Angle(\widehat{\widetilde{\mu}_1}, \mu_1) \to^p 0^\circ$ or 180°.

On the case $l = 2$, we move away the first PC direction, $\widetilde{A} - P_{\mu_1}\widetilde{A} \sim N(0, \Sigma_{l=2})$, where $\Sigma_{l=2} = diag(d^{\alpha_2}, d^{\alpha_3}, \ldots, d^{\alpha_m}, 1, \ldots, 1)$. Similarly, we can prove that$(\widetilde{A} - P_{\mu_1}\widetilde{A})^T(\widetilde{A} - P_{\mu_1}\widetilde{A})$ is the rank 1 matrix. Following the same proof above, we get $Angle(\widehat{\widetilde{\mu}_2}, \mu_2) \to^p 0^\circ$ or 180°.

Iteratively, for $l = 3, \ldots, t$, we look at $\widetilde{A} - P_{\mu_1,\mu_2,\ldots,\mu_{l-1}}\widetilde{A} \sim N(0, \Sigma_l)$, where $\Sigma_l = diag(d^{\alpha_l}, \ldots, d^{\alpha_m}, 1, \ldots, 1)$. Using the similar arguments, we conclude that $Angle(\widehat{\widetilde{\mu}}_l, \mu_l) \to^p 0^\circ$ or 180° for $l = 3, \ldots, t$.

In the case of $0 < \alpha_m < \cdots < \alpha_{t+1} < 1$, for $j \neq k$, by the Law of Large Numbers $\frac{\sum_{l=2}^{d} Z_{jl}Z_{kl}}{d} \to^p 0$, and by (4.4), it follows that the (j,k)th entry of the matrix $\widetilde{A}^T\widetilde{A}$ is

$$\begin{aligned} \frac{X_j^T X_k}{||X_j||||X_k||} &= \frac{\frac{d^\alpha}{d} Z_{j1}Z_{k1} + \frac{\sum_{l=2}^{d} Z_{jl}Z_{kl}}{d}}{\frac{||X_j||}{\sqrt{d}}\frac{||X_k||}{\sqrt{d}}} \\ &\to \frac{0+0}{1\times 1} \to 0 \textit{ in probability.} \end{aligned} \tag{4.6}$$

i.e. $\widetilde{A}^T\widetilde{A} \sim I_{n\times n}$, with the largest sample eigenvalue 1 and an arbitrary set of eigenvectors. Therefore the first element of $\widetilde{U}$ is a random direction. Therefore the spherical PC l direction, $\widehat{\widetilde{\mu}}_l$ is strongly inconsistent for $l = t+1, \ldots, m$. □

However conventional PCA is not robust to outliers.

Theorem 4.2. *Given a Gaussian HDLSS data set A, where the first $n-1$ samples $X_j \sim N_d(0, \Sigma)$ with $\Sigma = diag(d^{\alpha_1}, d^{\alpha_2}, \ldots, d^{\alpha_m}, 1, \ldots, 1)$, the nth sample $X_n \sim N(0, \Sigma_2)$, with $\Sigma_2 = d^\beta I_{n\times n}$, where $\alpha_1 > \alpha_2 > \cdots > \alpha_t > 1 > \alpha_{t+1} > \cdots > \alpha_m > 0$, $\beta \in R^+$ with n fixed and $d \to \infty$, it follows that*

- *when $1 < \alpha_l$ and $\beta < \alpha_l$, $l = 1, 2, \ldots, t$, $Angle(\widehat{\mu}_l, \mu_l) \to^p 0^\circ$ or 180°, i.e. the direction of $\widehat{\mu}_l$ is consistent to μ_l;*
- *for $\alpha_i < 1$ or $\alpha_l < \beta$, for $l = t+1, \ldots, m$, $Angle(\widehat{\mu}_l, \mu_l) \to^p 90^\circ$, i.e. the direction of $\widehat{\mu}_l$ is asymptotically perpendicular to μ_l, i.e. is strongly inconsistent.*

Proof. Similar to the proof in Theorem 3.1

- For the case of $1 < \alpha_l$ and $\beta < \alpha_l$,

$$\frac{AA^T}{d^{\alpha_l}} \sim \begin{pmatrix} \frac{(n-1)d^{\alpha_1}+d^\beta}{d^{\alpha_l}} & . & . & 0 \\ 0 & \frac{(n-1)d^{\alpha_2}+d^\beta}{d^{\alpha_l}} & . & 0 \\ . & . & . & . \\ . & . & . & \frac{(n-1)+d^\beta}{d^{\alpha_l}} \end{pmatrix}.$$

  i.e. $AA^T \sim diag((n-1)d^{\alpha_1}, (n-1)d^{\alpha_2}, (n-1)d^{\alpha_3}, \ldots, (n-1)d^{\alpha_t}, \ldots, 0)$, where the symbol $\sim$ means element-wise approximation asymptotically with $d \to \infty$. By eigen analysis, there are t dominant eigenvalue of $\frac{AA^T}{n}$ and $\widehat{\mu}_l$ is consistent with μ_l, for $l = 1, 2, \ldots t$.

- For the case of $\alpha_l < \beta$ with fixed n and $d \to \infty$, since $(n-1)d^{\alpha_l} + d^{\beta} \sim d^{\beta}$, in the sense that $\frac{(n-1)d^{\alpha_l}+d^{\beta}}{d^{\beta}} \to 1$

$$\frac{AA^T}{d^{\beta}} \sim I_d.$$

  Therefore the sample covariance matrix of A approximates Σ_2 as $d \to \infty$. Thus the first sample eigen vector has a random direction.
- The proof of the case of $\alpha_l < 1$ is very similar to that in Theorem 3.1. In particular, for any $j < n$, $k < n$, $j \neq k$,

$$\begin{aligned} \frac{X_j^T X_k}{d} &= \frac{\sum_{l=t+1}^{m} d^{\alpha_l} Z_{jl} Z_{kl} + \sum_{l=m+1}^{d} Z_{jl} Z_{kl}}{d} \\ &\to^p 0 \end{aligned}$$

  We can easily show that $\frac{A^T A}{d} \sim \left(\begin{smallmatrix} I_{n-1} & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Therefore $Angle(\widehat{\mu}_l, \mu_l) \to^p 90^\circ$ as $d \to \infty$ for $l = t+1, \ldots, m$. □

Spherical PCA gives the same clean asymptotic properties as in Theorem 4.1 when we have contaminated samples.

Theorem 4.3. *Given a Gaussian HDLSS data set A (the same as the data set A in Theorem 4.2), we have*

- *$Angle(\widehat{\widetilde{\mu}}_l, \mu_l) \to^p 0^\circ$ or 180°, for $l = 1, 2, \ldots, t$, i.e. the spherical PC l direction, $\widehat{\widetilde{\mu}}_l$, is consistent to μ_l;*
- *$Angle(\widehat{\widetilde{\mu}}_l, \mu_l) \to^p 90^\circ$, i.e. the spherical PC l direction, $\widehat{\widetilde{\mu}}_l$ is strongly inconsistent for $l = t+1, \ldots, m$.*

Proof. The main proof is very similar to the proof of Theorem 4.1. Here we only show the case with $\alpha_1 > 1$. In the case of $\alpha_l > 1$, the (j, n)th entry of the matrix $\widetilde{A}^T \widetilde{A}$ $(j \neq n)$ is

$$\begin{aligned} \frac{X_j^T X_n}{||X_j|| ||X_n||} &= \frac{\sum_{l=1}^{m} d^{\frac{\alpha_l+\beta}{2}} Z_{jl} Z_{nl} + d^{\frac{\beta}{2}} \sum_{l=m+1}^{d} Z_{jl} Z_{nl}}{||X_j|| ||X_n||} \\ &= \frac{\sum_{l=1}^{m} d^{\alpha_l/2} Z_{jl} Z_{nl} + \sum_{l=m+1}^{d} Z_{jl} Z_{nl}}{||X_j|| \frac{||X_n||}{d^{\beta/2}}} \\ &\to^p \frac{\sum_{l=1}^{m} d^{\alpha_l/2} Z_{jl} Z_{nl} + \sum_{l=m+1}^{d} Z_{jl} Z_{nl}}{||X_j||}. \end{aligned} \tag{4.7}$$

Using (4.2)

$$\begin{aligned} \frac{X_j^T X_n}{||X_j|| ||X_n||} &= \frac{\sum_{l=1}^{m} d^{\alpha_l/2} Z_{jl} Z_{nl} + \sum_{l=m+1}^{d} Z_{jl} Z_{nl}}{\sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl}^2 + O_p(d)}} \\ &= \frac{Z_{j1} Z_{n1} + \frac{\sum_{l=2}^{d} Z_{jl} Z_{nl}}{d^{\alpha/2}}}{\frac{\sqrt{\sum_{l=1}^{m} d^{\alpha_l} Z_{jl}^2 + O_p(d)}}{d^{\alpha/2}}} \end{aligned}$$

$$\rightarrow \quad \frac{Z_{j1}Z_{n1}}{|Z_{j1}|} \textit{ in probability.}$$

The (j,k)th $(j \neq n,\ k \neq n,\ j \neq k)$ entry of the matrix $\widetilde{A}^T\widetilde{A}$ is

$$\frac{X_j^T X_k}{||X_j||||X_k||} \sim \frac{Z_{j1}Z_{k1}}{|Z_{j1}||Z_{k1}|}, j \neq k.$$

The (j,k)th entry of the matrix $\widetilde{A}^T\widetilde{A}$ is

$$\frac{X_j^T X_k}{||X_j||||X_k||} = \begin{pmatrix} \frac{X_j^T X_k}{||X_j||||X_k||}, j \neq k, j,k \neq n \\ \frac{Z_{j1}Z_{n1}}{|Z_{j1}|}, j \neq n \\ 1, j = k, j \neq n \end{pmatrix}.$$

We can easily show that $\langle \widehat{\widetilde{\mu_1}}, \mu_1 \rangle = 1$. Iteratively, for $l = 2, \ldots, t$, $(\widetilde{A} - P_{\mu_1,\ldots,\mu l-1}\widetilde{A})^T(\widetilde{A} - P_{\mu_1,\ldots,\mu l-1}\widetilde{A})$ is a rank 1 matrix. Using arguments as in Theorem 3.2, we get $\langle \widehat{\widetilde{\mu}}_l, \mu_l \rangle = 1$, therefore $Angle(\widehat{\widetilde{\mu}}_l, \mu_l) \rightarrow^p 0^\circ$ or 180°, for $l = 1, 2, \ldots, t$, i.e. the spherical PC l direction, $\widehat{\widetilde{\mu}}_l$, is consistent to μ_l. □

5. Discussion

A key assumption of this paper made to give direct access to the critical robustness insights is that the data are properly centered. In practice, the centering can be an important issue. Centering using the L1 M-estimate is recommended (Locantore et al. [6]), because that is intuitively consistent with spherical PCA. An interesting potential approach suggested by a reviewer, is to tackle the centering issue by applying SPCA to the pairwise differences of the data. An interesting open problem is the impact of the estimation on the asymptotics, which we conjecture will be negligible. Detailed investigation of this can be done essentially using Taylor expansion methods on $(\widehat{c} - c)$, where $\widehat{c}$ is the sample version of the L1 M-estimate and c is the true popoulation center. This is not pursued here for two reasons. First the asymptotic behavior of $(\widehat{c} - c)$ needs to be analyzed, and to our knowledge this does not appear in the literature. Second, the relatively streamlined and insightful (about the robustness of the PCA direction, which is the point of this paper) proofs we currently have will tend to be obscured having by $(\widehat{c} - c)$ terms appearing in the analysis.

Also worthwhile would be extension of the theory in other directions, including more general distributional assumptions and outlier configurations. Another challenge for future work is the special case at the boundary $\alpha = 1$, where conventional PCA was explored in Jung, Sen and Marron [14]. We believe that parallel results can also be established under appropriate non-Gaussian models, using e.g. sufficient moment conditions, based on ideas from Yata and Aoshima [15]. Theorems 3.2 and 4.3 suggest good breakdown properties of SPCAs. Another interesting open problem is precise quantification of breakdown (Hampel et al. [11]).

Acknowledgement

This work was supported by 1 R01 MH101819-01.

References

[1] Jolliffe, I. (2002). *Principal Component Analysis*. Springer. MR2036084
[2] Devlin, S. J. and Gnanadesikan, R. (1981). Robust estimation of dispersion matrices and principal components. *JASA* **76**.
[3] Rousseeuw, P. J. (1985). Least median of squares regression. *Journal of the American Statistical Association* **79**. MR0770281
[4] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* **B**. MR0851060
[5] Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *The Annals of Statistics* **80**(391).
[6] Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999). Robust principal component analysis for functional data. *Test* **8**(1). MR1707596
[7] Oja, H. (2010). *Multivariate Nonparametric Methods with R*. Springer. MR2598854
[8] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics* **37**. MR2572454
[9] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* **32**(4). MR2089135
[10] Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society* **67**. MR2155347
[11] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley. MR0829458
[12] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley. MR2488795
[13] Huber, P. (1972). The 1972 Wald Lecture Robust Statistics: A review. *The Annals of Mathematical Statistics* **43**(4). MR0314180
[14] Jung, S., Sen, A. and Marron, J. S. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *The Journal of Multivariate Analysis* **109**. MR2922863
[15] Yata, K. and Aoshima, M. (2010). Low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis* **101**(9). MR2671201