## Southern Illinois University Carbondale OpenSIUC

Spring 4-4-2014

# ROBUST PRINCIPAL COMPONENT ANALYSIS 

Ayed Rheal Alanzi<br>Southern Illinois University Carbondale, auid1403@hotmail.com

Follow this and additional works at: http:/ / opensiuc.lib.siu.edu/gs_rp

## Recommended Citation

Alanzi, Ayed Rheal, "ROBUST PRINCIPAL COMPONENT ANALYSIS" (2014). Research Papers. Paper 457.
http://opensiuc.lib.siu.edu/gs_rp/457

# ROBUST PRINCIPAL COMPONENT ANALYSIS 

by<br>Ayed Rheal Alanzi<br>M.S., Malaya University, 2009

A Research Paper<br>Submitted in Partial Fulfillment of the Requirements for the Master of Science Degree

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
May, 2014

Copyright by Ayed Rheal Alanzi, 2014
All Rights Reserved

# RESEARCH PAPER APPROVAL 

Robust Principal Component Analysis

By
Ayed Rheal Alanzi

A Research Paper Submitted in Partial
Fulfillment of the Requirements
for the Degree of
Master of Science in the field of Mathematics

Approved by:
Chair, David Olive

Bhaskar Bhattacharya
Michael Sullivan

Graduate School
Southern Illinois University Carbondale April 2, 2014

## AN ABSTRACT OF THE RESEARCH PAPER OF

Ayed Rheal Alanzi, for the Master of Science in Mathematics, presented on April 2, 2014, at Southern Illinois University Carbondale.

TITLE: Robust Principal Component Analysis

PROFESSOR: Dr. David Olive

A common technique for robust dispersion estimators is to apply the classical estimator to some subset $U$ of the data. Applying principal component analysis to the subset $U$ can result in a robust principal component analysis with good properties.

KEY WORDs: multivariate location and dispersion, principal components, outliers, scree plot.

## TABLE OF CONTENTS

Abstract ..... iii
List of Tables ..... v
List of Figures ..... vi
Introduction ..... 1
1 Robust Principal Component Analysis ..... 8
2 Examples and Simulations ..... 10
3 Conclusions ..... 20
References ..... 22
Vita ..... 24

## LIST OF TABLES

2.1 Estimation of $\boldsymbol{\Sigma}$ with $\gamma=0.4, n=35 p$. . . . . . . . . . . . . . . . . . . . . . 10
2.2 Variance Explained by PCA and RPCA, $p=4$. . . . . . . . . . . . . . . . . . 17
2.3 Variance Explained by PCA and RPCA, SSD $=10^{7}$ SD, $p=50 \ldots 18$

## LIST OF FIGURES

2.1 First Two Principal Components for Buxton data. . . . . . . . . . . . . . . . . 11
2.2 First Two Robust Principal Components with Outliers Omitted. . . . . . . . . 12
2.3 Robust Scree Plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

## INTRODUCTION

Principal component analysis (PCA) is used to explain the dispersion structure with a few linear combinations of the original variables, called principal components. These linear combinations are uncorrelated if the sample covariance matrix $\boldsymbol{S}$ or the sample correlation matrix $\boldsymbol{R}$ is used as the dispersion matrix. The analysis is used for data reduction and interpretation. The notation $\boldsymbol{e}_{j}$ will be used for orthonormal eigenvectors: $\boldsymbol{e}_{j}^{T} \boldsymbol{e}_{j}=1$ and $\boldsymbol{e}_{j}^{T} \boldsymbol{e}_{k}=0$ for $j \neq k$. The eigenvalue eigenvector pairs of a symmetric matrix $\boldsymbol{\Sigma}$ will be $\left(\lambda_{1}, \boldsymbol{e}_{1}\right), \ldots,\left(\lambda_{p}, \boldsymbol{e}_{p}\right)$ where $\lambda_{1} \geq \lambda_{2} \geq \cdots \geq \lambda_{p}$. The eigenvalue eigenvector pairs of a matrix $\hat{\boldsymbol{\Sigma}}$ will be $\left(\hat{\lambda}_{1}, \hat{\boldsymbol{e}}_{1}\right), \ldots,\left(\hat{\lambda}_{p}, \hat{\boldsymbol{e}}_{p}\right)$ where $\hat{\lambda}_{1} \geq \hat{\lambda}_{2} \geq \cdots \geq \hat{\lambda}_{p}$. The generalized correlation matrix defined below is the population correlation matrix when second moments exist if $\boldsymbol{\Sigma}=c \operatorname{Cov}(\boldsymbol{x})$ for some constant $c>0$ where $\operatorname{Cov}(\boldsymbol{x})$ is the population covariance matrix.

Let $\boldsymbol{\Sigma}=\left(\sigma_{i j}\right)$ be a positive definite symmetric $p \times p$ dispersion matrix. A generalized correlation matrix $\boldsymbol{\rho}=\left(\rho_{i j}\right)$ where $\rho_{i j}=\frac{\sigma_{i j}}{\sqrt{\sigma_{i i} \sigma_{j j}}}$.

PCA is applied to data $\boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{n}$ which are iid from some distribution. If a $p \times 1$ random vector $\boldsymbol{x}$ has joint pdf

$$
\begin{equation*}
f(\boldsymbol{z})=k_{p}|\boldsymbol{\Sigma}|^{-1 / 2} g\left[(\boldsymbol{z}-\boldsymbol{\mu})^{T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right] \tag{1}
\end{equation*}
$$

then $\boldsymbol{x}$ has an elliptically contoured $E C_{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.
The following theorem holds since the eigenvalues and generalized correlation matrix are continuous functions of $\boldsymbol{\Sigma}$. When the distribution of the $\boldsymbol{x}_{i}$ is unknown, then a good dispersion estimator estimates $c \boldsymbol{\Sigma}$ on a large class of distributions where $c>0$ depends on the unknown distribution of $\boldsymbol{x}_{i}$. For example, if the $\boldsymbol{x}_{i} \sim E C_{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, then the sample covariance matrix $\boldsymbol{S}$ estimates $\operatorname{Cov}(\boldsymbol{x})=c_{X} \boldsymbol{\Sigma}$.

Theorem 1. Suppose the dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $\left(\lambda_{1}, \boldsymbol{e}_{1}\right), \ldots,\left(\lambda_{p}, \boldsymbol{e}_{p}\right)$ where $\lambda_{1} \geq \lambda_{2} \geq \cdots \geq \lambda_{p}$. Suppose $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c \boldsymbol{\Sigma}$ for some constant $c>0$. Let the eigenvalue eigenvector pairs of $\hat{\boldsymbol{\Sigma}}$ be $\left(\hat{\lambda}_{1}, \hat{\boldsymbol{e}}_{1}\right), \ldots,\left(\hat{\lambda}_{p}, \hat{\boldsymbol{e}}_{p}\right)$ where $\hat{\lambda}_{1} \geq \hat{\lambda}_{2} \geq \cdots \geq \hat{\lambda}_{p}$.

Then $\hat{\lambda}_{j}(\hat{\boldsymbol{\Sigma}}) \xrightarrow{P} c \lambda_{j}(\boldsymbol{\Sigma})=c \lambda_{j}, \hat{\boldsymbol{\rho}} \xrightarrow{P} \boldsymbol{\rho}$ and $\hat{\lambda}_{j}(\hat{\boldsymbol{\rho}}) \xrightarrow{P} \lambda_{j}(\boldsymbol{\rho})$ where $\lambda_{j}(\boldsymbol{A})$ is the $j$ th eigenvalue of $\boldsymbol{A}$ for $j=1, \ldots, p$.

Eigenvectors $\boldsymbol{e}_{j}$ are not continuous functions of $\boldsymbol{\Sigma}$, and if $\boldsymbol{e}_{j}$ is an eigenvector of $\boldsymbol{\Sigma}$ then so is $-\boldsymbol{e}_{j}$. The software produces $\hat{\boldsymbol{e}}_{j}$ which sometimes approximates $\boldsymbol{e}_{j}$ and sometimes approximates $-\boldsymbol{e}_{j}$ if the eigenvalue $\lambda_{j}$ is unique, since then the set of eigenvectors corresponding to $\lambda_{j}$ has the form $a \boldsymbol{e}_{j}$ for any nonzero constant $a$. The situation becomes worse if some of the eigenvalues are equal, since the possible eigenvectors then span a space of dimension equal to the multiplicity of the eigenvalue. Hence if the multiplicity is two and both $\boldsymbol{e}_{j}$ and $\boldsymbol{e}_{k}$ are eigenvectors corresponding to the eigenvalue $\lambda_{i}$, then $\boldsymbol{e}_{i}=\boldsymbol{x}_{i} /\left\|\boldsymbol{x}_{i}\right\|$ is also an eigenvector corresponding to $\lambda_{i}$ where $\boldsymbol{x}_{i}=a_{j} \boldsymbol{e}_{j}+a_{k} \boldsymbol{e}_{k}$ for constants $a_{j}$ and $a_{k}$ which are not both equal to 0 . The software produces $\hat{\boldsymbol{e}}_{j}$ and $\hat{\boldsymbol{e}}_{k}$ that are approximately in the span of $\boldsymbol{e}_{j}$ and $\boldsymbol{e}_{k}$ for large $n$ by the following theorem, which also shows that $\hat{\boldsymbol{e}}_{i}$ is asymptotically an eigenvector of $\boldsymbol{\Sigma}$ in that $\left(\boldsymbol{\Sigma}-\lambda_{i}\right) \hat{\boldsymbol{e}}_{i} \xrightarrow{P} \mathbf{0}$. It is possible that $\hat{\boldsymbol{e}}_{i, n}$ is arbitrarily close to $\boldsymbol{e}_{i}$ for some values of $n$ and arbitrarily close to $-\boldsymbol{e}_{i}$ for other values of $n$ so that $\hat{\boldsymbol{e}}_{i} \equiv \hat{\boldsymbol{e}}_{i, n}$ oscillates and does not converge in probability to either $\boldsymbol{e}_{i}$ or $-\boldsymbol{e}_{i}$.

Theorem 2. Assume the $p \times p$ symmetric dispersion matrix $\boldsymbol{\Sigma}$ is positive definite.
a) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\hat{\boldsymbol{\Sigma}} \boldsymbol{e}_{i}-\hat{\lambda}_{i} \boldsymbol{e}_{i} \xrightarrow{P} \mathbf{0}$.
b) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma} \hat{\boldsymbol{e}}_{i}-\lambda_{i} \hat{\boldsymbol{e}}_{i} \xrightarrow{P} \mathbf{0}$.

If $\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}=O_{P}\left(n^{-\delta}\right)$ where $0<\delta \leq 0.5$, then
c) $\lambda_{i} \boldsymbol{e}_{i}-\hat{\boldsymbol{\Sigma}} \boldsymbol{e}_{i}=O_{P}\left(n^{-\delta}\right)$, and
d) $\hat{\lambda}_{i} \hat{\boldsymbol{e}}_{i}-\boldsymbol{\Sigma} \hat{\boldsymbol{e}}_{i}=O_{P}\left(n^{-\delta}\right)$.
e) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c \boldsymbol{\Sigma}$ for some constant $c>0$, and if the eigenvalues $\lambda_{1}>\cdots>\lambda_{p}>0$ of $\boldsymbol{\Sigma}$ are unique, then the absolute value of the correlation of $\hat{\boldsymbol{e}}_{j}$ with $\boldsymbol{e}_{j}$ converges to 1 in probability: $\quad\left|\operatorname{corr}\left(\hat{\boldsymbol{e}}_{j}, \boldsymbol{e}_{j}\right)\right| \xrightarrow{P} 1$.

Proof. a) $\hat{\boldsymbol{\Sigma}} \boldsymbol{e}_{i}-\hat{\lambda}_{i} \boldsymbol{e}_{i} \xrightarrow{P} \boldsymbol{\Sigma} \boldsymbol{e}_{i}-\lambda_{i} \boldsymbol{e}_{i}=\mathbf{0}$.
b) Note that $\left(\boldsymbol{\Sigma}-\lambda_{i} \boldsymbol{I}\right) \hat{\boldsymbol{e}}_{i}=\left[\left(\boldsymbol{\Sigma}-\lambda_{i} \boldsymbol{I}\right)-\left(\hat{\boldsymbol{\Sigma}}-\hat{\lambda}_{i} \boldsymbol{I}\right)\right] \hat{\boldsymbol{e}}_{i}=o_{P}(1) O_{P}(1) \xrightarrow{P} \mathbf{0}$.
c) $\lambda_{i} \boldsymbol{e}_{i}-\hat{\boldsymbol{\Sigma}} \boldsymbol{e}_{i}=\boldsymbol{\Sigma} \boldsymbol{e}_{i}-\hat{\boldsymbol{\Sigma}} \boldsymbol{e}_{i}=O_{P}\left(n^{-\delta}\right)$.
d) $\hat{\lambda}_{i} \hat{\boldsymbol{e}}_{i}-\boldsymbol{\Sigma} \hat{\boldsymbol{e}}_{i}=\hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{e}}_{i}-\boldsymbol{\Sigma} \hat{\boldsymbol{e}}_{i}=O_{P}\left(n^{-\delta}\right)$.
e) Note that a) and b) hold if $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$ is replaced by $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c \boldsymbol{\Sigma}$. Hence for large $n$, $\hat{\boldsymbol{e}}_{i} \equiv \hat{\boldsymbol{e}}_{i, n}$ is arbitrarily close to either $\boldsymbol{e}_{i}$ or $-\boldsymbol{e}_{i}$, and the result follows.

Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. The $i$ th squared Mahalanobis distance is

$$
\begin{equation*}
D_{i}^{2}=D_{i}^{2}(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))=\left(\boldsymbol{x}_{i}-T(\boldsymbol{W})\right)^{T} \boldsymbol{C}^{-1}(\boldsymbol{W})\left(\boldsymbol{x}_{i}-T(\boldsymbol{W})\right) \tag{2}
\end{equation*}
$$

for each point $\boldsymbol{x}_{i}$. The population squared Mahalanobis distance corresponding to a population location vector $\boldsymbol{\mu}$ and nonsingular dispersion matrix $\boldsymbol{\Sigma}$ is $D_{\boldsymbol{x}}^{2}(\boldsymbol{\mu}, \boldsymbol{\Sigma})=D_{\boldsymbol{x}}^{2}=$ $(\boldsymbol{x}-\boldsymbol{\mu})^{T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$.

Rule of thumb 1. To use PCA, assume the DD plot of classical versus robust Mahalanobis distances and the subplots of the scatterplot matrix are linear. Want $n>10 p$ for classical PCA and $n>20 p$ for robust PCA that uses the FCH, RFCH or RMVN estimators described in Olive and Hawkins (2010). For classical PCA, use the correlation matrix $\boldsymbol{R}$ instead of the covariance matrix $\boldsymbol{S}$ if $\max _{i=1, \ldots, p} S_{i}^{2} / \min _{i=1, \ldots, p} S_{i}^{2}>2$. If $\boldsymbol{S}$ is used, also do a PCA using $\boldsymbol{R}$.

The trace of a matrix $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$, and if $\boldsymbol{A}$ is a $p \times p$ matrix, then $\operatorname{trace}(\boldsymbol{A})=\operatorname{tr}(\boldsymbol{A})=\sum_{i=1}^{p} \boldsymbol{A}_{i i}=\sum_{i=1}^{p} \lambda_{i}$. Note that $\operatorname{tr}(\operatorname{Cov}(\boldsymbol{x}))=\sigma_{1}^{2}+\cdots+\sigma_{p}^{2}$ and $\operatorname{tr}(\hat{\boldsymbol{\rho}})=p$.

Let dispersion estimator $\hat{\boldsymbol{\Sigma}}$ have eigenvalue eigenvector pairs $\left(\hat{\lambda}_{1}, \hat{\boldsymbol{e}}_{1}\right), \ldots,\left(\hat{\lambda}_{p}, \hat{\boldsymbol{e}}_{p}\right)$ where $\hat{\lambda}_{1} \geq \hat{\lambda}_{2} \geq \cdots \geq \hat{\lambda}_{p}$. Then the $p$ principal components corresponding to the $j$ th case $\boldsymbol{x}_{j}$ are $Z_{j 1}=\hat{\boldsymbol{e}}_{1}^{T} \boldsymbol{x}_{j}, \ldots, Z_{j p}=\hat{\boldsymbol{e}}_{p}^{T} \boldsymbol{x}_{j}$. Let the vector $\boldsymbol{z}_{j}=\left(Z_{j 1}, \ldots, Z_{j p}\right)^{T}$. The proportion of the trace explained by the first $k$ th principal components is $\sum_{i=1}^{k} \hat{\lambda}_{i} / \sum_{j=1}^{p} \hat{\lambda}_{j}=\sum_{i=1}^{k} \hat{\lambda}_{i} / \operatorname{tr}(\hat{\boldsymbol{\Sigma}})$. When a correlation or covariance matrix is being estimated, "trace" is replaced by "variance." The population analogs use the dispersion matrix $\boldsymbol{\Sigma}$ with eigenvalue eigenvector pairs ( $\lambda_{i}, \boldsymbol{e}_{i}$ )
for $i=1, \ldots, p$. The population principal components corresponding to the $j t h$ case are $Y_{j i}=\boldsymbol{e}_{i}^{T} \boldsymbol{x}_{j}$, and $Z_{j i}=\hat{Y}_{j i}$ for $i=1, \ldots, p$.

Note that the principal components can be collected into an $n \times p$ data matrix

$$
\boldsymbol{Z}=\left[\begin{array}{cccc}
Z_{1,1} & Z_{1,2} & \ldots & Z_{1, p} \\
Z_{2,1} & Z_{2,2} & \ldots & Z_{2, p} \\
\vdots & \vdots & \ddots & \vdots \\
Z_{n, 1} & Z_{n, 2} & \ldots & Z_{n, p}
\end{array}\right]=\left[\begin{array}{llll} 
& \boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \ldots \\
\boldsymbol{u}_{p}
\end{array}\right]=\left[\begin{array}{c}
\boldsymbol{z}_{1}^{T} \\
\vdots \\
\boldsymbol{z}_{n}^{T}
\end{array}\right]
$$

Then $\boldsymbol{u}_{i}$ corresponds to the $i$ th principal component.
The data matrix $\boldsymbol{W}$ corresponds to the usual axes where $\boldsymbol{e}_{i}$ is a vector of zeroes except for a one in the $i$ th position. Hence the $i$ th axis corresponds to the $i$ th variable $X_{i}$. The data matrix $\boldsymbol{Z}$ corresponds to axes that are parallel to the axes of the hyperellipsoid corresponding to the dispersion matrix $\hat{\boldsymbol{\Sigma}}$. These axes are a rotation of the usual axes about the origin.

If $\hat{\boldsymbol{\Sigma}}=\boldsymbol{S}$, then the definition of the estimated proportion of the total population variance may make little sense if the variables are measured on different scales. Assume the population covariance matrix is $I_{2}$. Then $\lambda_{j} /\left(\lambda_{1}+\lambda_{2}\right)=0.5$, but if $x_{j}$ is multiplied by 3 then $V\left(x_{j}\right)=9=\lambda_{j}$, and $\lambda_{j} /\left(\lambda_{1}+\lambda_{2}\right)=0.9$. Then $x_{j}$ seems much more important than the other variable just by scaling. This is why rule of thumb 1 says $\boldsymbol{R}$ should be used instead of $\boldsymbol{S}$ if $\max _{i=1, \ldots, p} S_{i}^{2} / \min _{i=1, \ldots, p} S_{i}^{2}>2$.

The hyperellipsoid $\left\{\boldsymbol{x} \mid D_{\boldsymbol{x}}^{2} \leq h^{2}\right\}=\left\{\boldsymbol{x}:(\boldsymbol{x}-\boldsymbol{\mu})^{T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq h^{2}\right\}$, where $h^{2}=u_{1-\alpha}$ and $P\left(D_{\boldsymbol{x}_{i}}^{2} \leq u_{1-\alpha}\right)=1-\alpha$, is the highest density region covering $1-\alpha$ of the mass for a large class of elliptically contoured distributions. The hyperellipsoid is centered at $\boldsymbol{\mu}$. If $\boldsymbol{\mu}=\mathbf{0}$, then points at squared distance $\boldsymbol{w}^{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{w}=h^{2}$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\boldsymbol{e}_{i}$ where the half length in the direction of $\boldsymbol{e}_{i}$ is $h \sqrt{\lambda_{i}}$.

The projection vector of a vector $\boldsymbol{x}$ onto a vector $\boldsymbol{e}$ is

$$
\frac{e e^{T} \boldsymbol{x}}{\boldsymbol{e}^{T} \boldsymbol{e}}
$$

Hence if $\boldsymbol{e}^{T} \boldsymbol{e}=1$, the projection vector is $\boldsymbol{v}=\left[\boldsymbol{e}^{T} \boldsymbol{x}\right] \boldsymbol{e}$ and $\|\boldsymbol{v}\|=\left|\boldsymbol{e}^{T} \boldsymbol{x}\right|$. So $\boldsymbol{e}^{T} \boldsymbol{x}$ is the signed length of the projection vector of $\boldsymbol{x}$ onto $\boldsymbol{e}$, and $\boldsymbol{e}^{T} \boldsymbol{x}$ is called the (scalar) projection of $\boldsymbol{x}$ onto $\boldsymbol{e}$.

The $\boldsymbol{e}_{i}$ are the directions of the axes through the origin that are parallel to the axes of the hyperellipsoid. Suppose $\boldsymbol{\mu}=\mathbf{0}$. Then the $i$ th principle component is the linear combination of the predictors that is the projection on the $i$ th axis of the hyperellipsoid. That is, get the projection vectors of the $\boldsymbol{x}_{i}$ onto $\boldsymbol{e}_{i}$ and find their signed lengths $\boldsymbol{e}_{i}^{T} \boldsymbol{x}_{i}$ from the origin. Then these scalars form the $i$ th principal components corresponding to the $n$ data cases $\boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{n}$. So the first principal component is the projection on the major axis, the second principal component is the projection on the next longest axis, ..., the $p$ th principal component is the projection on the minor axis. The axes are orthogonal, so the directions $\boldsymbol{e}_{i}$ are orthogonal. When $\boldsymbol{\mu} \neq \mathbf{0}$ the projections on $\boldsymbol{e}_{i}$ are projections on the axes through the origin that are parallel to the axes of the hyperellipsoid.

The first $k$ principal components can be regarded as a good $k$ dimensional approximation to the $p$ dimensional data. Suppose the data cloud approximates the hyperellipsoid $\left\{\boldsymbol{x} \mid D_{\boldsymbol{x}}^{2} \leq h^{2}\right\}$ where $h^{2}=D_{(n)}^{2}$, the largest squared distance, so the hyperellipsoid contains all of the data. Then a good one dimensional approximation is the projection on the major axis since this captures the dimension with the greatest variability or dispersion as measured by $\boldsymbol{\Sigma}$. A good two dimensional approximation uses the projection on the major axis and the projection on the next largest axis since these are the two orthogonal directions where the two projections have the greatest variability. Following Mardia, Kent and Bibby (1979, p. 220), if $\boldsymbol{S}$ (with centered data) or $\boldsymbol{R}$ is used as the dispersion matrix, then the vector space spanned by the first $k$ principal components has smaller mean square deviation from the $p$ variables than any other $k$-dimensional subspace.

Since $\boldsymbol{Z}$ represents a new coordinate system, the $i$ th case $\boldsymbol{x}_{i}=\left(\boldsymbol{x}_{i}^{T} \hat{\boldsymbol{e}}_{i}\right) \hat{\boldsymbol{e}}_{1}+\cdots+$ $\left(\boldsymbol{x}_{i}^{T} \hat{\boldsymbol{e}}_{p}\right) \hat{\boldsymbol{e}}_{p}=Z_{i, 1} \hat{\boldsymbol{e}}_{1}+\cdots+Z_{i, p} \hat{\boldsymbol{e}}_{p}$. Also $\boldsymbol{x}_{i}=\tilde{\boldsymbol{x}}_{i}(k)+\boldsymbol{r}_{i}(k)$ where $\tilde{\boldsymbol{x}}_{i}(k)=\sum_{j=1}^{k} Z_{i, j} \hat{\boldsymbol{e}}_{j}$ and the residual vector $\boldsymbol{r}_{i}(k)=\sum_{j=k+1}^{p} Z_{i, j} \hat{\boldsymbol{e}}_{j}$. The squared length of the residual vector is $\left\|\boldsymbol{r}_{i}(k)\right\|^{2}=\boldsymbol{r}_{i}(k)^{T} \boldsymbol{r}_{i}(k)=Z_{i, k+1}^{2}+\cdots+Z_{i, p}^{2}$.

Suppose $\boldsymbol{S}$ or $\boldsymbol{R}$ is used as the as the dispersion matrix and that $T=\mathbf{0}$ so the hyperellisoid is centered at the origin. The eigenvector corresponding to the largest eigenvalue determines the major axis of the hyperellipsoid. This axis forms the line through the origin such that the sum of squared distances from the $n$ data points $\boldsymbol{x}_{i}$ to this line is a minimum. If the data points are projected onto a hyperplane perpendicular to the major axis line, then the eigenvector corresponding to the next largest eigenvalue determines the second longest axis of the hyperellipsoid, and this axis is the line through the origin in the hyperplane that minimizes the sum of squared distances, and so on.

When the covariance matrix is used, that the first principal component $\boldsymbol{e}_{1}^{T} \boldsymbol{x}$ is the linear combination $\boldsymbol{g}_{1}^{T} \boldsymbol{x}$ that maximizes $\operatorname{Var}\left(\boldsymbol{g}_{1}^{T} \boldsymbol{x}\right)$ subject to $\boldsymbol{g}_{1}^{T} \boldsymbol{g}_{1}=1$, while the $j$ th principal component is the linear combination $\boldsymbol{g}_{j}^{T} \boldsymbol{x}$ that maximizes $\operatorname{Var}\left(\boldsymbol{g}_{j}^{T} \boldsymbol{x}\right)$ subject to $\boldsymbol{g}_{j}^{T} \boldsymbol{g}_{j}=1$ and $\operatorname{Cov}\left(\boldsymbol{g}_{j}^{T} \boldsymbol{x}, \boldsymbol{g}_{k}^{T} \boldsymbol{x}\right)=0$ for $k<j$.

Dimension reduction involves using the first $k$ principal components to approximate the data matrix without losing much important information. Want the proportion of the trace explained by the first $k$ principal components to be higher than 0.8 or 0.9 .

Rule of thumb 2. The value of $k$ should be such that

$$
\frac{\sum_{i=1}^{k} \hat{\lambda}_{i}}{\sum_{i=1}^{p} \hat{\lambda}_{i}} \geq 0.9
$$

The scree plot of component number versus eigenvalue is also useful for choosing $k$ since often there is a sharp bend in the scree plot when the components are no longer important. See Cattell (1966).

Following Johnson and Wichern (1988, p. 343, 347), let $\boldsymbol{x}=\left(X_{1}, \ldots, X_{p}\right)$ be a random vector such that the $\boldsymbol{x}_{i}$ and $\boldsymbol{x}$ have the same distribution. Let $Y_{i}=\boldsymbol{e}_{i}^{T} \boldsymbol{x}$ be the population
principal components based on the covariance matrix $\operatorname{Cov}(\boldsymbol{x})=\boldsymbol{\Sigma} \boldsymbol{x}$. Let $\boldsymbol{e}_{i}=\left(e_{1 i}, \ldots, e_{p i}\right)^{T}$. Then $e_{k i}$ is proportional to the correlation between $Y_{i}$ and $X_{k}$, in fact,

$$
\operatorname{corr}\left(Y_{i}, X_{k}\right)=\frac{e_{k i} \sqrt{\lambda_{i}}}{\sqrt{\sigma_{k k}}}
$$

for $i, k=1, \ldots, p$. If the correlation matrix $\boldsymbol{\rho}$ is used instead of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $\operatorname{corr}\left(Y_{i}, X_{k}\right)=$ $e_{k i} \sqrt{\lambda_{i}}$.

Following Johnson and Wichern (1988, p. 252-253), some software that uses $\boldsymbol{S}$ or $\boldsymbol{R}$ centers the data by using $\boldsymbol{x}_{i}-\overline{\boldsymbol{x}}$. Centering does not change $\boldsymbol{S}$ or $\boldsymbol{R}$ but makes the $i$ th principal component equal to $\hat{\boldsymbol{e}}_{i}^{T}(\boldsymbol{x}-\overline{\boldsymbol{x}})$ for observation $\boldsymbol{x}$.

Warning: If $\hat{\lambda}_{p} \approx 0$, then $\hat{\boldsymbol{\Sigma}}$ is nearly singular, and there could be an unnoticed linear dependency in the data set, e.g. $X_{p} \approx \sum_{i=1}^{p-1} c_{i} X_{i}$. Then one or more of the variables is redundant and should be deleted. Following Johnson and Wichern (1988, p. 360), suppose $p=4$ and $X_{1}, X_{2}$ and $X_{3}$ are midterm exam scores while $X_{4}$ is the total of the midterm scores so that $X_{4}=X_{1}+X_{2}+X_{3}$. Due to rounding, $\hat{\lambda}_{4}$ could be nonzero, but very close to zero.

## CHAPTER 1

## ROBUST PRINCIPAL COMPONENT ANALYSIS

A robust "plug in" method uses an analysis based on the ( $\hat{\lambda}_{i}, \hat{e}_{i}$ ) computed from a robust dispersion estimator $\boldsymbol{C}$. The RPCA method performs the classical principal component analysis on the RMVN subset $U$ of cases that are given weight 1, using either the sample covariance matrix $\boldsymbol{C}_{U}=\boldsymbol{S}_{U}$ or the sample correlation matrix $\boldsymbol{R}_{U}$.

The following assumption (E1) gives a class of distributions where the Olive and Hawkins (2010) FCH, RFCH and RMVN robust estimators can be proven to be $\sqrt{n}$ consistent. Cator and Lopuhaä $(2010,2012)$ show that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution.

Assumption (E1): The $\boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{n}$ are iid from a "unimodal" $E C_{p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\operatorname{Cov}\left(\boldsymbol{x}_{i}\right)$ where $g$ is continuously differentiable with finite 4th moment: $\int\left(\boldsymbol{x}^{T} \boldsymbol{x}\right)^{2} g\left(\boldsymbol{x}^{T} \boldsymbol{x}\right) d \boldsymbol{x}<\infty$.

Under assumption (E1), $\boldsymbol{C}_{U}$ and $\boldsymbol{R}_{U}$ are $\sqrt{n}$ consistent highly outlier resistant estimators of $c \boldsymbol{\Sigma}=d \operatorname{Cov}(\boldsymbol{x})$ and the population correlation matrix $\boldsymbol{D} \operatorname{Cov}(\boldsymbol{x}) \boldsymbol{D}=\boldsymbol{\rho}$, respectively, where $\boldsymbol{D}=\operatorname{diag}\left(1 / \sqrt{\sigma}{ }_{11}, \ldots, 1 / \sqrt{\sigma}_{\mathrm{pp}}\right)$ and the $\sigma_{i i}$ are the diagonal entries of $\operatorname{Cov}(\boldsymbol{x})=\boldsymbol{\Sigma} \boldsymbol{x}=c_{X} \boldsymbol{\Sigma}$. Let $\lambda_{i}(\boldsymbol{A})$ be the eigenvalues of $\boldsymbol{A}$ where $\lambda_{1}(\boldsymbol{A}) \geq \lambda_{2}(\boldsymbol{A}) \geq \cdots \geq$ $\lambda_{p}(\boldsymbol{A})$. Let $\hat{\lambda}_{i}(\hat{\boldsymbol{A}})$ be the eigenvalues of $\hat{\boldsymbol{A}}$ where $\hat{\lambda}_{1}(\hat{\boldsymbol{A}}) \geq \hat{\lambda}_{2}(\hat{\boldsymbol{A}}) \geq \cdots \geq \hat{\lambda}(\hat{\boldsymbol{A}})$.

Theorem 3. Under (E1), the correlation of the eigenvalues computed from the classical PCA and RPCA converges to 1 in probability.

Proof: The eigenvalues are continuous functions of the dispersion estimator, hence consistent estimators of dispersion give consistent estimators of the population eigenvalues. See Eaton and Tyler (1991) and Bhatia, Elsner and Krause (1990). Let $\lambda_{i}(\boldsymbol{\Sigma})=\lambda_{i}$ be the eigenvalues of $\boldsymbol{\Sigma}$ so $c_{X} \lambda_{i}$ are the eigenvalues of $\operatorname{Cov}(\boldsymbol{x})=\boldsymbol{\Sigma} \boldsymbol{x}$. Under $(\mathrm{E} 1), \lambda_{i}(\boldsymbol{S}) \xrightarrow{P} c_{X} \lambda_{i}$
and $\lambda_{i}\left(\boldsymbol{C}_{U}\right) \xrightarrow{P} c \lambda_{i}=\frac{c}{c_{X}} c_{X} \lambda_{i}=d c_{X} \lambda_{i}$. Hence the population eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $d \boldsymbol{\Sigma} \boldsymbol{x}$ differ by the positive multiple $d$, and the population correlation of the two sets of eigenvalues is equal to one.

Now let $\lambda_{i}(\boldsymbol{\rho})=\lambda_{i}$. Under (E1), both $\boldsymbol{R}$ and $\boldsymbol{R}_{U}$ converge to $\boldsymbol{\rho}$ in probability, so $\hat{\lambda}_{i}(\boldsymbol{R}) \xrightarrow{P} \lambda_{i}$ and $\hat{\lambda}_{i}\left(\boldsymbol{R}_{U}\right) \xrightarrow{P} \lambda_{i}$ for $i=1, \ldots, p$. Hence the two population sets of eigenvalues are the same and thus have population correlation equal to one. QED

Note that if $\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{e}=\lambda \boldsymbol{e}$, then

$$
d \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{e}=d \lambda \boldsymbol{e}
$$

Thus $\hat{\lambda}_{i}(\boldsymbol{S}) \xrightarrow{P} \lambda_{i}\left(\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)$ and $\hat{\lambda}_{i}\left(\boldsymbol{C}_{U}\right) \xrightarrow{P} d \lambda_{i}\left(\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)$ for $i=1, \ldots, p$. Since plotting software fills space, two scree plots of two sets of eigenvalues that differ by a constant positive multiple will look nearly the same, except for the labels of the vertical axis, and the "trace explained" by the largest $k$ eigenvalues will be the same for the two sets of eigenvalues. Theorem 2 implies that for a large class of elliptically contoured distributions and for large $n$, the classical and robust scree plots should be similar visually, and the "trace explained" by the classical PCA and the robust PCA should also be similar.

The eigenvectors are not continuous functions of the dispersion estimator, and the sample size may need to be massive before the robust and classical eigenvectors or principal components have high absolute correlation. In the software, sign changes in the eigenvectors are common, since $\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{e}=\lambda \boldsymbol{e}$ implies that $\boldsymbol{\Sigma}_{\boldsymbol{x}}(-\boldsymbol{e})=\lambda(-\boldsymbol{e})$.

## CHAPTER 2

## EXAMPLES AND SIMULATIONS

Table 2.1. Estimation of $\boldsymbol{\Sigma}$ with $\gamma=0.4, n=35 p$

| p | type | $n$ | $p m$ | Q |
| :---: | :---: | :---: | :---: | :---: |
| 5 | 1 | 135 | 16 | 0.153 |
| 5 | 2 | 135 | 6 | 0.213 |
| 10 | 1 | 350 | 21 | 0.326 |
| 10 | 2 | 350 | 6 | 0.326 |
| 15 | 1 | 525 | 26 | 0.856 |
| 15 | 2 | 525 | 7 | 0.675 |
| 20 | 1 | 700 | 33 | 0.798 |
| 20 | 2 | 700 | 8 | 0.792 |
| 25 | 1 | 875 | 39 | 1.014 |
| 25 | 2 | 875 | 10 | 1.867 |

The robust estimator used was the RMVN estimator of Olive and Hawkins (2010) and Zhang, Olive and Ye (2012). This estimator was shown to be $\sqrt{n}$ consistent and highly outlier resistant for a large class of elliptically contoured distributions.

A simulation was done to check that RMVN estimates $\boldsymbol{\Sigma}$ if the clean data is MVN and $\gamma$ is the percentage of outliers. The clean cases were multivariate normal (MVN): $\boldsymbol{x} \sim N_{p}(\mathbf{0}, \operatorname{diag}(1,2, \ldots, p))$. Outlier types were $\boldsymbol{x} \sim N_{p}\left((0, \ldots, 0, p m)^{T}, 0.0001 \boldsymbol{I}_{p}\right)$, a near point mass at the major axis, and the mean shift $\boldsymbol{x} \sim N_{p}(p m 1, \operatorname{diag}(1,2, \ldots, p))$ where $\mathbf{1}=(1, \ldots, 1)^{T}$. On clean MVN data, $n \geq 20 p$ gave good results for $2 \leq p \leq 100$. For the
contaminated MVN data, the first $n \gamma$ cases were outliers, and the classical estimator $\boldsymbol{S}_{c}$ was computed on the clean cases. The diagonal elements of $\boldsymbol{S}_{c}$ and $\hat{\boldsymbol{\Sigma}}_{R M V N}$ should both be estimating $(1,2, \ldots, p)^{T}$. The average diagonal elements of both matrices were computed for 20 runs, and the criterion $Q$ was the sum of the absolute differences of the $p$ diagonal elements from the two averaged matrices. Since $\gamma=0.4$ and the initial subsets for the RMVN estimator are half sets, the simulations used $n=35 p$. The values of $Q$ shown in Table 2.1 correspond to good estimation of the diagonal elements. Values of $p m$ slightly smaller than the tabled values led to poor estimation of the diagonal elements.


Figure 2.1. First Two Principal Components for Buxton data.

Example 1. Buxton (1920) gives various measurements on 87 men including height, head length, nasal height, bigonal breadth and cephalic index. Five heights were recorded to be about 19 mm with the true heights recorded under head length. Performing a classical principal components analysis on these five variables using the covariance matrix resulted in a first principal component corresponding to a major axis that passed through the outliers. See Figure 2.1 where the second principal component is plotted versus the first. The robust PCA, or the classical PCA performed after the outliers are removed, resulted in a first principal component that was approximately - height with


Figure 2.2. First Two Robust Principal Components with Outliers Omitted.
$\hat{\boldsymbol{e}}_{1} \approx(-1.000,0.002,-0.023,-0.002,-0.009)^{T}$ while the second robust principal component was based on the eigenvector $\hat{\boldsymbol{e}}_{2} \approx(-0.005,0.848,-0.054,-0.048,0.525)^{T}$. The plot of the first two robust principal components, with the outliers deleted, is shown in Figure 2.2. These two components explain about $86 \%$ of the variance.

The $R$ function prcomp can be used to compute output. Suppose the data matrix is $z$. The commands

```
zz <- prcomp(z)
```

zz
will create and display output. The term $z z \$ s d$ gives the square roots of the eigenvalues while the term $z z \$$ rot displays the eigenvectors using the covariance matrix. Hence Figure 2.1 can be made with the following commands.
z <- cbind (buxy, buxx)
zz <- prcomp(z)
PC1 <- z\% $\% \% z z \$ r o t[, 1]$
PC2 <- $\mathrm{z} \% * \% z z \$ r o t[, 2]$


Figure 2.3. Robust Scree Plot.

It usually makes more sense to use the correlation matrix. The mpack function rprcomp does robust principal components. The two functions use "scale=T" or "cor=T" to use a correlation matrix.

```
zzcor <- prcomp(z,scale=T)
zrcor <- rprcomp(z,cor=T)
```

Then

```
zrcor$out$sd^2
```

gives the eigenvalues and zrcor\$out $\$$ rot gives the eigenvectors. Scree plots can be made with the following commands, and Figure 2.3 shows the robust scree plot which suggests that the last principal component can be deleted.

```
EIG <- zzcor$sd^2
plot(EIG)
#robust scree plot
```

```
REIG <- zrcor$out$sd^2
plot(REIG)
```

The outliers are known from the DD plot so the robust principal component analysis can be done with and without the outliers. The data matrix $z w$ is the clean data without the outliers.

```
zw <-z[-c(61,62,63,64,65),]
zzcorc <- prcomp(zw,scale=T)
# clean data with corr matrix
> zzcorc
Standard deviations:
```

[1] 1.31843581 .17239911 .01552660 .78673490 .4867867
Rotation:

|  | PC1 | PC2 | PC3 | PC4 | PC5 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| buxy | 0.01551 | 0.71466 | 0.02247 | -0.68890 | -0.11806 |
| len | 0.70308 | -0.06778 | 0.07744 | -0.16901 | 0.68302 |
| nasal | 0.15038 | 0.68868 | 0.02042 | 0.70385 | 0.08539 |
| bigonal | 0.11646 | -0.04882 | 0.96504 | 0.02261 | -0.22855 |
| cephalic | -0.68502 | 0.08950 | 0.24854 | -0.03071 | 0.67825 |
| zrcor <- rprcomp (z,cor=T) |  |  |  |  |  |
| $>$ zrcor |  |  |  |  |  |
| \$out |  |  |  |  |  |

Standard deviations:
[1] 1.33234001 .15488790 .99886430 .81827410 .4730769
Rotation:

|  | PC1 | PC2 | PC3 | PC4 | PC5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| buxy | -0.10724 | -0.69431 | -0.11325 | 0.69184 | -0.12238 |

```
len }0.609909-0.06324 0.02560 0.17129 0.69085
nasal 0.04094 -0.70310-0.08718 -0.70093 0.07123
bigonal 0.02638-0.13994 0.98660 0.01120-0.07884
cephalic -0.70527-0.00317 0.07443 0.02432 0.70460
> zrcorc <- rprcomp(zw,cor=T)
> zrcorc
$out
Standard deviations:
[1] 1.3369152 1.1466891 1.0016463 0.8123854 0.4842482
Rotation:
\begin{tabular}{lcllll} 
& PC1 & PC2 & PC3 & PC4 & PC5 \\
buxy & -0.21306 & 0.67557 & -0.01727 & -0.68852 & -0.15446 \\
len & 0.67272 & 0.21639 & 0.05560 & -0.15178 & 0.68884 \\
nasal & -0.22213 & 0.66958 & 0.05174 & 0.68978 & 0.15441 \\
bigonal & -0.01374 & -0.02995 & 0.99668 & -0.03546 & -0.06543 \\
cephalic & -0.67270 & -0.21807 & 0.02363 & -0.16076 & 0.68813
\end{tabular}
```

Note that the square roots of the eigenvalues, given by "Standard deviations," do not change much for the following three estimators: the classical estimator applied to the clean data, and the robust estimator applied to the full data or the clean data. The first eigenvector is roughly proportional to length - cephalic while the second eigenvector is roughly proportional to buxy + nasal. The third principal component is highly correlated with bigonal, the fourth principal component is proportional to buxy - nasal, and the fifth principal component to length + cephalic.

Consider several estimators described in Olive and Hawkins (2010). In simulations for principal component analysis, FCH, RMVN, OGK and Fake-MCD seem to estimate $c \boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x}=\boldsymbol{A} \boldsymbol{z}+\boldsymbol{\mu}$ where $\boldsymbol{z}=\left(z_{1}, \ldots, z_{p}\right)^{T}$ and the $z_{i}$ are iid from a continuous distribution
with variance $\sigma^{2}$. Here $\boldsymbol{\Sigma} \boldsymbol{x}=\operatorname{Cov}(\boldsymbol{x})=\sigma^{2} \boldsymbol{A} \boldsymbol{A}^{T}$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c \boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of $z_{i}$ is also symmetric. DGK and Fake-MCD (with fixed random number seed) are affine equivariant. FCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

The simulations used 1000 runs where $\boldsymbol{x}=\boldsymbol{A} \boldsymbol{z}$ and $\boldsymbol{z} \sim N_{p}\left(\mathbf{0}, \boldsymbol{I}_{p}\right), \boldsymbol{z} \sim L N\left(\mathbf{0}, \boldsymbol{I}_{p}\right)$ where the marginals are iid lognormal( 0,1 ), or $\boldsymbol{z} \sim M V T_{p}(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy $(0,1)$. The choice $\boldsymbol{A}=\operatorname{diag}(\sqrt{1}, \ldots, \sqrt{p})$ results in $\boldsymbol{\Sigma}=\operatorname{diag}(1, \ldots, p)$. Note that the population eigenvalues will be proportional to $(p, p-1, \ldots, 1)^{T}$ and the population "variance explained" by the $i$ th principal component is $\lambda_{i} / \sum_{j=1}^{p} \lambda_{j}=2(p+1-i) /[p(p+1)]$. For $p=4$, these numbers are $0.4,0.3$ and 0.2 for the first three principal components. If the "correlation" option is used, then the population "correlation matrix" is the identity matrix $\boldsymbol{I}_{p}$, the $i$ th population eigenvalue is proportional to $1 / p$ and the population "variance explained" by the $i$ th principal component is $1 / p$.

Table 2.2 shows the mean "variance explained" (M) along with the standard deviations (S) for the first three principal components. Also $a_{i}$ and $p_{i}$ are the average absolute value of the correlation between the $i$ th eigenvectors or the $i$ th principal components of the classical and robust methods. Two rows were used for each " $n$-data type" combination. The $a_{i}$ are shown in the top row while the $p_{i}$ are in the lower row. The values of $a_{i}$ and $p_{i}$ were similar. The standard deviations were slightly smaller for the classical PCA for normal data. The classical method failed to estimate $(0.4,0.3,0.2)$ for the Cauchy data. For the lognormal data, RPCA gave better estimates, and the $p_{i}$ were not high except for $n=10000$.

To compare affine equivariant and non-equivariant estimators, Maronna and Zamar (2002) suggest using $\boldsymbol{A}_{i, i}=1$ and $\boldsymbol{A}_{i, j}=\rho$ for $i \neq j$ and $\rho=0,0.5,0.7,0.9$, and 0.99. Then $\boldsymbol{\Sigma}=\boldsymbol{A}^{2}$. If $\rho$ is high, or if $p$ is high and $\rho \geq 0.5$, then the data are concentrated about the line with direction $1=(1, \ldots, 1)^{T}$. For $p=50$ and $\rho=0.99$, the population variance

Table 2.2. Variance Explained by PCA and RPCA, $p=4$

| n | type | M/S | vexpl | rvexpl | $a_{1} / p_{1}$ | $a_{2} / p_{2}$ | $a_{3} / p_{3}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 40 | N | M | 0.445,0.289,0.178 | 0.472,0.286,0.166 | 0.895 | 0.821 | 0.825 |
|  |  | S | 0.050,0.037,0.032 | 0.062,0.043,0.037 | 0.912 | 0.813 | 0.804 |
| 100 | N | M | $0.419,0.295,0.191$ | 0.425,0.293,0.189 | 0.952 | 0.926 | 0.963 |
|  |  | S | 0.033,0.030,0.024 | 0.040,0.032,0.027 | 0.956 | 0.923 | 0.953 |
| 200 | N | M | 0.410,0.296,0.196 | 0.410,0.296,0.196 | 0.988 | 0.978 | 0.979 |
|  |  | S | 0.024, $0.024,0.017$ | 0.027,0.024,0.019 | 0.991 | 0.973 | 0.980 |
| 400 | N | M | 0.404,0.298,0.198 | 0.406,0.298,0.198 | 0.994 | 0.991 | 0.996 |
|  |  | S | 0.019,0.017,0.014 | 0.021,0.019,0.015 | 0.995 | 0.990 | 0.994 |
| 1000 | N | M | 0.399,0.301,0.199 | 0.399,0.300,0.199 | 0.998 | 0.998 | 0.999 |
|  |  | S | 0.013,0.010,0.009 | 0.014,0.011,0.010 | 0.999 | 0.997 | 0.998 |
| 40 | C | M | 0.765,0.159,0.056 | $0.514,0.275,0.147$ | 0.563 | 0.519 | 0.511 |
|  |  | S | 0.165,0.112,0.051 | 0.078,0.055,0.040 | 0.776 | 0.383 | 0.239 |
| 100 | C | M | 0.762,0.156,0.060 | 0.455,0.286,0.173 | 0.585 | 0.527 | 0.528 |
|  |  | S | 0.173,0.112,0.055 | 0.054,0.041,0.034 | 0.797 | 0.377 | 0.269 |
| 200 | C | M | 0.743,0.172,0.062 | 0.432,0.290, 0.184 | 0.620 | 0.555 | 0.580 |
|  |  | S | 0.185,0.125,0.055 | 0.042,0.0313,0.029 | 0.800 | 0.445 | 0.300 |
| 400 | C | M | 0.756,0.162,0.060 | 0.413,0.296,0.194 | 0.608 | 0.562 | 0.575 |
|  |  | S | 0.172,0.113,0.054 | 0.030,0.025,0.022 | 0.796 | 0.397 | 0.308 |
| 1000 | C | M | 0.751,0.168,0.058 | 0.408,0.297,0.196 | 0.629 | 0.563 | 0.582 |
|  |  | S | 0.159,0.107,0.047 | 0.023,0.019,0.015 | 0.811 | 0.437 | 0.325 |
| 40 | L | M | 0.539,0.256,0.139 | $0.521,0.268,0.146$ | 0.610 | 0.509 | 0.530 |
|  |  | S | 0.127,0.075,0.054 | 0.099,0.061,0.047 | 0.643 | 0.439 | 0.398 |
| 100 | L | M | 0.482,0.270,0.165 | 0.459,0.279,0.172 | 0.647 | 0.555 | 0.566 |
|  |  | S | 0.180,0.063, 0.052 | 0.077,0.047,0.041 | 0.654 | 0.492 | 0.474 |
| 200 | L | M | 0.463,0.272,0.173 | 0.436,0.285, 0.182 | 0.668 | 0.544 | 0.633 |
|  |  | S | 0.110,0.059,0.054 | 0.056,0.041, 0.034 | 0.642 | 0.519 | 0.565 |
| 400 | L | M | 0.437,0.282,0.185 | 0.416,0.290,0.194 | 0.748 | 0.639 | 0.739 |
|  |  | S | 0.080,0.048, 0.044 | 0.049,0.035,0.033 | 0.727 | 0.594 | 0.690 |
| 1000 | L | M | 0.423,0.289,0.188 | 0.425,0.293,0.189 | 0.871 | 0.797 | 0.928 |
|  |  | S | 0.073,0.042,0.039 | 0.032,0.024,0.025 | 0.837 | 0.778 | 0.893 |
| 10000 | L | M | 0.400, $0.301,0.200$ | 0.403,0.293,0.204 | 0.982 | 0.967 | 0.991 |
|  |  | S | 0.027,0.023,0.018 | 0.013,0.011,0.009 | 0.976 | 0.967 | 0.989 |

explained by the first principal component is 0.999998 . If the "correlation" option is used, then there is still one extremely dominant principal component unless both $p$ and $\rho$ are small.

Table 2.3. Variance Explained by PCA and RPCA, SSD $=10^{7} \mathrm{SD}, p=50$

| n | type | vexpl | SSD | rvexpl | SSD | $a_{1}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 200 | N | 0.999998 | 1.958 | 0.999998 | 2.867 | 0.687 |
| 400 | N | 0.999981 | 1.600 | 0.999981 | 1.632 | 0.883 |
| 800 | N | 0.999981 | 1.214 | 0.999981 | 1.275 | 0.872 |
| 1000 | N | 0.999998 | 0.917 | 0.999998 | 0.971 | 0.944 |
| 200 | C | 0.999954 | 109.3 | 0.999981 | 4.352 | 0.460 |
| 400 | C | 0.999913 | 601.4 | 0.999981 | 2.716 | 0.450 |
| 800 | C | 0.999974 | 363.6 | 0.999981 | 2.058 | 0.435 |
| 1000 | C | 0.999996 | 161.3 | 0.999998 | 1.482 | 0.112 |
| 200 | L | 0.999982 | 2.024 | 0.999979 | 3.292 | 0.486 |
| 400 | L | 0.999981 | 2.047 | 0.999979 | 2.134 | 0.506 |
| 800 | L | 0.999981 | 1.131 | 0.999979 | 1.657 | 0.468 |
| 1000 | L | 0.999998 | 0.919 | 0.999998 | 1.508 | 0.175 |

Table 2.3 shows the mean "variance explained" along with the standard deviations multiplied by $10^{7}$ for the first principal component. The $a_{1}$ value is given but $p_{1}$ was always 1.0 to many decimal places even with Cauchy data. Hence the eigenvectors from the robust and classical methods could have low absolute correlation, but the data was so tightly clustered that the first principal components from the robust and classical methods
had absolute correlation near 1 .

## CHAPTER 3

## CONCLUSIONS

Jolliffe (2010) is an authoritative text on PCA. Cattell (1966) and Bentler and Yuan (1998) are good references for scree plots. M $\phi$ ller, von Frese and Bro (2005) discuss PCA, principal component regression and drawbacks of $M$ estimators. Waternaux (1976) gives some large sample theory for PCA. In particular, if the $\boldsymbol{x}_{i}$ are iid from a multivariate distribution with fourth moments and a covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ such that the eigenvalues are distinct and positive, then $\sqrt{n}\left(\hat{\lambda}_{i}-\lambda_{i}\right) \xrightarrow{D} N\left(0, \kappa_{i}+2 \lambda_{i}^{2}\right)$ where $\kappa_{i}$ is the kurtosis of the marginal distribution of $x_{i}$, for $i=1, \ldots, p$.

The literature for robust PCA is large, but the "high breakdown" methods are impractical or not backed by theory. Some of these methods may be useful as outlier diagnostics. The theory of Boente (1987) for mildly outlier resistant principal components is not based on DGK estimators since the weighting function on the $D_{i}$ is continuous. Spherical principal components is a mildly outlier resistant bounded influence approach suggested by Locantore, Marron, Simpson, Tripoli, Zhang and Cohen (1999). Boente and Fraiman (1999) claim that basis of the eigenvectors is consistently estimated by spherical principal components for elliptically contoured distributions. Also see Maronna, Martin and Yohai (2006, p. 212-213) and Taskinen, Koch and Oja (2012).

Simulations were done in $R$. The MASS library was used to compute FMCD and the robustbase library was used to compute OGK. The mpack function covrmvn computes the FCH, RMVN and MB estimators while covfch computes the FCH, RFCH and MB estimators. The following functions were used in the three simulations and have more outlier configurations than the two described in the simulation. Function covesim was used to produce Table 2.1 and pcasim for Tables 2.2 and 2.3.

For a nonsingular matrix, the inverse of the matrix, the determinant of the matrix and the eigenvalues of the matrix are continuous functions of the matrix. Hence if $\hat{\boldsymbol{\Sigma}}$
is a consistent estimator of $\boldsymbol{\Sigma}$, then the inverse, determinant and eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are consistent estimators of the inverse, determinant and eigenvalues of $\boldsymbol{\Sigma}$. See, for example, Bhatia, Elsner and Krause (1990), Stewart (1969) and Severini (2005, p. 348-349).

## REFERENCES

[1] Bentler, P.M., and Yuan, K.H. (1998), "Tests for Linear Trend in the Smallest Eigenvalues of the Correlation Matrix," Psychometrika, 63, 131-144.
[2] Bhatia, R., Elsner, L., and Krause, G. (1990), "Bounds for the Variation of the Roots of a Polynomial and the Eigenvalues of a Matrix," Linear Algebra and Its Applications, 142, 195-209.
[3] Boente, G. (1987), "Asymptotic Theory for Robust Principal Components," Journal of Multivariate Analysis, 21, 67-78.
[4] Boente, G., and Fraiman, R. (1999), "Discussion of 'Robust Principal Component Analysis for Functional Data' by Locantore et al," Test, 8, 28-35.
[5] Buxton, L.H.D. (1920), "The Anthropology of Cyprus," The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 50, 183-235.
[6] Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," Journal of Multivariate Analysis, 101, 2372-2388.
[7] Cator, E.A., and Lopuhaä, H.P. (2012), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," Bernoulli, 18, 520-551.
[8] Cattell, R.B. (1966), "The Scree Test for the Number of Factors," Multivariate Behavioral Research, 1, 245-276.
[9] Eaton, M.L., and Tyler, D.E. (1991), "On Wielands's Inequality and its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix," The Annals of Statistics, 19, 260-271.
[10] Johnson, R.A., and Wichern, D.W. (1988), Applied Multivariate Statistical Analysis, 2nd ed., Prentice Hall, Englewood Cliffs, NJ.
[11] Jolliffe, I.T. (2010), Principal Component Analysis, 2nd ed., Springer, New York, NY.
[12] 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," (with discussion), Test, 8, 1-73.
[13] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), Multivariate Analysis, Academic Press, London.
[14] Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), Robust Statistics: Theory and Methods, Wiley, Hoboken, NJ.
[15] Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," Technometrics, 44, 307-317.
[16] M $\phi$ ller, S.F., von Frese, J., and Bro, R. (2005), "Robust Methods for Multivariate Data Analysis," Journal of Chemometrics, 19, 549-563.
[17] Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://lagrange.math.siu.edu/Olive/pphbmld.pdf).
[18] Severini, T.A. (2005), Elements of Distribution Theory, Cambridge University Press, New York, NY.
[19] Stewart, G.M. (1969), "On the Continuity of the Generalized Inverse," SIAM Journal on Applied Mathematics, 17, 33-45.
[20] Taskinen, S., Koch, I., and Oja, H. (2012), "Robustifying Principal Component Analysis with Spatial Sign Vectors," Statistics \& Probability Letters, 82, 765-774.
[21] Waternaux, C.M. (1976), "Asymptotic Distribution of the Sample Roots for a Nonnormal Population," Biometrika, 63, 639-645.
[22] Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," International Journal of Statistics and Probability, 1, 119-136.

# VITA 

Graduate School<br>Southern Illinois University

Ayed Rheal Alanzi
Date of Birth: February 05, 1983
Business Administration Department, Majmaah University. P.O Box 66, Majmaah 11952

- Kingdom of Saudi Arabia.
auid1403@hotmail.com

Malaya University
Master of Science, Statistics, August 2009

Research Paper Title:
Robust Principal Component Analysis

Major Professor: Dr. David Olive

