Robust Principal Component Analysis

David J. Olive*

Southern Illinois University

March 25, 2014

Abstract

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.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@siu.edu.

1 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 $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The eigenvalue eigenvector pairs of a matrix $\hat{\boldsymbol{\Sigma}}$ will be $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ 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 $\boldsymbol{\Sigma} = c\ \mathrm{Cov}(\boldsymbol{x})$ for some constant $c > 0$ where $\mathrm{Cov}(\boldsymbol{x})$ is the population covariance matrix.

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

PCA is applied to data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ which are iid from some distribution. If a $p \times 1$ random vector $\boldsymbol{x}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{1}$$

then $\boldsymbol{x}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ 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 $\boldsymbol{x}_i \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, then the sample covariance matrix $\boldsymbol{S}$ estimates $\text{Cov}(\boldsymbol{x}) = c_X \boldsymbol{\Sigma}$.

Theorem 1. Suppose the dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ 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 $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ 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 $\hat{\lambda}_j \left(\hat{\boldsymbol{\rho}}\right) \xrightarrow{P} \lambda_j \left(\boldsymbol{\rho}\right)$ where $\lambda_j(\boldsymbol{A})$ is the jth eigenvalue of $\boldsymbol{A}$ for $j = 1, ..., p$.

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 λ_j is unique, since then the set of eigenvectors corresponding to λ_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 λ_i, then $\boldsymbol{e}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|$ is also an eigenvector corresponding to λ_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 $(\boldsymbol{\Sigma} - \lambda_i)\hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$. 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 $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i - \hat{\lambda}_i \boldsymbol{e}_i \xrightarrow{P} \boldsymbol{0}$.

b) If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i - \lambda_i \hat{\boldsymbol{e}}_i \xrightarrow{P} \boldsymbol{0}$.

If $\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O_P(n^{-\delta})$ where $0 < \delta \leq 0.5$, then

c) $\lambda_i \boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = O_P(n^{-\delta})$, and

d) $\hat{\lambda}_i \hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = O_P(n^{-\delta})$.

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: $|\text{corr}(\hat{\boldsymbol{e}}_j, \boldsymbol{e}_j)| \xrightarrow{P} 1$.

Proof. a) $\hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i - \hat{\lambda}_i \boldsymbol{e}_i \xrightarrow{P} \boldsymbol{\Sigma}\boldsymbol{e}_i - \lambda_i \boldsymbol{e}_i = \boldsymbol{0}$.

b) Note that $(\boldsymbol{\Sigma} - \lambda_i \boldsymbol{I})\hat{\boldsymbol{e}}_i = [(\boldsymbol{\Sigma} - \lambda_i \boldsymbol{I}) - (\hat{\boldsymbol{\Sigma}} - \hat{\lambda}_i \boldsymbol{I})]\hat{\boldsymbol{e}}_i = o_P(1)O_P(1) \xrightarrow{P} \boldsymbol{0}$.

c) $\lambda_i \boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = \boldsymbol{\Sigma}\boldsymbol{e}_i - \hat{\boldsymbol{\Sigma}}\boldsymbol{e}_i = O_P(n^{-\delta})$.

d) $\hat{\lambda}_i \hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{e}}_i - \boldsymbol{\Sigma}\hat{\boldsymbol{e}}_i = O_P(n^{-\delta})$.

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 ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \tag{2}$$

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 $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2_{\boldsymbol{x}} = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$.

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 trace$(\boldsymbol{A}) = tr(\boldsymbol{A}) = \sum_{i=1}^p \boldsymbol{A}_{ii} = \sum_{i=1}^p \lambda_i$. Note that $tr(\text{Cov}(\boldsymbol{x})) =$

$\sigma_1^2 + \cdots + \sigma_p^2$ and $tr(\hat{\boldsymbol{\rho}}) = p$.

Let dispersion estimator $\hat{\boldsymbol{\Sigma}}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then the p *principal components* corresponding to the jth case $\boldsymbol{x}_j$ are $Z_{j1} = \hat{\boldsymbol{e}}_1^T \boldsymbol{x}_j, ..., Z_{jp} = \hat{\boldsymbol{e}}_p^T \boldsymbol{x}_j$. Let the vector $\boldsymbol{z}_j = (Z_{j1}, ..., Z_{jp})^T$. The *proportion of the trace explained* by the first kth principal components is $\sum_{i=1}^k \hat{\lambda}_i / \sum_{j=1}^p \hat{\lambda}_j = \sum_{i=1}^k \hat{\lambda}_i / tr(\hat{\boldsymbol{\Sigma}})$. 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 $i = 1, ..., p$. The population principal components corresponding to the j case are $Y_{ji} = \boldsymbol{e}_i^T \boldsymbol{x}_j$, and $Z_{ji} = \hat{Y}_{ji}$ for $i = 1, ..., p$.

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).

2 Robust Principal Component Analysis

A robust "plug in" method uses an analysis based on the $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ 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 $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Under assumption (E1), $\boldsymbol{C}_U$ and $\boldsymbol{R}_U$ are $\sqrt{n}$ consistent highly outlier resistant estimators of $c\boldsymbol{\Sigma} = d\text{Cov}(\boldsymbol{x})$ and the population correlation matrix $\boldsymbol{D}\text{Cov}(\boldsymbol{x})\boldsymbol{D} = \boldsymbol{\rho}$, respectively, where $\boldsymbol{D} = \text{diag}(1/\sqrt{\sigma}_{11}, ..., 1/\sqrt{\sigma}_{\text{pp}})$ and the σ_{ii} are the diagonal entries of $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = c_X\boldsymbol{\Sigma}$. 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 $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Under (E1), $\lambda_i(\boldsymbol{S}) \xrightarrow{P} c_X\lambda_i$ and $\lambda_i(\boldsymbol{C}_U) \xrightarrow{P} c\lambda_i = \dfrac{c}{c_X} c_X\lambda_i = d\ c_X\ \lambda_i$. 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 $\hat{\lambda}_i(\boldsymbol{R}_U) \xrightarrow{P} \lambda_i$ for $i = 1, ..., p$. 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 $\hat{\lambda}_i(\boldsymbol{S}) \xrightarrow{P} \lambda_i(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\hat{\lambda}_i(\boldsymbol{C}_U) \xrightarrow{P} d\lambda_i(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ for $i = 1, ..., p$. 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 $\boldsymbol{\Sigma}_{\boldsymbol{x}} \ (-\boldsymbol{e}) = \lambda(-\boldsymbol{e})$.

3 Examples and Simulations

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.

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 19mm with the true heights recorded under head length. Performing a classical principal components analysis on these five variables using the covariance matrix

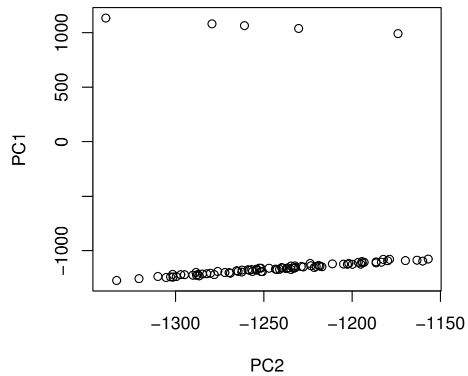

Figure 1: First Two Principal Components for Buxton data

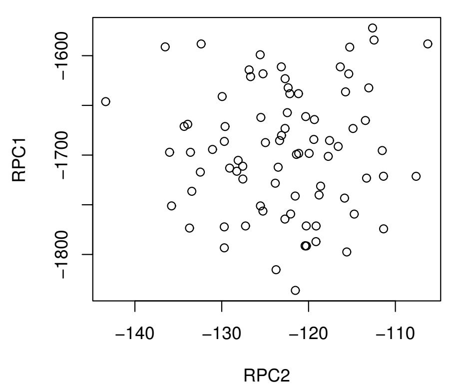

Figure 2: First Two Robust Principal Components with Outliers Omitted

resulted in a first principal component corresponding to a major axis that passed through the outliers. See Figure 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 $\hat{\boldsymbol{e}}_1 \approx (-1.000, 0.002, -0.023, -0.002, -0.009)^T$ 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. These two components explain about 86% of the variance.

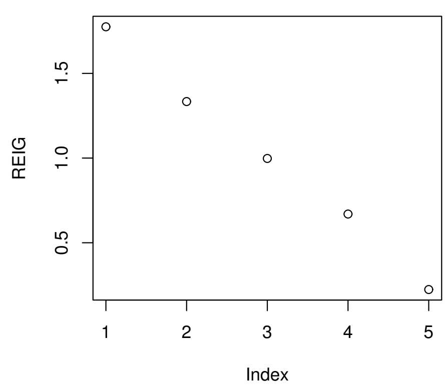

Figure 3: Robust Scree Plot

Figure 3 shows the robust scree plot which suggests that the last principal component can be deleted.

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 *zw* 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] 1.3184358 1.1723991 1.0155266 0.7867349 0.4867867

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] 1.3323400 1.1548879 0.9988643 0.8182741 0.4730769

Rotation:

          PC1      PC2      PC3      PC4      PC5

buxy     -0.10724 -0.69431 -0.11325  0.69184 -0.12238

len       0.69909 -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:

          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
```

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 $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the z_i are iid from a continuous distribution with variance σ^2. Here $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. 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 $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\boldsymbol{z} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{z} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). The choice $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$ results in $\boldsymbol{\Sigma} = diag(1, ..., p)$. Note that the population eigenvalues will be proportional to $(p, p-1, ..., 1)^T$ and the population "variance explained" by the ith 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 ith population eigenvalue is proportional to $1/p$ and the population "variance explained" by the ith principal component is $1/p$.

Table 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 ith eigenvectors or the ith 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

Table 1: 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
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
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
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
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
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
10000	L	M	0.400,0.301,0.200	0.402,0.300,0.199	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

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 ρ is high, or if p is high and $\rho \geq 0.5$, then the data are concentrated about the line with direction $\mathbf{1} = (1, ..., 1)^T$. For $p = 50$ and $\rho = 0.99$, the population variance 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 ρ are small.

Table 2: Variance Explained by PCA and RPCA, SSD = 10^7 SD, $p = 50$

n	type	vexpl	SSD	rvexpl	SSD	a_1
200	N	0.999998	1.958	0.999998	2.867	0.687
1000	N	0.999998	0.917	0.999998	0.971	0.944
1000	C	0.999996	161.3	0.999998	1.482	0.112
1000	L	0.999998	0.919	0.999998	1.508	0.175

Table 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.

4 Conclusions

To use PCA, assume the DD plot of classical versus robust Mahalanobis distances and the subplots of the scatterplot matrix are linear. Want $n > 10p$ for classical PCA and $n > 20p$ 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 $\max_{i=1,...,p} S_i^2 / \min_{i=1,...,p} S_i^2 > 2$. If $\boldsymbol{S}$ is used, also do a PCA using $\boldsymbol{R}$.

Jolliffe (2010) is an authoritative text on PCA. Cattell (1966) and Bentler and Yuan (1998) are good references for scree plots. Mϕller, 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 $\sqrt{n}(\hat{\lambda}_i - \lambda_i) \xrightarrow{D} N(0, \kappa_i + 2\lambda_i^2)$ where κ_i is the kurtosis of the marginal distribution of x_i, for $i = 1, ..., p$.

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 1 and `pcasim` for Tables 2 and 3. See Zhang (2011) for more extensive simulations.

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

Bentler, P.M., and Yuan, K.H. (1998), "Tests for Linear Trend in the Smallest Eigenvalues of the Correlation Matrix," *Psychometrika,* 63, 131-144.

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.

Boente, G. (1987), "Asymptotic Theory for Robust Principal Components," *Journal of*

Multivariate Analysis, 21, 67-78.

Boente, G., and Fraiman, R. (1999), "Discussion of 'Robust Principal Component Analysis for Functional Data' by Locantore et al," *Test*, 8, 28-35.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," *Journal of Multivariate Analysis,* 101, 2372-2388.

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.

Cattell, R.B. (1966), "The Scree Test for the Number of Factors," *Multivariate Behavioral Research,* 1, 245-276.

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.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Jolliffe, I.T. (2010), *Principal Component Analysis*, 2nd ed., Springer, New York, NY.

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.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, Wiley, Hoboken, NJ.

Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317.

Mϕller, S.F., von Frese, J., and Bro, R. (2005), "Robust Methods for Multivariate Data Analysis," *Journal of Chemometrics,* 19, 549-563.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://lagrange.math.siu.edu/Olive/pphbmld.pdf).

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Stewart, G.M. (1969), "On the Continuity of the Generalized Inverse," *SIAM Journal on Applied Mathematics*, 17, 33-45.

Taskinen, S., Koch, I., and Oja, H. (2012), "Robustifying Principal Component Analysis with Spatial Sign Vectors," *Statistics & Probability Letters*, 82, 765-774.

Waternaux, C.M. (1976), "Asymptotic Distribution of the Sample Roots for a Nonnormal Population," *Biometrika*, 63, 639-645.

Zhang, J. (2011), *Applications of a Robust Dispersion Estimator*, Ph.D. Thesis, Southern Illinois University, online at (http://lagrange.math.siu.edu/Olive/szhang.pdf).

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.