11 - Principal Components Analysis

Junvie Pailden

SIUE, F2017, Stat 589

September 25, 2017

General Objectives

- Explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.
  1. Data Reduction
  2. Interpretation
- Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number of k of the principal components.

- If so, there is as much information in the k components as there is in the original p variables.
- The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set consisting of n measurements on k principal components.

Population Principal Components

- Principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$.
- Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes.
- The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.
- Principal components depend solely on the covariance matrix $\mathbf{\Sigma}$ (or the correlation matrix $\boldsymbol{\rho}$) of $X_1, X_2, \ldots, X_p$.
- **NO Multivariate Normal Assumption Required**

Population Principal Components: Notation

- Let the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
- Consider the linear combinations

$$
\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

- We obtain

$$
\begin{aligned}
Var(Y_i) &= \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i \quad i = 1, 2, \ldots, p \\
Cov(Y_i, Y_k) &= \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k \;\; i, k = 1, 2, \ldots, p
\end{aligned}
$$

- The principal components are those uncorrelated linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances in $Var(Y_i)$ are as large as possible.

First principal component = linear combination $\mathbf{a}_1'\mathbf{X}$ that maximizes

$$Var(\mathbf{a}_1'\mathbf{X}) \text{ subject to } \mathbf{a}_1'\mathbf{a}_1 = 1$$

Second principal component = linear combination $\mathbf{a}_2'\mathbf{X}$ that maximizes

$$Var(\mathbf{a}_2'\mathbf{X}) \text{ subject to } \mathbf{a}_2'\mathbf{a}_2 = 1$$

$$Cov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = 0$$

At the ith step ,

ith principal component = linear combination $\mathbf{a}_i'\mathbf{X}$ that maximizes

$$Var(\mathbf{a}_i'\mathbf{X}) \text{ subject to } \mathbf{a}_i'\mathbf{a}_i = 1 \text{ and}$$

$$Cov(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0 \text{ for } k < i$$

Result 8.1

Let $\mathbf{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. Let $\mathbf{\Sigma}$ have the eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_2), (\lambda_1, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$

Then the ith *principal component* is given by

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \;\; i = 1, 2, \ldots, p$$

With these choices,

$$\begin{aligned} Var(Y_i) &= \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i \;\; i = 1, 2, \ldots, p \\ Cov(Y_i, Y_k) &= \mathbf{e}_i'\Sigma\mathbf{e}_k = 0 \;\; i \neq k \end{aligned}$$

Result 8.1 (cont)

If some λ_i are equal, the choices of the corresponding coefficient vectors, $\mathbf{e}_i$ and hence Y_i are not unique.

Result 8.1, the principal components are uncorrelated and have variances equal to the eigenvalues of $\boldsymbol{\Sigma}$.

Result 8.2

Let $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have covariance matrix $\mathbf{\Sigma}$, with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}$,. . .,$Y_p = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1+\lambda_2+\cdots+\lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

Result 8.2 says that

$$\begin{aligned} \text{Total population variance} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

Result 8.2 (cont)

and, the proportion of total variance due to (explained by) the kth principal component is

$$\begin{pmatrix} \text{Proportion of total} \\ \text{population variance} \\ \text{due to } kth \text{ principal} \\ \text{component} \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- If most (for instance, 80 to 90%) of the total population variance, for large p, can be attributed to the first one, two, or three components, then these components can “replace” the original p variables without much of loss of information. Each component of the coefficient vector $\mathbf{e}_i' = [e_{i1}, \ldots, e_{ik}, \ldots, e_{ip}]$ also merits inspection.
- The magnitude of e_{ik} measures the importance of the kth variable to the ith principal component, irrespective of the other variables. In particular, e_{ik} is proportional to the correlation coefficient between Y_i and X_k.

Result 8.3

If $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}$,...,$Y_p = \mathbf{e}_p'\mathbf{X}$ are the principal components obtained from the covariance matrix $\mathbf{\Sigma}$, then

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \ldots, p$$

are the correlation coefficients between the components Y_i and the variables X_k.

Example 8.1 Calculation the population principal components I

Let X_1, X_2, and X_3 have the covariance matrix

```
(Sigma <- matrix(c(1,-2,0,-2,5,0,0,0,2), nrow=3, byrow=T))
```

```
#      [,1] [,2] [,3]
# [1,]    1   -2    0
# [2,]   -2    5    0
# [3,]    0    0    2
```

```
eigen.Sigma <- eigen(Sigma)
(lambda <- eigen.Sigma$values)
```

```
# [1] 5.83 2.00 0.17
```

Example 8.1 Calculation the population principal components II

```
(eigen.Sigma$vectors)
```

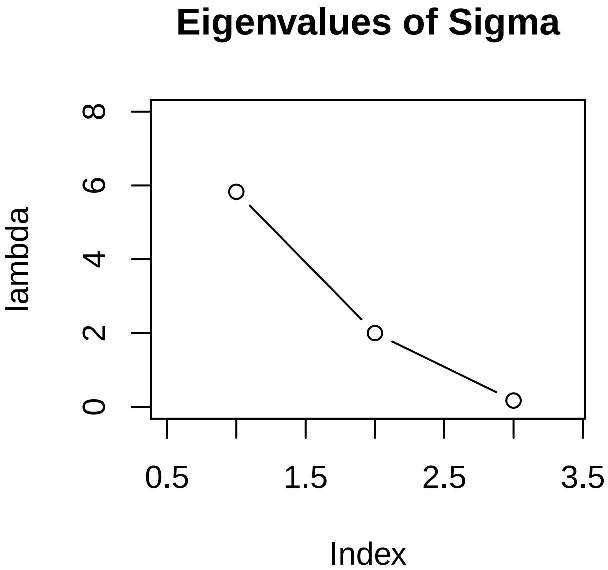
```
#       [,1] [,2] [,3]
# [1,] -0.38    0 0.92
# [2,]  0.92    0 0.38
# [3,]  0.00    1 0.00
```

```
(prop.var <- cumsum(lambda)/sum(lambda) )
```

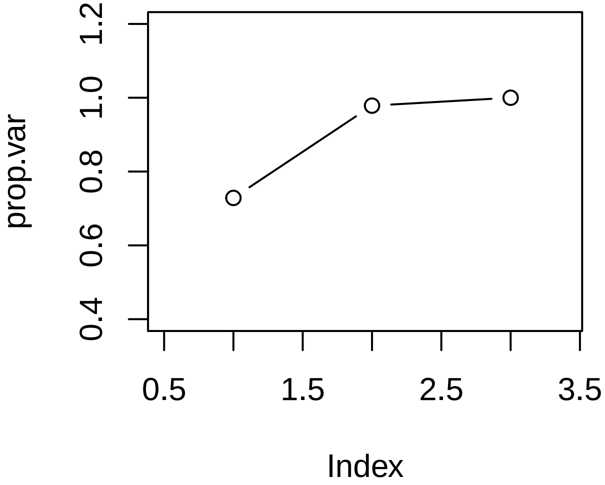
```
# [1] 0.73 0.98 1.00
```

Plots

```
plot(lambda, type="b", main="Eigenvalues of Sigma",
     xlim = c(0.5, 3.4),ylim = c(0,8))
plot(prop.var, type="b", main="Proportion of Variation",
     xlim = c(0.5, 3.4), ylim = c(0.4,1.2))
```


Eigenvalues of Sigma
lambda
0
2
4
6
8
0.5
1.5
2.5
3.5
Index

Proportion of Variation



Principal Components Obtained from Standardized Variables

Principal components may also be obtained for the standardized variables

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}},\ Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}, \ldots, Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

In matrix notation,

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \mu),$$
$$E(\mathbf{Z}) = 0, \text{and } , Cov(Z) = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho},$$

where $\mathbf{V}^{1/2}$ is the diagonal standard deviation matrix.

Result 8.4

The ith principal component of the standardized variables $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_p]_{\text{with}} Cov(\mathbf{Z}) = \boldsymbol{\rho}$, is given by

$$Y_i = e_i' Z = e_i'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \mu), \; i = 1, 2, \ldots, p$$

Moreover,

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$$

and

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i} \quad i, k = 1, 2, \ldots, p.$$

$$\begin{pmatrix} \text{Proportion of (standardized)} \\ \text{population variance} \\ \text{due to } kth \text{ principal} \\ \text{component} \end{pmatrix} = \frac{\lambda_k}{p}, \;\; k = 1, 2, \ldots, p$$

Summarizing Sample Variation by Principal Components

- Suppose the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ represent n independent drawings from some p-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- These data yield the sample mean vector $\bar{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$.
- Our objective in this section will be to construct uncorrelated linear combinations of the measured characteristics that account for much of the variation in the sample.
- The uncorrelated combinations with the largest variances will be called the *sample principal components*.

- Recall that the n values of any linear combination

$$\mathbf{a}_1'\mathbf{x} = a_{11}x_{j1} + a_{12}x_{j2} + \cdots + a_{1p}x_{jp}, \quad j = 1, 2, \ldots, n$$

  have the sample mean $\mathbf{a}_1'\bar{\mathbf{x}}$ and sample variance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$.
- Also, the pairs of values $(\mathbf{a}_1'\mathbf{x}_j, \mathbf{a}_2'\mathbf{x}_j)$ have sample covariance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$.

If $\mathbf{S} = \{s_{ik}\}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, the ith sample principal component is given by

$$\hat{y}_i = \hat{\mathbf{e}}_i' = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \quad i = 1, 2, \ldots, p$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and $\mathbf{x}$ is any observation on the variables $X_1, X_2, \ldots, X_p$. Also,

$$\text{Sample variance}\,(\hat{y}_k) = \hat{\lambda}_k, \; k = 1, 2, \ldots, p$$
$$\text{Sample covariance}\,(\hat{y}_i, \hat{y}_k) = 0, \; i \neq k$$

In addition,

$$\text{Total sample variance } = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

Sample Variation by Principal Components

- We shall denote the sample principal components by $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$.
- The observations $\mathbf{x}_j$ are often "centered" by subtracting $\bar{\mathbf{x}}$.
- This has no effect on the sample covariance matrix $\mathbf{S}$ and gives the ith principal component

$$\hat{y}_i = \hat{\mathbf{e}}_i'(x - \bar{x}), \quad i = 1, 2, \ldots, p$$

for any observation vector $\mathbf{x}$.

- If we consider the values of the ith component

$$\hat{y}_{ji} = \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, 2, \ldots, n$$

generating by substituting each observation $\mathbf{x}_j$ for the arbitrary $\mathbf{x}$, then

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}) = \frac{1}{n}\hat{\mathbf{e}}_i'\left(\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})\right) = \frac{1}{n}\hat{\mathbf{e}}_i'\mathbf{0} = 0$$

The Number of Principal Components

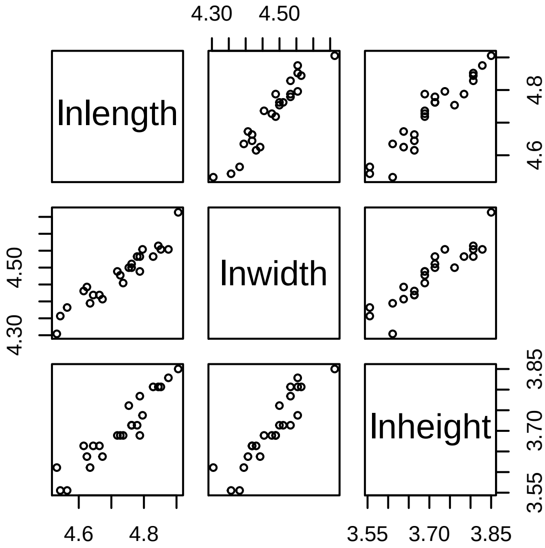
- There is always the question of how many components to retain.
- A useful visual aid to determining an appropriate number of principal components is a ***scree plot***.
- With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i - the magnitude of an eigenvalue versus its number.
- To determine the appropriate number of components, we look for an elbow (bend) in the scree plot.
- The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.

Example 8.4: Summarizing the data with one sample principal component

In a study of size and relationships for painted male turtles, Joicoeur and Moistmann measured carapace length, width, and height. The authors suggests a logarithmic transformation in studies of size-and-shape relationships. Perform a principal component analysis.

```r
turtles <- read.table("T6-9.DAT", header=F)[25:48,-4]
X <- log(as.matrix(turtles))
colnames(X) <- c("lnlength", "lnwidth", "lnheight")
```

```r
# Make plots of data:
pairs(X, cex = 0.7)
```



```r
# Compute means and covariance matrix
colMeans(X)
```

```
# lnlength  lnwidth lnheight
#      4.7      4.5      3.7
```

```r
(S <- cov(X))
```

```
#          lnlength lnwidth lnheight
# lnlength   0.0111  0.0080   0.0082
# lnwidth    0.0080  0.0064   0.0060
# lnheight   0.0082  0.0060   0.0068
```

Compute principal components for original data. Use `prcomp` built-in function in R

```r
(turtles.pcomp <- prcomp(X))
```

```
# Standard deviations (1, .., p=3):
# [1] 0.153 0.024 0.019
#
# Rotation (n x k) = (3 x 3):
#            PC1   PC2   PC3
# lnlength 0.68 -0.16  0.71
# lnwidth  0.51 -0.59 -0.62
# lnheight 0.52  0.79 -0.32
```

```r
# eigenvalues
turtles.pcomp$sdev^2
```

```
# [1] 0.02330 0.00060 0.00036
```

```r
summary(turtles.pcomp)
```

```
# Importance of components%s:
#                          PC1    PC2    PC3
# Standard deviation     0.153 0.0245 0.0190
# Proportion of Variance 0.961 0.0247 0.0148
# Cumulative Proportion  0.961 0.9852 1.0000
```

Checking

Check calculations of principal components by computing eigenvalues/eigenvectors and proportion explained from the original data. Use the covariance matrix S

```
eigen.turtles <- eigen(S)
eigen.turtles$values
```

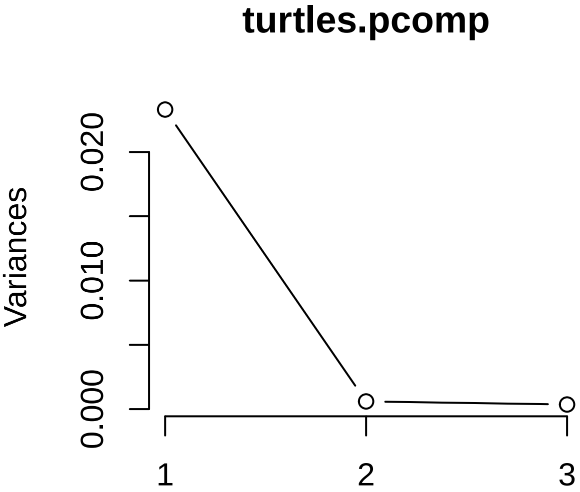
```
# [1] 0.02330 0.00060 0.00036
```

```
cumsum(eigen.turtles$values)/sum(eigen.turtles$values)
```

```
# [1] 0.96 0.99 1.00
```

Screeplot order eigenvalues from largest to smallest

```
screeplot(turtles.pcomp, type="lines")
```



```
eigen.turtles$vectors
```

```
#       [,1]  [,2]  [,3]
# [1,] 0.68 -0.16  0.71
# [2,] 0.51 -0.59 -0.62
# [3,] 0.52  0.79 -0.32
```

Turtle Measurements via Principal Components

- The scree plot in the previous slide has a very distinct elbow that occurs at $i = 2$. There is clearly a dominant principal component.
- The first principal component, which explains 96% of the total variance, has an interesting subject-matter interpretations.

$$\begin{aligned} \hat{y}_1 &= .68 \ln(\text{length}) + .51 \ln(\text{width}) + .52 \ln(\text{height}) \\ &= \ln\left[(\text{length})^{.68} \times (\text{width})^{.51} \times (\text{height})^{.52}\right] \end{aligned}$$

- The first principal component $\hat{y}_1$ may be viewed as the *volume* of a box with adjusted dimensions.
- For instance, the adjusted height, $(\text{height})^{.52}$, can account (in some sense) for the rounded shape of the carapace.
- The values of the first principal component can be computed as

$$\hat{\mathbf{y}}_1 = \begin{bmatrix} \hat{y}_{11} \\ \hat{y}_{21} \\ \vdots \\ \hat{y}_{n1} \end{bmatrix} = \mathbf{X}\hat{\mathbf{e}}_1 = \mathbf{X}[.68, .51, .52]$$

```
(ev1 <- turtles.pcomp$rotation[, 1])
```

```
# lnlength  lnwidth lnheight
#     0.68     0.51     0.52
```

```
PC1 <- X%*%ev1
summary(PC1)
```

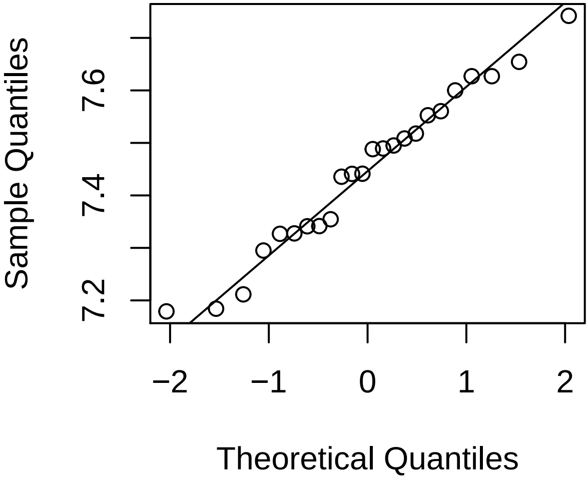
```
#        V1
#  Min.   :7.2
#  1st Qu.:7.3
#  Median :7.5
#  Mean   :7.4
#  3rd Qu.:7.6
#  Max.   :7.7
```

```r
stem(PC1)
```

```
#
#   The decimal point is 1 digit(s) to the left of the |
#
#   70 | 88
#   72 | 1933445
#   74 | 4449901256
#   76 | 03354
```

```r
qqnorm(PC1)
qqline(PC1)
```

Normal Q−Q Plot



Standardizing the Sample Principal Components

- Sample principal components are, in general, not invariant with respect to changes in scale.
- Variables measured on different scales or on a common scale with widely differing ranges are often standardized.
- For $j = 1, 2. \ldots, n$, the standardized observation of the jth observation in the sample is

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \begin{bmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \begin{bmatrix} z_{j1} \\ z_{j2} \\ \vdots \\ z_{jp} \end{bmatrix}$$

Standardizing the Sample Principal Components

- The $n \times p$ data matrix of standardized observations

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

- Verify that $\bar{\mathbf{z}} = \frac{1}{n}\left(\mathbf{1}'\mathbf{Z}\right)' = \frac{1}{n}\mathbf{Z}'\mathbf{1} = \mathbf{0}$ and $\mathbf{S}_z = \mathbf{R}$, where $\mathbf{R}$ is the correlation matrix.
- The ith principal component, $i = 1, 2, \ldots, p$, is

$$\begin{aligned} \hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{z} &= \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p \\ \text{Sample variance}(\hat{y}_i) &= \hat{\lambda}_i \\ \text{Sample covariance}(\hat{y}_i, \hat{y}_k) &= 0, \;\; i \neq k \end{aligned}$$

where $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ is the ith eigenvalue-eigenvector pair of $\mathbf{R}$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$.

Principal Components for stock return data - Exercise 8.10

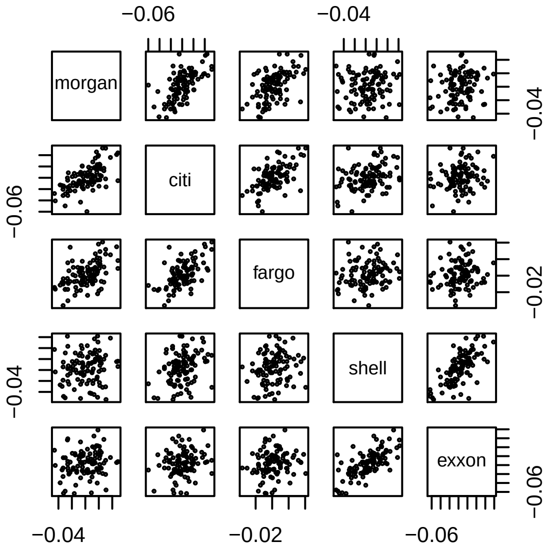
The weekly rates of return for five stocks listed on the New York Stock Exchange are given in T8-4.DAT.

- Construct the sample covariance matrix S, and find the sample principal components.
- Determine the proportion of the total sample variance explained by the first three principal components. Interpret the components.
- Given the results, do you feel that the stock rates-of-terun data can be summarized in fewer than five dimensions? Explain.

```r
stock <- read.table("T8-4.DAT", header=F,
    col.names = c("morgan", "citi", "fargo",
                  "shell", "exxon"))
stock <- as.matrix(stock)
colMeans(stock)
```

```
#  morgan    citi   fargo   shell   exxon
# 0.00106 0.00066 0.00163 0.00405 0.00404
```

```r
pairs(stock, cex = 0.3)
```



Compute principal components for standardized data scale = TRUE will result in all variables being scaled to have unit variance (i.e. a variance of 1, and hence a standard deviation of 1).

```
(stock.pcomp <- prcomp(stock, scale = T))
```

```
# Standard deviations (1, .., p=5):
# [1] 1.56 1.19 0.71 0.63 0.51
#
# Rotation (n x k) = (5 x 5):
#          PC1   PC2    PC3   PC4    PC5
# morgan -0.47  0.37 -0.604  0.36  0.384
# citi   -0.53  0.24 -0.136 -0.63 -0.496
# fargo  -0.47  0.32  0.772  0.29  0.071
# shell  -0.39 -0.59  0.093 -0.38  0.595
# exxon  -0.36 -0.61 -0.109  0.49 -0.498
```

```
# eigenvalues
stock.pcomp$sdev^2
```

Checking

Check calculations of principal components by computing eigenvalues/eigenvectors and proportion explained from the standardized data

```
(R <- cor(stock))
```

```
#        morgan citi fargo shell exxon
# morgan   1.00 0.63  0.51  0.11  0.15
# citi     0.63 1.00  0.57  0.32  0.21
# fargo    0.51 0.57  1.00  0.18  0.15
# shell    0.11 0.32  0.18  1.00  0.68
# exxon    0.15 0.21  0.15  0.68  1.00
```

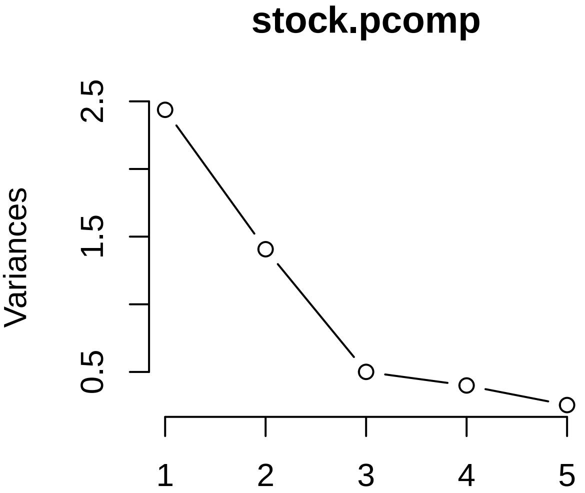
```
eigen.stock <- eigen(R)
eigen.stock$values
```

```
# [1] 2.44 1.41 0.50 0.40 0.26
```

Screeplot - Stocks Data

Screeplot order eigenvalues from largest to smallest.

```
screeplot(stock.pcomp, npcs = 5, type = "lines")
```



Stock Data via Principal Components

- Using the standardized variables, the first two sample principal components

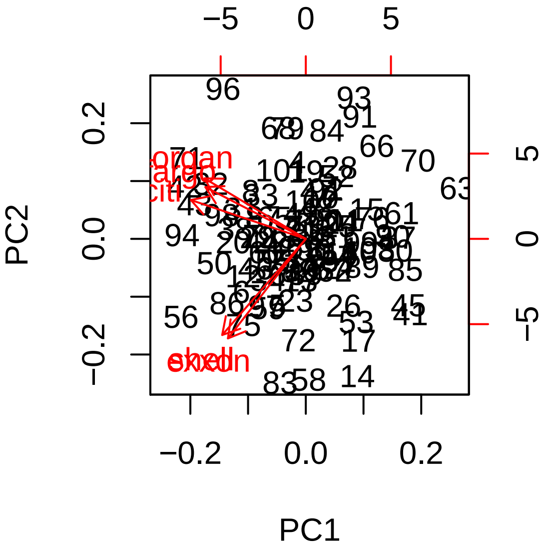
$$\hat{y}_1 = \hat{\mathbf{e}}_1\mathbf{z} = .469z_1 + .532z_2 + .465z_3 + .387z_4 + .361z_5$$
$$\hat{y}_2 = \hat{\mathbf{e}}_2\mathbf{z} = -.368z_1 - .236z_2 - .315z_3 + .585z_4 + .606z_5$$

- These components account for

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}\right)100\% = \left(\frac{2.437 + 1.407}{5}\right)100\% = 77\%$$

  of the total (standardized) sample variance.
- The first component is a roughly equally weighted sum, or "index" of the five stocks. This component might be called a *market component*.

```
biplot(stock.pcomp)
```



- The second component represents a contrast between the banking stocks (JP Morgan, Citibank, Wells Fargo) and the oil stocks (Royal Dutch Shell, Exxon-Mobil). It might be called an *industry component*.
- We see that most of the variation in these stock returns is due to market activity and uncorrelated industry activity.