Principal Components Analysis

Nathaniel E. Helwig

Assistant Professor of Psychology and Statistics
University of Minnesota (Twin Cities)

Updated 16-Mar-2017

Copyright

Outline of Notes

1) Background
- Overview
- Data, Cov, Cor
- Orthogonal Rotation

2) Population PCs
- Definition
- Calculation
- Properties

3) Sample PCs
- Definition
- Calculation
- Properties

4) PCA in Practice
- Covariance or Correlation?
- Decathlon Example
- Number of Components

Background

Definition and Purposes of PCA

Principal Components Analysis (PCA) finds linear combinations of variables that best explain the covariation structure of the variables.

There are two typical purposes of PCA:

1. Data reduction: explain covariation between p variables using $r < p$ linear combinations
2. Data interpretation: find features (i.e., components) that are important for explaining covariation

Data Matrix

The data matrix refers to the array of numbers

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

where x_{ij} is the j-th variable collected from the i-th item (e.g., subject).

- items/subjects are rows
- variables are columns

$\mathbf{X}$ is a data matrix of order $n \times p$ (# items by # variables).

Population Covariance Matrix

The population covariance matrix refers to the symmetric array

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_{pp} \end{pmatrix}$$

where

- $\sigma_{jj} = E([X_j - \mu_j]^2)$ is the population variance of the j-th variable
- $\sigma_{jk} = E([X_j - \mu_j][X_k - \mu_k])$ is the population covariance between the j-th and k-th variables
- $\mu_j = E(X_j)$ is the population mean of the j-th variable

Sample Covariance Matrix

The sample covariance matrix refers to the symmetric array

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{pmatrix}$$

where

- $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the sample variance of the j-th variable
- $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ is the sample covariance between the j-th and k-th variables
- $\bar{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$ is the sample mean of the j-th variable

Covariance Matrix from Data Matrix

We can calculate the (sample) covariance matrix such as

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c$$

where $\mathbf{X}_c = \mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}' = \mathbf{C}\mathbf{X}$ with

- $\bar{\mathbf{x}}' = (\bar{x}_1, \ldots, \bar{x}_p)$ denoting the vector of variable means
- $\mathbf{C} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ denoting a centering matrix

Note that the centered matrix $\mathbf{X}_c$ has the form

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ x_{31} - \bar{x}_1 & x_{32} - \bar{x}_2 & \cdots & x_{3p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

Population Correlation Matrix

The population correlation matrix refers to the symmetric array

$$\mathbf{P} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots & 1 \end{pmatrix}$$

where

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}$$

is the Pearson correlation coefficient between variables X_j and X_k.

Sample Correlation Matrix

The sample correlation matrix refers to the symmetric array

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

where

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

is the Pearson correlation coefficient between variables $\mathbf{x}_j$ and $\mathbf{x}_k$.

Correlation Matrix from Data Matrix

We can calculate the (sample) correlation matrix such as

$$\mathbf{R} = \frac{1}{n-1}\mathbf{X}_s'\mathbf{X}_s$$

where $\mathbf{X}_s = \mathbf{CXD}^{-1}$ with

- $\mathbf{C} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ denoting a centering matrix
- $\mathbf{D} = \text{diag}(s_1, \ldots, s_p)$ denoting a diagonal scaling matrix

Note that the standardized matrix $\mathbf{X}_s$ has the form

$$\mathbf{X}_s = \begin{pmatrix} (x_{11} - \bar{x}_1)/s_1 & (x_{12} - \bar{x}_2)/s_2 & \cdots & (x_{1p} - \bar{x}_p)/s_p \\ (x_{21} - \bar{x}_1)/s_1 & (x_{22} - \bar{x}_2)/s_2 & \cdots & (x_{2p} - \bar{x}_p)/s_p \\ (x_{31} - \bar{x}_1)/s_1 & (x_{32} - \bar{x}_2)/s_2 & \cdots & (x_{3p} - \bar{x}_p)/s_p \\ \vdots & \vdots & \ddots & \vdots \\ (x_{n1} - \bar{x}_1)/s_1 & (x_{n2} - \bar{x}_2)/s_2 & \cdots & (x_{np} - \bar{x}_p)/s_p \end{pmatrix}$$

Rotating Points in Two Dimensions

Suppose we have $\mathbf{z} = (x, y)' \in \mathbb{R}^2$, i.e., points in 2D Euclidean space.

A 2×2 orthogonal rotation of (x, y) of the form

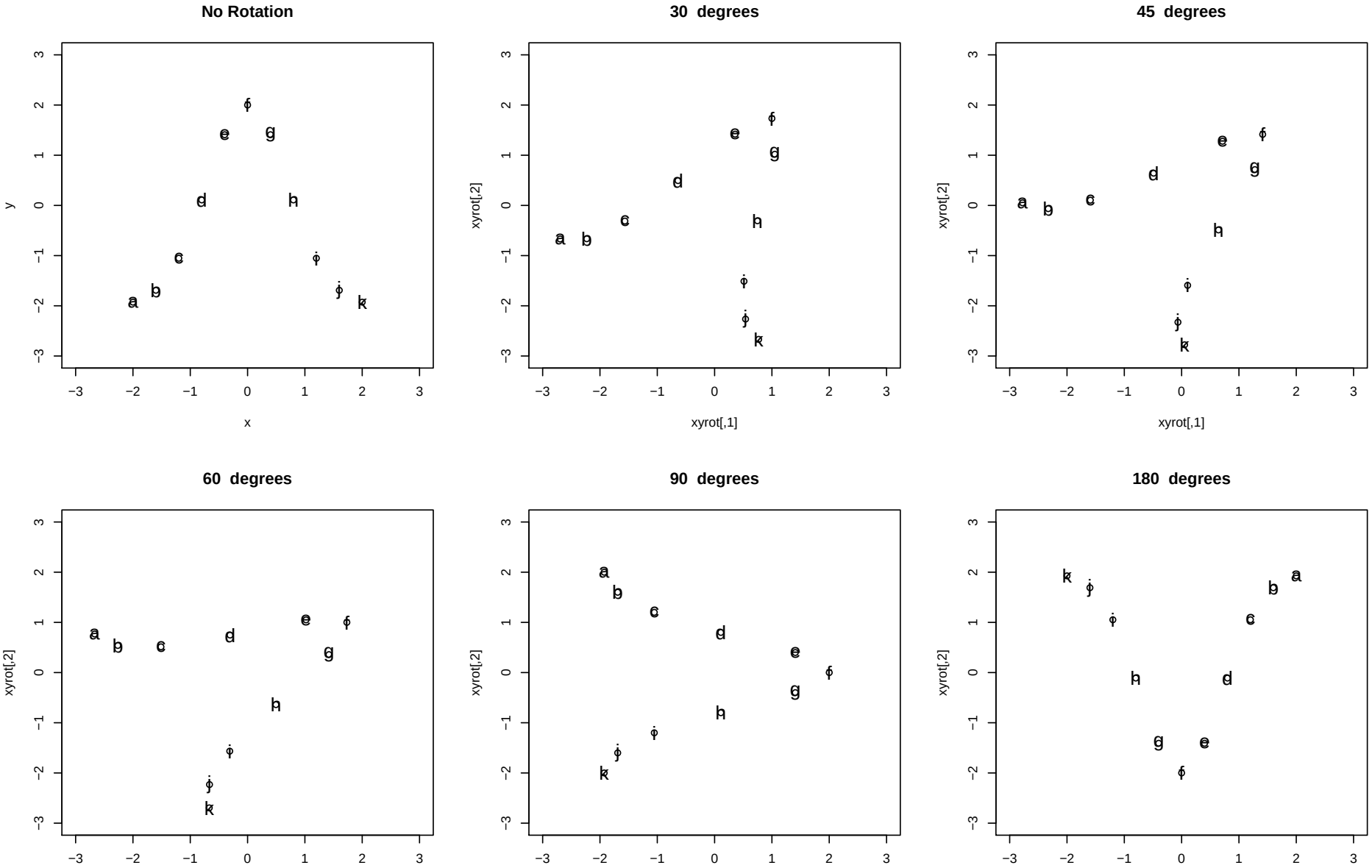
$$\begin{pmatrix} x^* \\ y^* \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

rotates (x, y) counter-clockwise around the origin by an angle of θ and

$$\begin{pmatrix} x^* \\ y^* \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

rotates (x, y) clockwise around the origin by an angle of θ.

Visualization of 2D Clockwise Rotation



Visualization of 2D Clockwise Rotation (R Code)

```
rotmat2d <- function(theta){
  matrix(c(cos(theta),sin(theta),-sin(theta),cos(theta)),2,2)
}
x <- seq(-2,2,length=11)
y <- 4*exp(-x^2) - 2
xy <- cbind(x,y)
rang <- c(30,45,60,90,180)
dev.new(width=12,height=8,noRStudioGD=TRUE)
par(mfrow=c(2,3))
plot(x,y,xlim=c(-3,3),ylim=c(-3,3),main="No Rotation")
text(x,y,labels=letters[1:11],cex=1.5)
for(j in 1:5){
  rmat <- rotmat2d(rang[j]*2*pi/360)
  xyrot <- xy%*%rmat
  plot(xyrot,xlim=c(-3,3),ylim=c(-3,3))
  text(xyrot,labels=letters[1:11],cex=1.5)
  title(paste(rang[j]," degrees"))
}
```

Orthogonal Rotation in Two Dimensions

Note that the 2×2 rotation matrix

$$\mathbf{R} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

is an orthogonal matrix for all θ:

$$\begin{aligned} \mathbf{R}'\mathbf{R} &= \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \\ &= \begin{pmatrix} \cos^2(\theta) + \sin^2(\theta) & \cos(\theta)\sin(\theta) - \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) - \cos(\theta)\sin(\theta) & \cos^2(\theta) + \sin^2(\theta) \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

Orthogonal Rotation in Higher Dimensions

Suppose we have a data matrix $\mathbf{X}$ with p columns.

- Rows of $\mathbf{X}$ are coordinates of points in p-dimensional space
- Note: when $p = 2$ we have situation on previous slides

A $p \times p$ orthogonal rotation is an orthogonal linear transformation.

- $\mathbf{R}'\mathbf{R} = \mathbf{R}\mathbf{R}' = \mathbf{I}_p$ where $\mathbf{I}_p$ is $p \times p$ identity matrix
- If $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{R}$ is rotated data matrix, then $\tilde{\mathbf{X}}\tilde{\mathbf{X}}' = \mathbf{X}\mathbf{X}'$
- Orthogonal rotation preserves relationships between points

Population Principal Components

Linear Combinations of Random Variables

$\boldsymbol{X} = (X_1, \ldots, X_p)'$ is a random vector with covariance matrix $\boldsymbol{\Sigma}$, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\boldsymbol{\Sigma}$.

Consider forming new variables $Y_1, \ldots, Y_p$ by taking p different linear combinations of the X_j variables:

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

where $\mathbf{b}_k' = (b_{1k}, \ldots, b_{pk})$ is the k-th linear combination vector.

- $\mathbf{b}_k$ are called the loadings for the k-th principal component

Defining Principal Components in the Population

Note that the random variable $Y_k = \mathbf{b}_k' \boldsymbol{X}$ has the properties:

$$\mathrm{Var}(Y_k) = \mathbf{b}_k' \boldsymbol{\Sigma} \mathbf{b}_k$$
$$\mathrm{Cov}(Y_k, Y_\ell) = \mathbf{b}_k' \boldsymbol{\Sigma} \mathbf{b}_\ell$$

The principal components are the <u>uncorrelated</u> linear combinations $Y_1, \ldots, Y_p$ whose variances are as large as possible.

$$\mathbf{b}_1 = \underset{\|\mathbf{b}_1\|=1}{\mathrm{argmax}} \{\mathrm{Var}(\mathbf{b}_1' \boldsymbol{X})\}$$

$$\mathbf{b}_2 = \underset{\|\mathbf{b}_2\|=1}{\mathrm{argmax}} \{\mathrm{Var}(\mathbf{b}_2' \boldsymbol{X})\} \quad \text{subject to} \quad \mathrm{Cov}(\mathbf{b}_1' \boldsymbol{X}, \mathbf{b}_2' \boldsymbol{X}) = 0$$

$$\mathbf{b}_\ell = \underset{\|\mathbf{b}_\ell\|=1}{\mathrm{argmax}} \{\mathrm{Var}(\mathbf{b}_\ell' \boldsymbol{X})\} \quad \text{subject to} \quad \mathrm{Cov}(\mathbf{b}_k' \boldsymbol{X}, \mathbf{b}_\ell' \boldsymbol{X}) = 0 \ \ \forall \ \ k < \ell$$

Visualizing Principal Components for Bivariate Normal

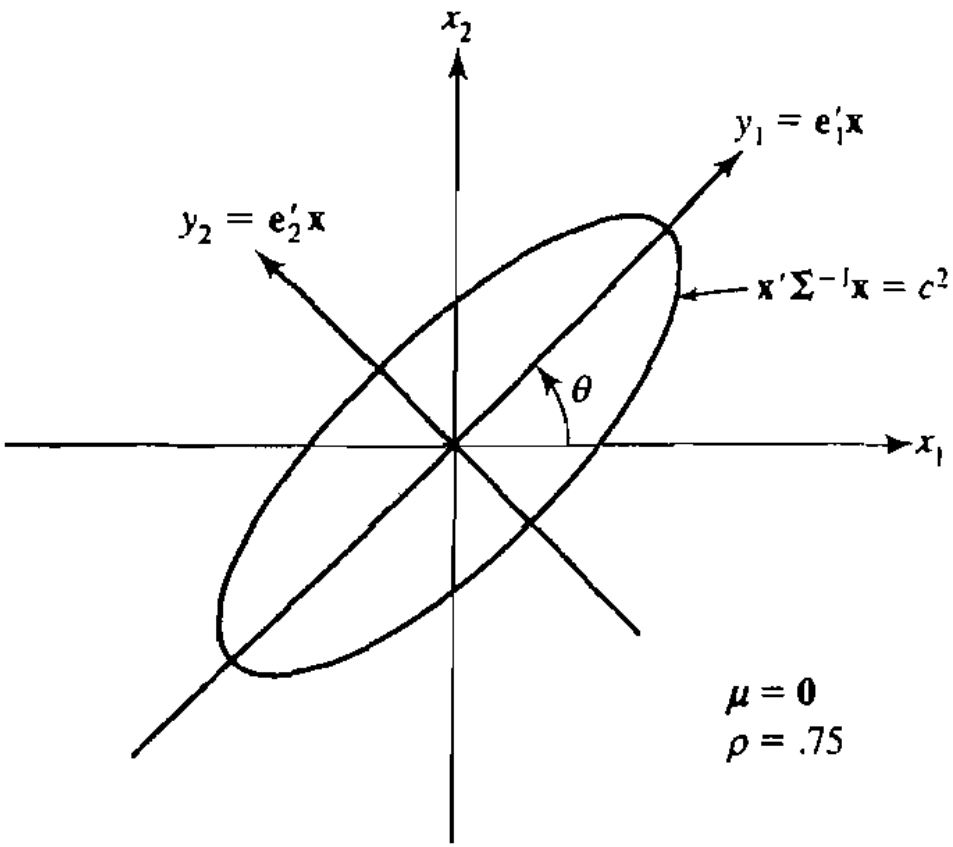


Figure 8.1 The constant density ellipse $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

Figure: Figure 8.1 from Applied Multivariate Statistical Analysis, 6th Ed (Johnson & Wichern). Note that $\mathbf{e}_1$ and $\mathbf{e}_2$ denote the eigenvectors of $\boldsymbol{\Sigma}$.

PCA Solution via the Eigenvalue Decomposition

We can write the population covariance matrix $\mathbf{\Sigma}$ such as

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k'$$

where

- $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ contains the eigenvectors ($\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_p$)
- $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ contains the (non-negative) eigenvalues

The PCA solution is obtained by setting $\mathbf{b}_k = \mathbf{v}_k$ for $k = 1, \ldots, p$:

- $\text{Var}(Y_k) = \text{Var}(\mathbf{v}_k'\boldsymbol{X}) = \mathbf{v}_k'\mathbf{\Sigma}\mathbf{v}_k = \mathbf{v}_k'\mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{v}_k = \lambda_k$
- $\text{Cov}(Y_k, Y_\ell) = \text{Cov}(\mathbf{v}_k'\boldsymbol{X}, \mathbf{v}_\ell'\boldsymbol{X}) = \mathbf{v}_k'\mathbf{\Sigma}\mathbf{v}_\ell = \mathbf{v}_k'\mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{v}_\ell = 0$ if $k \neq \ell$

Variance Explained by Principal Components

$Y_1, \ldots, Y_p$ has the same total variance as $X_1, \ldots, X_p$:

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \mathrm{tr}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{V}\mathbf{\Lambda}\mathbf{V}') = \mathrm{tr}(\mathbf{\Lambda}) = \sum_{j=1}^{p} \mathrm{Var}(Y_j)$$

The proportion of the total variance accounted for by the k-th PC is

$$R_k^2 = \frac{\lambda_k}{\sum_{\ell=1}^{p} \lambda_\ell}$$

If $\sum_{k=1}^{r} R_k^2 \approx 1$ for some $r < p$, we do not lose much transforming the original variables into fewer new (principal component) variables.

Covariance of Variables and Principal Components

The covariance between X_j and Y_k has the form

$$\mathrm{Cov}(X_j, Y_k) = \mathrm{Cov}(\mathbf{e}_j'\boldsymbol{X}, \mathbf{v}_k'\boldsymbol{X}) = \mathbf{e}_j'\boldsymbol{\Sigma}\mathbf{v}_k = \mathbf{e}_j'(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}')\mathbf{v}_k = \mathbf{e}_j'\mathbf{v}_k\lambda_k = v_{jk}\lambda_k$$

where

- $\mathbf{e}_j$ is a vector of zeros with a one in the j-th position
- $\mathbf{v}_k = (v_{1k}, \ldots, v_{pk})'$ is the k-th eigenvector of $\boldsymbol{\Sigma}$

This implies that the correlation between X_j and Y_k has the form

$$\mathrm{Cor}(X_j, Y_k) = \frac{\mathrm{Cov}(X_j, Y_k)}{\sqrt{\mathrm{Var}(X_j)}\sqrt{\mathrm{Var}(Y_k)}} = \frac{v_{jk}\lambda_k}{\sqrt{\sigma_{jj}}\sqrt{\lambda_k}} = \frac{v_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}$$

Sample Principal Components

Linear Combinations of Observed Random Variables

$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is an observed random vector and $\mathbf{x}_i \overset{\text{iid}}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Consider forming new variables $y_{i1}, \ldots, y_{ip}$ by taking p different linear combinations of the x_{ij} variables:

$$
\begin{aligned}
y_{i1} &= \mathbf{b}_1'\mathbf{x}_i = b_{11}x_{i1} + b_{21}x_{i2} + \cdots + b_{p1}x_{ip} \\
y_{i2} &= \mathbf{b}_2'\mathbf{x}_i = b_{12}x_{i1} + b_{22}x_{i2} + \cdots + b_{p2}x_{ip} \\
&\vdots \\
y_{ip} &= \mathbf{b}_p'\mathbf{x}_i = b_{1p}x_{i1} + b_{2p}x_{i2} + \cdots + b_{pp}x_{ip}
\end{aligned}
$$

where $\mathbf{b}_k' = (b_{1k}, \ldots, b_{pk})$ is the k-th linear combination vector.

Sample Properties of Linear Combinations

Note that the sample mean and variance of the y_{ik} variables are:

$$\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_{ik} = \mathbf{b}_k'\bar{\mathbf{x}}$$

$$s_{y_k}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{b}_k'\mathbf{x}_i - \mathbf{b}_k'\bar{\mathbf{x}})(\mathbf{b}_k'\mathbf{x}_i - \mathbf{b}_k'\bar{\mathbf{x}})'$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{b}_k'(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{b}_k = \mathbf{b}_k'\mathbf{S}\mathbf{b}_k$$

and the sample covariance between y_{ik} and $y_{i\ell}$ is given by

$$s_{y_k y_\ell} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ik}-\bar{y}_k)(y_{i\ell}-\bar{y}_\ell) = \mathbf{b}_k'\mathbf{S}\mathbf{b}_\ell$$

Defining Principal Components in the Sample

The principal components are the uncorrelated linear combinations $y_{i1}, \ldots, y_{ip}$ whose sample variances are as large as possible.

$$\mathbf{b}_1 = \underset{\|\mathbf{b}_1\|=1}{\operatorname{argmax}} \{\mathbf{b}_1' \mathbf{S} \mathbf{b}_1\}$$

$$\mathbf{b}_2 = \underset{\|\mathbf{b}_2\|=1}{\operatorname{argmax}} \{\mathbf{b}_2' \mathbf{S} \mathbf{b}_2\} \quad \text{subject to} \quad \mathbf{b}_1' \mathbf{S} \mathbf{b}_2 = 0$$

$$\vdots$$

$$\mathbf{b}_\ell = \underset{\|\mathbf{b}_\ell\|=1}{\operatorname{argmax}} \{\mathbf{b}_\ell' \mathbf{S} \mathbf{b}_\ell\} \quad \text{subject to} \quad \mathbf{b}_k' \mathbf{S} \mathbf{b}_\ell = 0 \ \ \forall \ \ k < \ell$$

PCA Solution via the Eigenvalue Decomposition

We can write the sample covariance matrix $\mathbf{S}$ such as

$$\mathbf{S} = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}' = \sum_{k=1}^{p} \hat{\lambda}_k \hat{\mathbf{v}}_k \hat{\mathbf{v}}_k'$$

where

- $\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p]$ contains the eigenvectors ($\hat{\mathbf{V}}'\hat{\mathbf{V}} = \hat{\mathbf{V}}\hat{\mathbf{V}}' = \mathbf{I}_p$)
- $\hat{\mathbf{\Lambda}} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ contains the (non-negative) eigenvalues

The PCA solution is obtained by setting $\mathbf{b}_k = \hat{\mathbf{v}}_k$ for $k = 1, \ldots, p$:

- $s_{y_k}^2 = \hat{\mathbf{v}}_k'\mathbf{S}\hat{\mathbf{v}}_k = \hat{\mathbf{v}}_k'\hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}'\hat{\mathbf{v}}_k = \hat{\lambda}_k$
- $s_{y_k y_\ell} = \hat{\mathbf{v}}_k'\mathbf{S}\hat{\mathbf{v}}_\ell = \hat{\mathbf{v}}_k'\hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}'\hat{\mathbf{v}}_\ell = 0$ if $k \neq \ell$

Variance Explained by Principal Components

$\{y_{i1}, \ldots, y_{ip}\}_{i=1}^n$ has the same total variance as $\{x_{i1}, \ldots, x_{ip}\}_{i=1}^n$:

$$\sum_{j=1}^{p} s_k^2 = \mathrm{tr}(\mathbf{S}) = \mathrm{tr}(\hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}') = \mathrm{tr}(\hat{\mathbf{\Lambda}}) = \sum_{j=1}^{p} s_{y_k}^2$$

The proportion of the total variance accounted for by the k-th PC is

$$\hat{R}_k^2 = \frac{\hat{\lambda}_k}{\sum_{\ell=1}^{p} \hat{\lambda}_\ell}$$

If $\sum_{k=1}^{r} \hat{R}_k^2 \approx 1$ for some $r < p$, we do not lose much transforming the original variables into fewer new (principal component) variables.

Covariance of Variables and Principal Components

The sample covariance between x_{ij} and y_{ik} has the form

$$\begin{aligned}\mathrm{Cov}(x_{ij}, y_{ik}) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_{ik} - \bar{y}_k) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{e}_j'(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\mathbf{v}}_k \\ &= \mathbf{e}_j'\mathbf{S}\hat{\mathbf{v}}_k = \mathbf{e}_j'\hat{\mathbf{v}}_k\hat{\lambda}_k = \hat{v}_{jk}\hat{\lambda}_k\end{aligned}$$

where

- $\mathbf{e}_j$ is a vector of zeros with a one in the j-th position
- $\hat{\mathbf{v}}_k = (\hat{v}_{1k}, \ldots, \hat{v}_{pk})'$ is the k-th eigenvector of $\mathbf{S}$

This implies that the (sample) correlation between x_{ij} and y_{ik} is

$$\mathrm{Cor}(x_{ij}, y_{ik}) = \frac{\mathrm{Cov}(x_{ij}, y_{ik})}{\sqrt{\mathrm{Var}(x_{ij})}\sqrt{\mathrm{Var}(y_{ik})}} = \frac{\hat{v}_{jk}\hat{\lambda}_k}{s_j\hat{\lambda}_k^{1/2}} = \frac{\hat{v}_{jk}\hat{\lambda}_k^{1/2}}{s_j}$$

Large Sample Properties

Assume that $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that the eigenvalues of $\boldsymbol{\Sigma}$ are strictly positive and unique: $\lambda_1 > \cdots > \lambda_p > 0$.

As $n \to \infty$, we have that

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \approx \mathrm{N}(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$$
$$\sqrt{n}(\hat{\mathbf{v}}_k - \mathbf{v}_k) \approx \mathrm{N}(\mathbf{0}, \mathbf{V}_k)$$

where $\mathbf{V}_k = \lambda_k \sum_{\ell \neq k} \frac{\lambda_\ell}{(\lambda_\ell - \lambda_k)^2} \mathbf{v}_\ell \mathbf{v}_\ell'$

Furthermore, as $n \to \infty$, we have that $\hat{\lambda}_k$ and $\hat{\mathbf{v}}_k$ are independent.

Principal Components Analysis in Practice

PCA Solution is Related to SVD (Covariance Matrix)

Let $\hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'$ denote the SVD of $\mathbf{X}_c = \mathbf{CX}$.

The PCA (covariance) solution is directly related to the SVD of $\mathbf{X}_c$:

$$\mathbf{Y} = \hat{\mathbf{U}}\hat{\mathbf{D}} \qquad \text{and} \qquad \mathbf{B} = \hat{\mathbf{V}}$$

Note that columns of $\hat{\mathbf{V}}$ are the...

- Right singular vectors of the mean-centered data matrix $\mathbf{X}_c$
- Eigenvectors of the covariance matrix $\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c = \hat{\mathbf{V}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}'$ where $\hat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$ with $\hat{\lambda}_k = \frac{\hat{d}_{kk}^2}{n-1}$ and $\hat{\mathbf{D}} = \mathrm{diag}(\hat{d}_{11}, \ldots, \hat{d}_{pp})$.

PCA Solution is Related to SVD (Correlation Matrix)

Let $\tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}'$ denote the SVD of $\mathbf{X}_{\mathrm{s}} = \mathbf{CXD}^{-1}$ with $\mathbf{D} = \mathrm{diag}(s_1, \ldots, s_p)$.

The PCA (correlation) solution is directly related to the SVD of $\mathbf{X}_{\mathrm{cs}}$:

$$\mathbf{Y} = \tilde{\mathbf{U}}\tilde{\mathbf{D}} \qquad \text{and} \qquad \mathbf{B} = \tilde{\mathbf{V}}$$

Note that columns of $\tilde{\mathbf{V}}$ are the...

- Right singular vectors of the centered and scaled data matrix $\mathbf{X}_{\mathrm{s}}$
- Eigenvectors of the correlation matrix $\mathbf{R} = \frac{1}{n-1}\mathbf{X}_{\mathrm{s}}'\mathbf{X}_{\mathrm{s}} = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}'$ where $\tilde{\mathbf{\Lambda}} = \mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p)$ with $\tilde{\lambda}_k = \frac{\tilde{d}_{kk}^2}{n-1}$ and $\tilde{\mathbf{D}} = \mathrm{diag}(\tilde{d}_{11}, \ldots, \tilde{d}_{pp})$.

Covariance versus Correlation Considerations

Problem: there is no simple relationship between SVDs of $\mathbf{X}_c$ and $\mathbf{X}_s$.

- No simple relationship between PCs obtained from **S** and **R**
- Rescaling variables can fundamentally change our results

Note that PCA is trying to explain the variation in **S** or **R**

- If units of p variables are comparable, covariance PCA may be more informative (because units of measurement are retained)
- If units of p variables are incomparable, correlation PCA may be more appropriate (because units of measurement are removed)

Men's Olympic Decathlon Data from 1988

Data from men's 1988 Olympic decathlon

- Total of $n = 34$ athletes
- Have $p = 10$ variables giving score for each decathlon event
- Have overall decathlon score also (`score`)

```
> decathlon[1:9,]
         run100 long.jump  shot high.jump run400 hurdle discus pole.vault javelin run1500 score
Schenk    11.25      7.43 15.48      2.27  48.90  15.13  49.28        4.7   61.32  268.95  8488
Voss      10.87      7.45 14.97      1.97  47.71  14.46  44.36        5.1   61.76  273.02  8399
Steen     11.18      7.44 14.20      1.97  48.29  14.81  43.66        5.2   64.16  263.20  8328
Thompson  10.62      7.38 15.02      2.03  49.06  14.72  44.80        4.9   64.04  285.11  8306
Blondel   11.02      7.43 12.92      1.97  47.44  14.40  41.20        5.2   57.46  256.64  8286
Plaziat   10.83      7.72 13.58      2.12  48.34  14.18  43.06        4.9   52.18  274.07  8272
Bright    11.18      7.05 14.12      2.06  49.34  14.39  41.68        5.7   61.60  291.20  8216
De.Wit    11.05      6.95 15.34      2.00  48.21  14.36  41.32        4.8   63.00  265.86  8189
Johnson   11.15      7.12 14.52      2.03  49.15  14.66  42.36        4.9   66.46  269.62  8180
```

Resigning Running Events

For the running events (`run100`, `run400`, `run1500`, and `hurdle`), lower scores correspond to better performance, whereas higher scores represent better performance for other events.

To make interpretation simpler, we will resign the running events:

```
> decathlon[,c(1,5,6,10)] <- (-1)*decathlon[,c(1,5,6,10)]
> decathlon[1:9,]
         run100 long.jump  shot high.jump run400 hurdle discus pole.vault javelin run1500 score
Schenk   -11.25      7.43 15.48      2.27 -48.90 -15.13  49.28        4.7   61.32 -268.95  8488
Voss     -10.87      7.45 14.97      1.97 -47.71 -14.46  44.36        5.1   61.76 -273.02  8399
Steen    -11.18      7.44 14.20      1.97 -48.29 -14.81  43.66        5.2   64.16 -263.20  8328
Thompson -10.62      7.38 15.02      2.03 -49.06 -14.72  44.80        4.9   64.04 -285.11  8306
Blondel  -11.02      7.43 12.92      1.97 -47.44 -14.40  41.20        5.2   57.46 -256.64  8286
Plaziat  -10.83      7.72 13.58      2.12 -48.34 -14.18  43.06        4.9   52.18 -274.07  8272
Bright   -11.18      7.05 14.12      2.06 -49.34 -14.39  41.68        5.7   61.60 -291.20  8216
De.Wit   -11.05      6.95 15.34      2.00 -48.21 -14.36  41.32        4.8   63.00 -265.86  8189
Johnson  -11.15      7.12 14.52      2.03 -49.15 -14.66  42.36        4.9   66.46 -269.62  8180
```

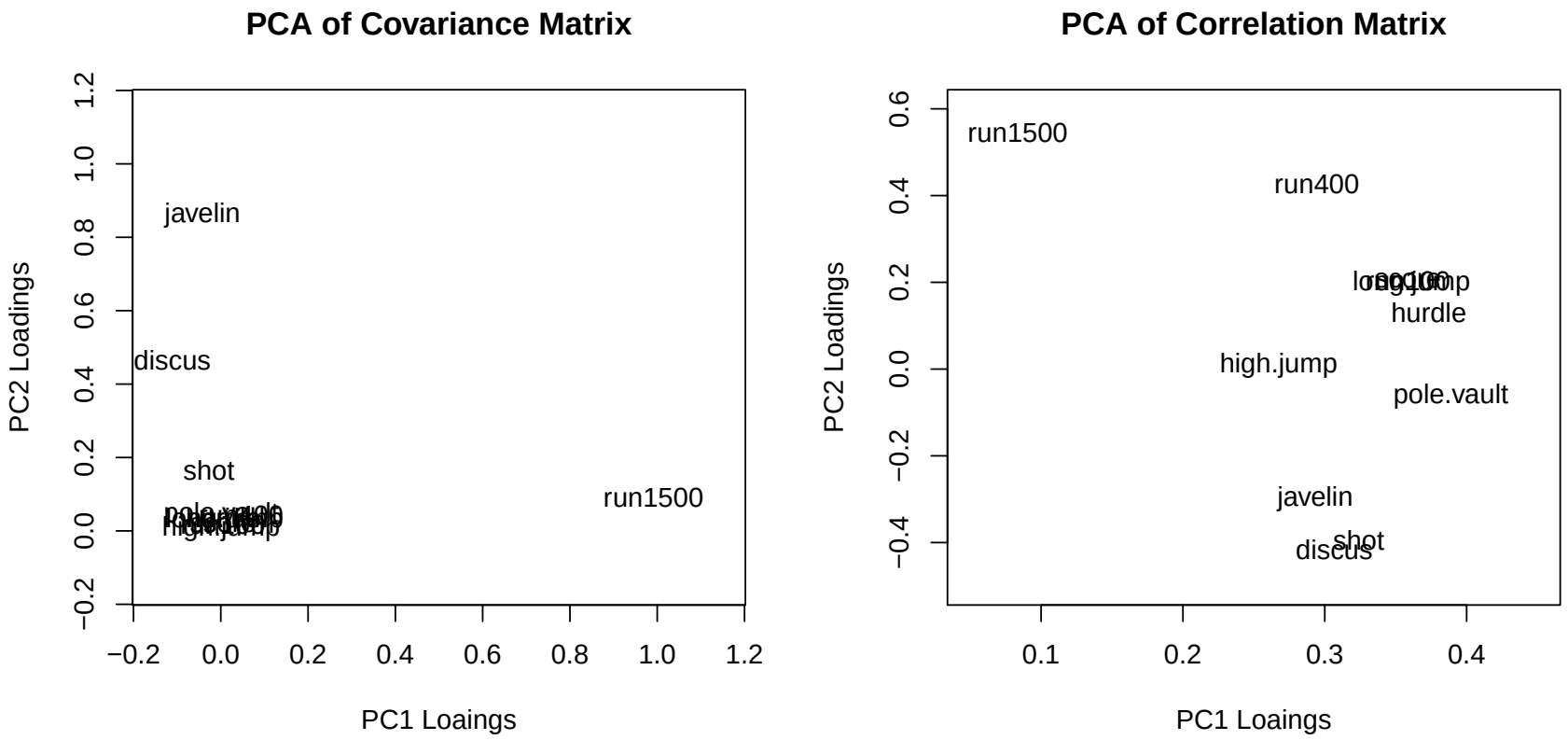
PCA on Covariance Matrix

```
# PCA on covariance matrix (default)
> pcaCOV <- princomp(x=decathlon[,1:10])
> names(pcaCOV)
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"
[6] "scores"   "call"

# resign PCA solution
> pcsign <- sign(colSums(pcaCOV$loadings^3))
> pcaCOV$loadings <- pcaCOV$loadings %*% diag(pcsign)
> pcaCOV$scores <- pcaCOV$scores %*% diag(pcsign)

# Note: R uses MLE of covariance matrix
> n <- nrow(decathlon)
> sum((pcaCOV$sdev - sqrt(eigen((n-1)/n*cov(decathlon[,1:10]))$values))^2)
[1] 8.861398e-28
```

PCA on Correlation Matrix

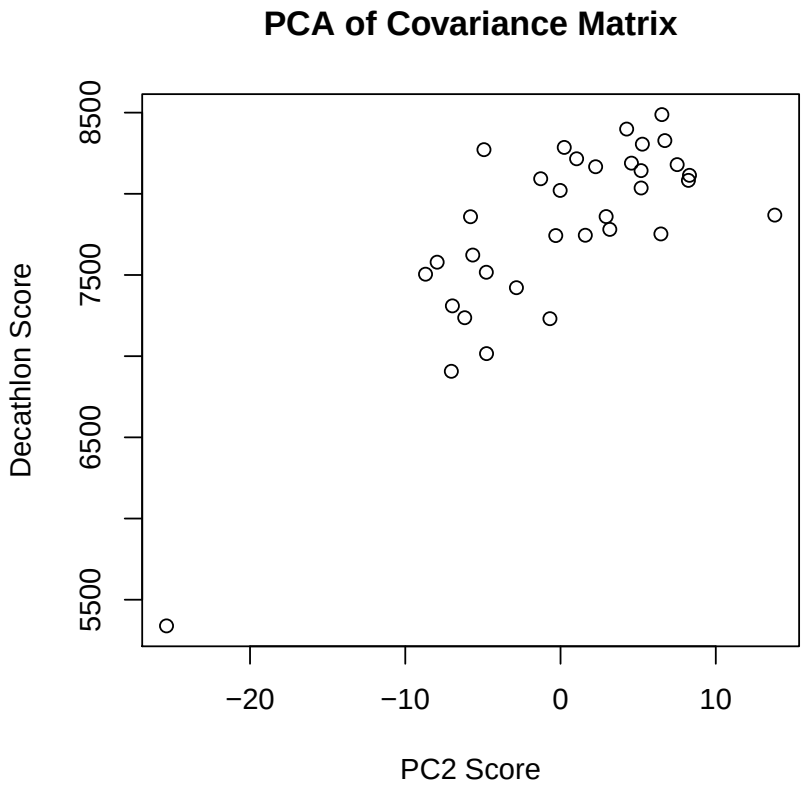
```
# PCA on correlation matrix
> pcaCOR <- princomp(x=decathlon[,1:10], cor=TRUE)
> names(pcaCOR)
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"
[6] "scores"   "call"

# resign PCA solution
> pcsign <- sign(colSums(pcaCOR$loadings^3))
> pcaCOR$loadings <- pcaCOR$loadings %*% diag(pcsign)
> pcaCOR$scores <- pcaCOR$scores %*% diag(pcsign)

# Note: PC standard deviations are sqrts correlation matrix eigenvalues
> sum((pcaCOR$sdev - sqrt(eigen(cor(decathlon[,1:10]))$values))^2)
[1] 2.249486e-30
```

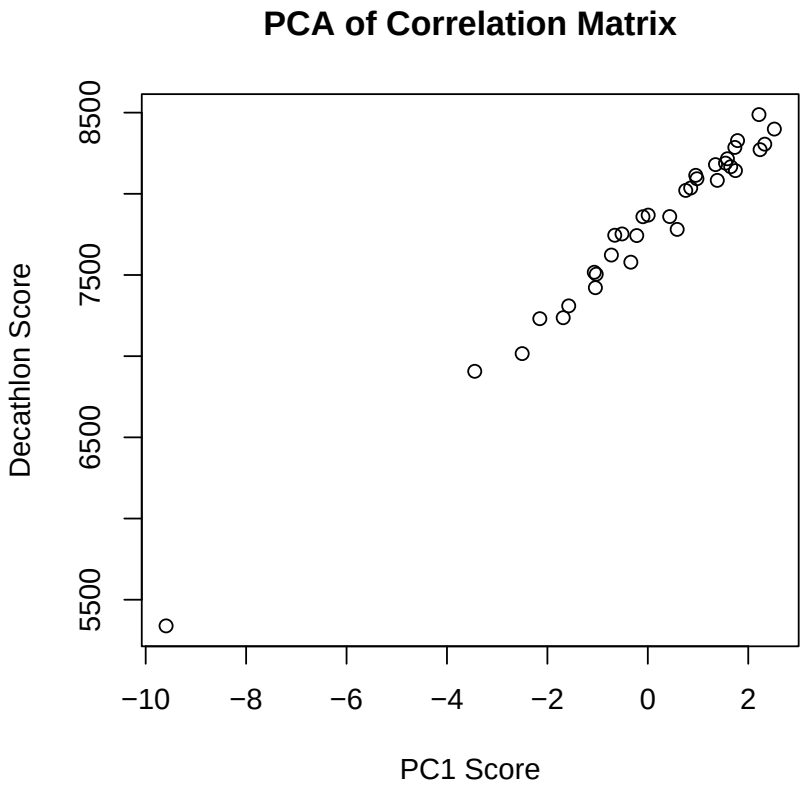
Plot Covariance and Correlation PCA Results




```
> dev.new(width=10, height=5, noRStudioGD=TRUE)
> par(mfrow=c(1,2))
> plot(pcaCOV$loadings[,1:2], xlab="PC1 Loaings", ylab="PC2 Loadings",
+      type="n", main="PCA of Covariance Matrix", xlim=c(-0.15, 1.15), ylim=c(-0.15, 1.15))
> text(pcaCOV$loadings[,1:2], labels=colnames(decathlon))
> plot(pcaCOR$loadings[,1:2], xlab="PC1 Loaings", ylab="PC2 Loadings",
+      type="n", main="PCA of Correlation Matrix", xlim=c(0.05, 0.45), ylim=c(-0.5,0.6))
> text(pcaCOR$loadings[,1:2], labels=colnames(decathlon))
```

Correlation of PC Scores and Overall Decathlon Score






```
> round(cor(decathlon$score, pcaCOV$scores),3)
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6] [,7]  [,8]  [,9] [,10]
[1,] 0.214 0.792 0.297 0.395 0.113 0.207 0.06 0.061 0.016 0.127
> round(cor(decathlon$score, pcaCOR$scores),3)
      [,1]  [,2]  [,3]  [,4]  [,5]   [,6]  [,7]   [,8]  [,9]  [,10]
[1,] 0.991 0.017 0.079 0.064 -0.04 -0.025 0.013 -0.009 0.011 -0.002
```

Dimensionality Problem

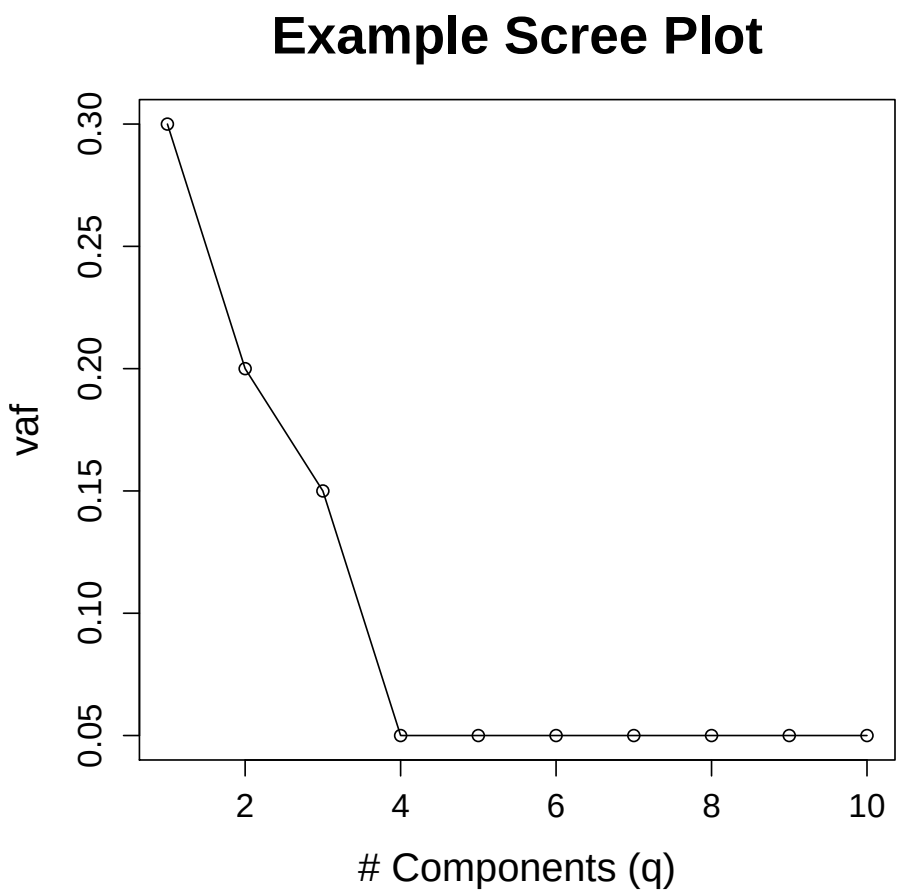
In practice, the optimal number of components is often unknown.

In some cases, possible/feasible values may be known a priori from theory and/or past research.

In other cases, we need to use some data-driven approach to select a reasonable number of components.

Scree Plots

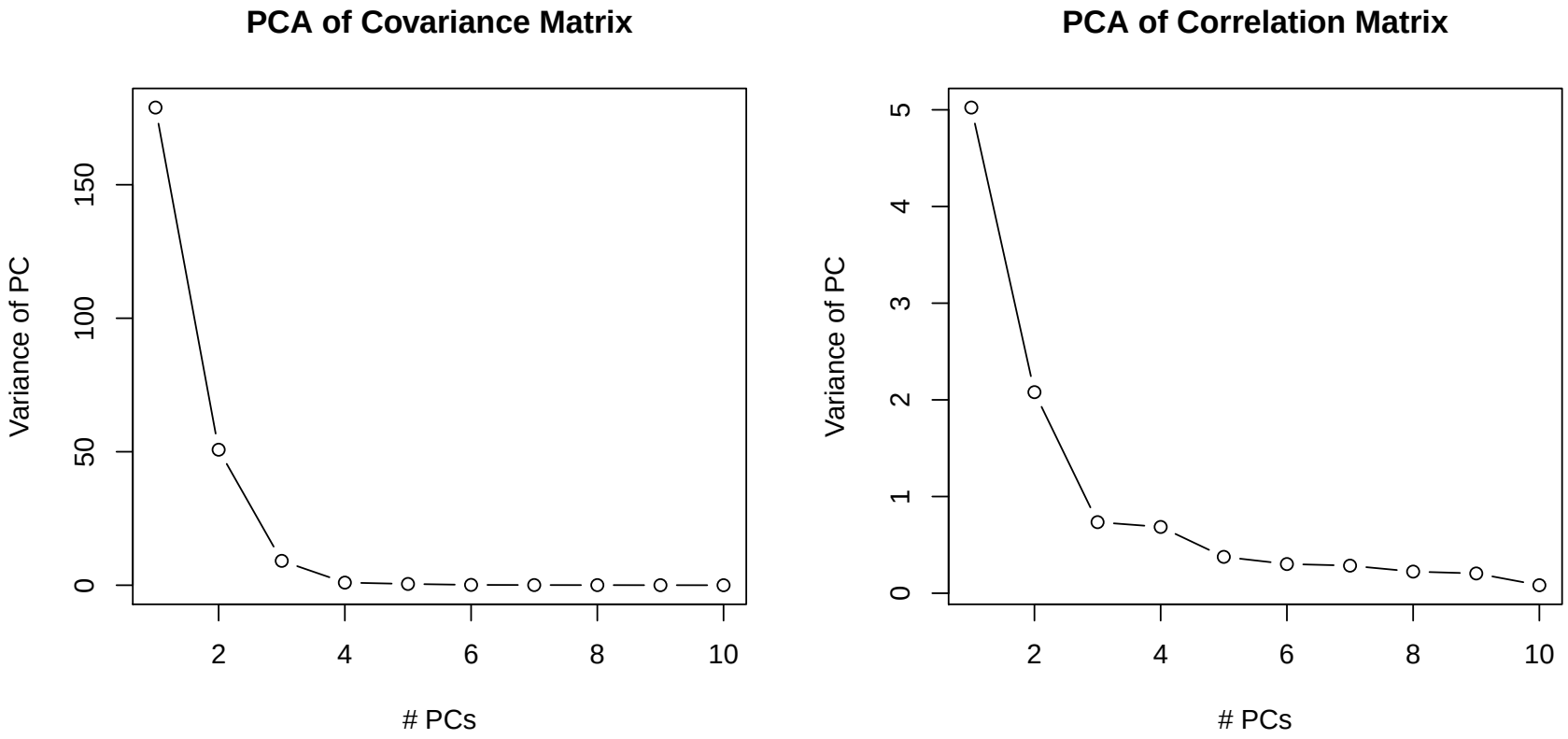
A scree plot displays the variance explained by each component.



We look for the "elbow" of the plot, i.e., point where line bends.

Could do formal test on derivative of scree line, but common sense approach often works fine.

Scree Plots For Decathlon PCA




```
> dev.new(width=10, height=5, noRStudioGD=TRUE)
> par(mfrow=c(1,2))
> plot(1:10, pcaCOV$sdev^2, type="b", xlab="# PCs", ylab="Variance of PC",
+      main="PCA of Covariance Matrix")
> plot(1:10, pcaCOR$sdev^2, type="b", xlab="# PCs", ylab="Variance of PC",
+      main="PCA of Correlation Matrix")
```