Chapter 12:

Principal Component Analysis

- §12.1 Introduction
- §12.2 Geometric and algebraic bases of principal components
- §12.4 Plotting of principal components
- §12.5 Principal components from the correlation matrix
- §12.6 Decide how many components to retain
- §12.8 Interpretation of principal components

12.1 Introduction

12.1 Introduction

- *Example:* We might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.

12.1 Introduction

- *Example:* We might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.
- In *Principal Component Analysis (PCA)*, we seek to maximize the variance of a linear combination of p variables from a single sample

$$z = \mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p$$

12.1 Introduction

- *Example:* We might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.
- In *Principal Component Analysis (PCA)*, we seek to maximize the variance of a linear combination of p variables from a single sample

$$z = \mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p$$

- Those linear combinations are called *principal components (PC)*. There are totally p such components.

12.1 Introduction

- *Example:* We might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.
- In *Principal Component Analysis (PCA)*, we seek to maximize the variance of a linear combination of p variables from a single sample

$$z = \mathbf{a}'\mathbf{y} = a_1y_1 + a_2y_2 + \cdots + a_py_p$$

- Those linear combinations are called *principal components (PC)*. There are totally p such components.
- The first PC is the linear combination with maximal variance, i.e., a dimension along which the observations are maximally spread out.

12.1 Introduction

- *Example:* We might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.
- In *Principal Component Analysis (PCA)*, we seek to maximize the variance of a linear combination of p variables from a single sample

$$z = \mathbf{a}'\mathbf{y} = a_1y_1 + a_2y_2 + \cdots + a_py_p$$

- Those linear combinations are called *principal components (PC)*. There are totally p such components.
- The first PC is the linear combination with maximal variance, i.e., a dimension along which the observations are maximally spread out.
- The second PC is the linear combination with maximal variance in a direction orthogonal to the first PC, and so on.

12.1 Introduction

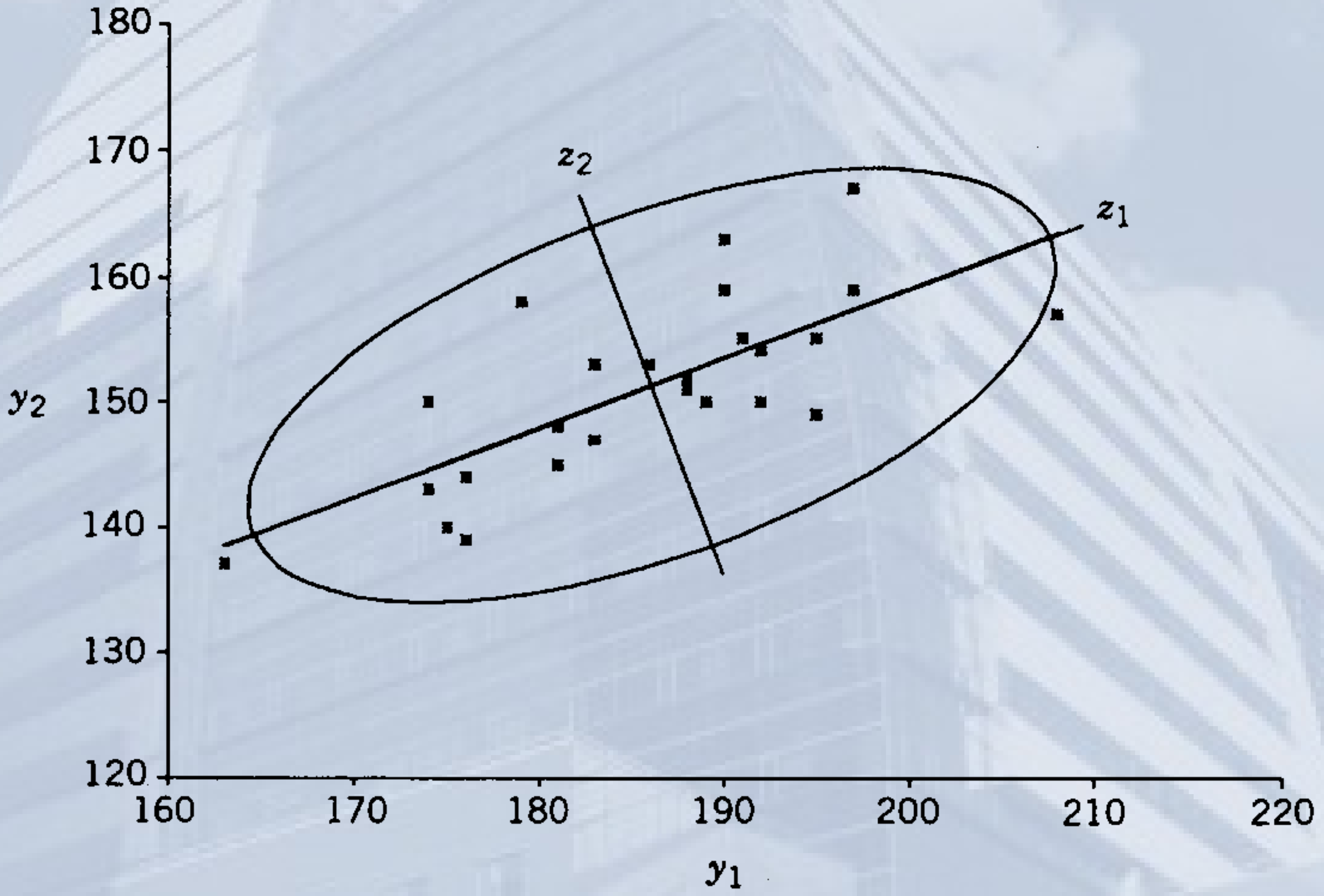

Figure 12.1. Principal component transformation for the sons data.

12.2.1 Geometric Approach

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.
- Center the data $\mathbf{y}_i - \bar{\mathbf{y}}$ to make $\bar{\mathbf{y}}_{new} = \mathbf{0}$ ($\bar{\mathbf{y}}$ at origin).

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.
- Center the data $\mathbf{y}_i - \bar{\mathbf{y}}$ to make $\bar{\mathbf{y}}_{new} = \mathbf{0}$ ($\bar{\mathbf{y}}$ at origin).
- Rotate axes to find a new coordinate system, where the data are maximally spread out along new directions.

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.
- Center the data $\mathbf{y}_i - \bar{\mathbf{y}}$ to make $\bar{\mathbf{y}}_{new} = \mathbf{0}$ ($\bar{\mathbf{y}}$ at origin).
- Rotate axes to find a new coordinate system, where the data are maximally spread out along new directions.
- For each observation $\mathbf{y}$, find p principal components

$$z_1 = \mathbf{a}_1'\mathbf{y}, \quad z_2 = \mathbf{a}_2'\mathbf{y}, \quad \ldots \quad , z_p = \mathbf{a}_p'\mathbf{y}$$

such that

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.
- Center the data $\mathbf{y}_i - \bar{\mathbf{y}}$ to make $\bar{\mathbf{y}}_{new} = \mathbf{0}$ ($\bar{\mathbf{y}}$ at origin).
- Rotate axes to find a new coordinate system, where the data are maximally spread out along new directions.
- For each observation $\mathbf{y}$, find p principal components

$$z_1 = \mathbf{a}_1'\mathbf{y}, \quad z_2 = \mathbf{a}_2'\mathbf{y}, \quad \ldots \quad , z_p = \mathbf{a}_p'\mathbf{y}$$

such that

- $\text{Var}(z_1) > \text{Var}(z_2) > \cdots > \text{Var}(z_p)$

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.
- Center the data $\mathbf{y}_i - \bar{\mathbf{y}}$ to make $\bar{\mathbf{y}}_{new} = \mathbf{0}$ ($\bar{\mathbf{y}}$ at origin).
- Rotate axes to find a new coordinate system, where the data are maximally spread out along new directions.
- For each observation $\mathbf{y}$, find p principal components

$$z_1 = \mathbf{a}_1'\mathbf{y}, \quad z_2 = \mathbf{a}_2'\mathbf{y}, \quad \ldots \quad , z_p = \mathbf{a}_p'\mathbf{y}$$

  such that

  - $\text{Var}(z_1) > \text{Var}(z_2) > \cdots > \text{Var}(z_p)$
  - $\text{Cov}(z_i, z_j) = 0$ for $i \neq j$

12.2.1 Geometric Approach

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a p-dimmensional multivariate sample.
- Center the data $\mathbf{y}_i - \bar{\mathbf{y}}$ to make $\bar{\mathbf{y}}_{new} = \mathbf{0}$ ($\bar{\mathbf{y}}$ at origin).
- Rotate axes to find a new coordinate system, where the data are maximally spread out along new directions.
- For each observation $\mathbf{y}$, find p principal components

$$z_1 = \mathbf{a}_1'\mathbf{y}, \quad z_2 = \mathbf{a}_2'\mathbf{y}, \quad \ldots \quad , z_p = \mathbf{a}_p'\mathbf{y}$$

  such that

  - $\text{Var}(z_1) > \text{Var}(z_2) > \cdots > \text{Var}(z_p)$
  - $\text{Cov}(z_i, z_j) = 0$ for $i \neq j$

- The p PCs are uncorrelated, but not necessarily independent (need normality assumption).

12.2.1 Geometric Approach

12.2.1 Geometric Approach

- Let **S** be the single sample covariance matrix.

12.2.1 Geometric Approach

- Let $\mathbf{S}$ be the single sample covariance matrix.
- Recall eigen-decomposition: $\mathbf{S} = \mathbf{CDC}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ and $\mathbf{C}$ is an orthogonal matrix of corresponding eigenvectors.

12.2.1 Geometric Approach

- Let $\mathbf{S}$ be the single sample covariance matrix.
- Recall eigen-decomposition: $\mathbf{S} = \mathbf{C}\mathbf{D}\mathbf{C}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ and $\mathbf{C}$ is an orthogonal matrix of corresponding eigenvectors.
- We take $\mathbf{a}_i$ to be the ith eigenvector of $\mathbf{S}$, then

$$\begin{aligned} \mathrm{Var}(z_i) &= \mathrm{Var}(\mathbf{a}_i'\mathbf{y}) = \lambda_i \\ \mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}(\mathbf{a}_i'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = 0 \end{aligned}$$

12.2.1 Geometric Approach

- Let $\mathbf{S}$ be the single sample covariance matrix.
- Recall eigen-decomposition: $\mathbf{S} = \mathbf{CDC}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ and $\mathbf{C}$ is an orthogonal matrix of corresponding eigenvectors.
- We take $\mathbf{a}_i$ to be the ith eigenvector of $\mathbf{S}$, then

$$\begin{aligned} \mathrm{Var}(z_i) &= \mathrm{Var}(\mathbf{a}_i'\mathbf{y}) = \lambda_i \\ \mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}(\mathbf{a}_i'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = 0 \end{aligned}$$

- The z_1 has the largest variance λ_1, followed by z_2, z_3, etc.

12.2.1 Geometric Approach

- Let $\mathbf{S}$ be the single sample covariance matrix.
- Recall eigen-decomposition: $\mathbf{S} = \mathbf{CDC}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ and $\mathbf{C}$ is an orthogonal matrix of corresponding eigenvectors.
- We take $\mathbf{a}_i$ to be the ith eigenvector of $\mathbf{S}$, then

$$\begin{aligned} \mathrm{Var}(z_i) &= \mathrm{Var}(\mathbf{a}_i'\mathbf{y}) = \lambda_i \\ \mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}(\mathbf{a}_i'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = 0 \end{aligned}$$

- The z_1 has the largest variance λ_1, followed by z_2, z_3, etc.
- We represent the p-dimmentional data with first k principal components that account for a large proportion of the total variation.

12.2.1 Geometric Approach

- Let $\mathbf{S}$ be the single sample covariance matrix.
- Recall eigen-decomposition: $\mathbf{S} = \mathbf{CDC}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ and $\mathbf{C}$ is an orthogonal matrix of corresponding eigenvectors.
- We take $\mathbf{a}_i$ to be the ith eigenvector of $\mathbf{S}$, then

$$\begin{aligned} \mathrm{Var}(z_i) &= \mathrm{Var}(\mathbf{a}_i'\mathbf{y}) = \lambda_i \\ \mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}(\mathbf{a}_i'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = 0 \end{aligned}$$

- The z_1 has the largest variance λ_1, followed by z_2, z_3, etc.
- We represent the p-dimmentional data with first k principal components that account for a large proportion of the total variation.
- PCA can be applied to any distribution of $\mathbf{y}$.

12.2.1 Geometric Approach

- Let $\mathbf{S}$ be the single sample covariance matrix.
- Recall eigen-decomposition: $\mathbf{S} = \mathbf{CDC}'$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ and $\mathbf{C}$ is an orthogonal matrix of corresponding eigenvectors.
- We take $\mathbf{a}_i$ to be the ith eigenvector of $\mathbf{S}$, then

$$\begin{aligned} \mathrm{Var}(z_i) &= \mathrm{Var}(\mathbf{a}_i'\mathbf{y}) = \lambda_i \\ \mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}(\mathbf{a}_i'\mathbf{y}, \mathbf{a}_j'\mathbf{y}) = 0 \end{aligned}$$

- The z_1 has the largest variance λ_1, followed by z_2, z_3, etc.
- We represent the p-dimmentional data with first k principal components that account for a large proportion of the total variation.
- PCA can be applied to any distribution of $\mathbf{y}$.
- Show Example 12.2.1 (p.384).

12.2.2 Algebraic Approach

12.2.2 Algebraic Approach

- The variance of a PC is

$$\text{Var}(z) = \text{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

12.2.2 Algebraic Approach

- The variance of a PC is

$$\text{Var}(z) = \text{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

- The $z_1 = \mathbf{a}_1'\mathbf{y}$ can be obtained by maximizing

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}.$$

12.2.2 Algebraic Approach

- The variance of a PC is

$$\text{Var}(z) = \text{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

- The $z_1 = \mathbf{a}_1'\mathbf{y}$ can be obtained by maximizing

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}.$$

- It can be shown that maximum λ = largest eigenvalue of $\mathbf{S}$, and $\mathbf{a}_1$ is the corresponding eigenvector.

12.2.2 Algebraic Approach

- The variance of a PC is

$$\text{Var}(z) = \text{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

- The $z_1 = \mathbf{a}_1'\mathbf{y}$ can be obtained by maximizing

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}.$$

- It can be shown that maximum λ = largest eigenvalue of $\mathbf{S}$, and $\mathbf{a}_1$ is the corresponding eigenvector.
- Unlike discriminant analysis or canonical correlation, $\mathbf{S}$ can be singular ($\mathbf{S}^{-1}$ does not exist) in PCA, and zero eigenvalues and their eigenvectors can be ignored.

12.2.2 Algebraic Approach

- The variance of a PC is

$$\mathsf{Var}(z) = \mathsf{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

- The $z_1 = \mathbf{a}_1'\mathbf{y}$ can be obtained by maximizing

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}.$$

- It can be shown that maximum $\lambda =$ largest eigenvalue of $\mathbf{S}$, and $\mathbf{a}_1$ is the corresponding eigenvector.
- Unlike discriminant analysis or canonical correlation, $\mathbf{S}$ can be singular ($\mathbf{S}^{-1}$ does not exist) in PCA, and zero eigenvalues and their eigenvectors can be ignored.
- Show Example 12.2.2 (p.386)

12.5 PCA from Correlation Matrix R

Motivation

12.5 PCA from Correlation Matrix $\mathbf{R}$

Motivation

- Generally, extracting components from $\mathbf{S}$ rather than $\mathbf{R}$ remains closer to the spirit and intent of PCA, especially if the components are to be used in further computations.

12.5 PCA from Correlation Matrix $\mathbf{R}$

Motivation

- Generally, extracting components from $\mathbf{S}$ rather than $\mathbf{R}$ remains closer to the spirit and intent of PCA, especially if the components are to be used in further computations.
- In some cases, the PCs will be more interpretable if $\mathbf{R}$ is used.

12.5 PCA from Correlation Matrix R

Motivation

- Generally, extracting components from **S** rather than **R** remains closer to the spirit and intent of PCA, especially if the components are to be used in further computations.
- In some cases, the PCs will be more interpretable if **R** is used.
- For example, if the variances differ widely or if the measurement units are not commensurate, the PCs of **S** will be dominated by the variables with large variances. The other variables will contribute very little. For a more balanced representation in such cases, the PCs of **R** may be used.

12.5 PCA from Correlation Matrix R

Properties

12.5 PCA from Correlation Matrix R

Properties

- The **R** has different eigenvalues than **S**.

12.5 PCA from Correlation Matrix R

Properties

- The **R** has different eigenvalues than **S**.
- The PCs from **R** differ from those obtained from **S**.

12.5 PCA from Correlation Matrix R

Properties

- The **R** has different eigenvalues than **S**.
- The PCs from **R** differ from those obtained from **S**.
- The percent of variance accounted for by each PC of **R** will differ from the percent for **S**.

12.5 PCA from Correlation Matrix R

Properties

- The **R** has different eigenvalues than **S**.
- The PCs from **R** differ from those obtained from **S**.
- The percent of variance accounted for by each PC of **R** will differ from the percent for **S**.
- The PCs from **R** are scale invariant.

12.5 PCA from Correlation Matrix R

Properties

- The **R** has different eigenvalues than **S**.
- The PCs from **R** differ from those obtained from **S**.
- The percent of variance accounted for by each PC of **R** will differ from the percent for **S**.
- The PCs from **R** are scale invariant.
- Different **R** matrices may produce the same PCs.

12.5 PCA from Correlation Matrix R

Properties

- The **R** has different eigenvalues than **S**.
- The PCs from **R** differ from those obtained from **S**.
- The percent of variance accounted for by each PC of **R** will differ from the percent for **S**.
- The PCs from **R** are scale invariant.
- Different **R** matrices may produce the same PCs.
- Show example on page 397.

12.6 How many components to keep?

12.6 How many components to keep?

- Retain as many components as needed to achieve a pre-specified percentage of variance, say 80%.

12.6 How many components to keep?

- Retain as many components as needed to achieve a pre-specified percentage of variance, say 80%.
- Retain components with eigenvalues above the average.

12.6 How many components to keep?

- Retain as many components as needed to achieve a pre-specified percentage of variance, say 80%.
- Retain components with eigenvalues above the average.
- Use *scree graph*, a plot of λ_i versus i, and look for a natural break between the ÒlargeÓ eigenvalues and the ÒsmallÓ eigenvalues.

12.6 How many components to keep?

- Retain as many components as needed to achieve a pre-specified percentage of variance, say 80%.
- Retain components with eigenvalues above the average.
- Use *scree graph*, a plot of λ_i versus i, and look for a natural break between the ÒlargeÓ eigenvalues and the ÒsmallÓ eigenvalues.

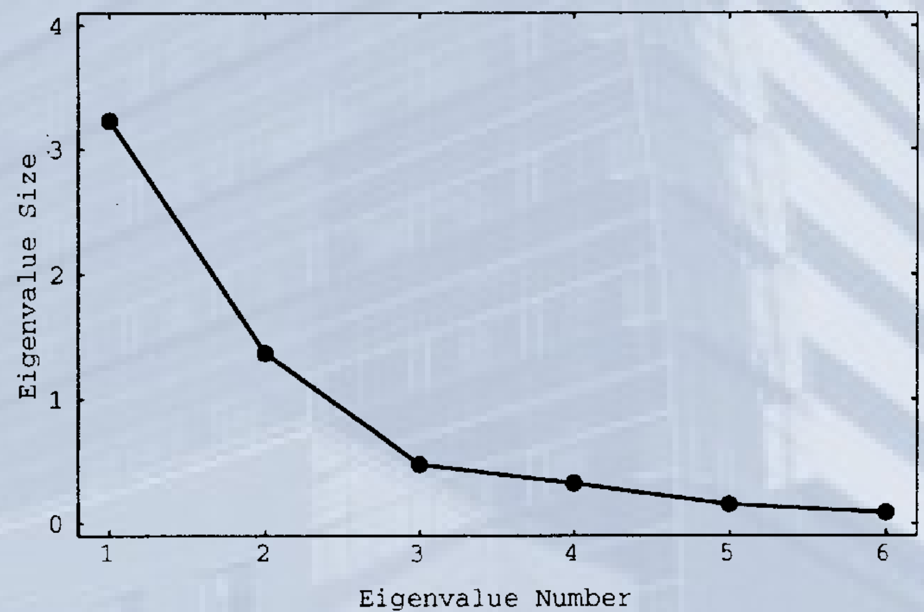

Figure 12.8. Scree graph for eigenvalues of modified football data.

12.6 How many components to keep?

- Retain as many components as needed to achieve a pre-specified percentage of variance, say 80%.
- Retain components with eigenvalues above the average.
- Use *scree graph*, a plot of λ_i versus i, and look for a natural break between the ÒlargeÓ eigenvalues and the ÒsmallÓ eigenvalues.

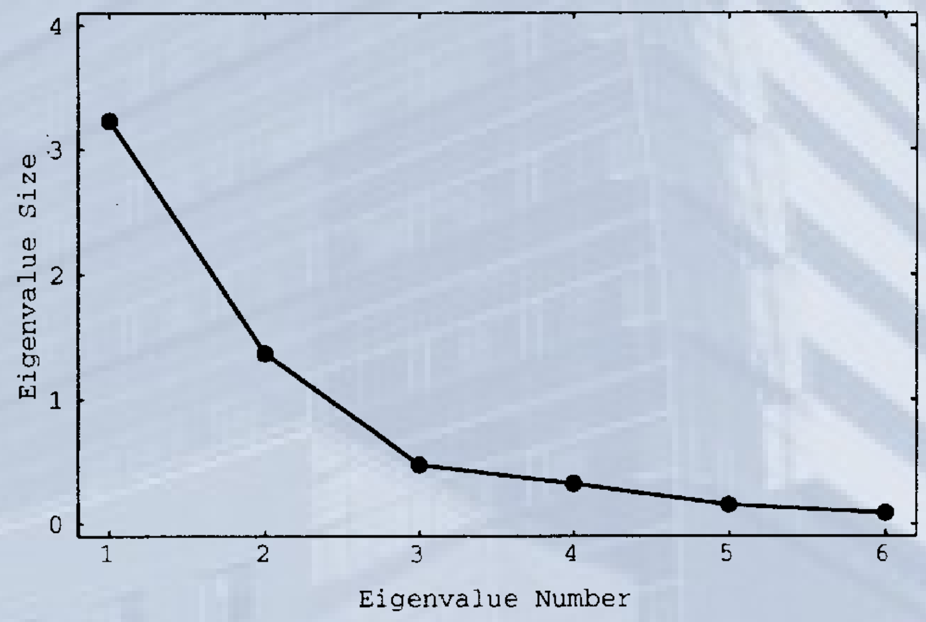

Figure 12.8. Scree graph for eigenvalues of modified football data.

- Show Example 12.6 (p.400).

Interpretation of PC

Interpretation of PC

- The PC's from $\mathbf{S}$ and $\mathbf{R}$ are NOT compatible, and will have different interpretations.

Interpretation of PC

- The PC's from $\mathbf{S}$ and $\mathbf{R}$ are NOT compatible, and will have different interpretations.
- The absolute value of a coefficient in PC shows the contribution of corresponding variable.

Interpretation of PC

- The PC's from $\mathbf{S}$ and $\mathbf{R}$ are NOT compatible, and will have different interpretations.
- The absolute value of a coefficient in PC shows the contribution of corresponding variable.
- If one variable has a much larger variance than other variables, this variable will dominate the first component, which will account for most of the variance.

Interpretation of PC

- The PC's from $\mathbf{S}$ and $\mathbf{R}$ are NOT compatible, and will have different interpretations.
- The absolute value of a coefficient in PC shows the contribution of corresponding variable.
- If one variable has a much larger variance than other variables, this variable will dominate the first component, which will account for most of the variance.
- A component will duplicate a variable when the variable is uncorrelated with the other variables, i.e., $z_i = \mathbf{a}_i'\mathbf{y} = y_i,\ i = 1, \ldots, p$.

Interpretation of PC

- The PC's from **S** and **R** are NOT compatible, and will have different interpretations.
- The absolute value of a coefficient in PC shows the contribution of corresponding variable.
- If one variable has a much larger variance than other variables, this variable will dominate the first component, which will account for most of the variance.
- A component will duplicate a variable when the variable is uncorrelated with the other variables, i.e., $z_i = \mathbf{a}_i'\mathbf{y} = y_i,\ i = 1, \ldots, p$.
- All elements of the first eigenvector $\mathbf{a}_1$ are positive if all correlations or covariances are positive.

Interpretation of PC

- The PC's from **S** and **R** are NOT compatible, and will have different interpretations.
- The absolute value of a coefficient in PC shows the contribution of corresponding variable.
- If one variable has a much larger variance than other variables, this variable will dominate the first component, which will account for most of the variance.
- A component will duplicate a variable when the variable is uncorrelated with the other variables, i.e., $z_i = \mathbf{a}_i' \mathbf{y} = y_i, \; i = 1, \ldots, p$.
- All elements of the first eigenvector $\mathbf{a}_1$ are positive if all correlations or covariances are positive.
- Show Example 12.8.1 (p.403).