Principal Component Analysis

- When a very large number of variables is measured on each subject, interpreting the data may be difficult.

- It is often possible to reduce the dimensionality of the data by finding a smaller set of linear combinations of the variables that preserve most of the variability across subjects.

- These linear combinations are called *principal components*.

- Finding the principal components is often an early step in a more complex analysis (e.g., regression, clustering, classification).

Objectives

- **Reduce Dimensionality:** Instead of analyzing variation in a large number, say p, variables as they vary from subject to subject, analyze variation in a much smaller number of principal component scores.

- **Develop Summary Indices:** Find meaningful, or useful, linear combinations of the original variables, such as food quality, consumer satisfaction, or economic indices.

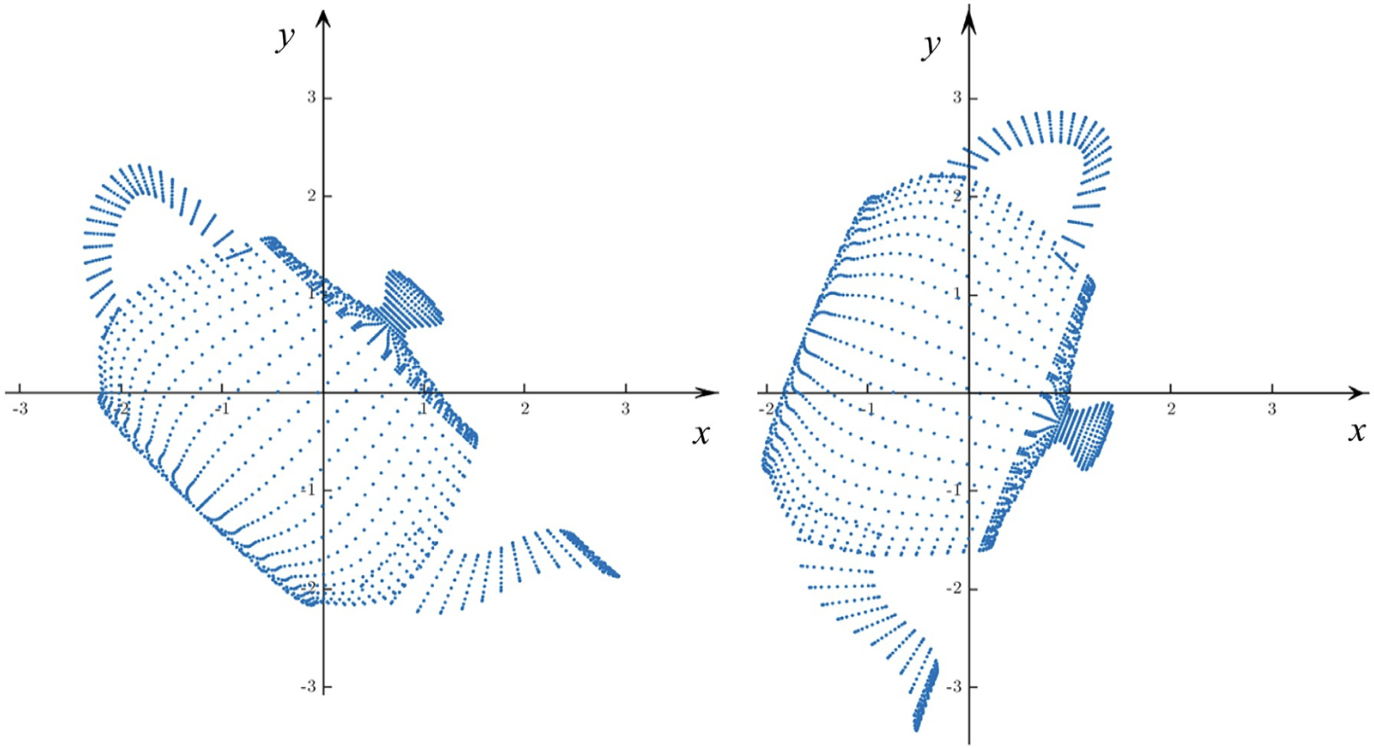
- **Cluster Analysis:** Visually display differences between groups or clusters.

- **Data Screening:** Detect outliers (extreme data vectors) or strong associations among variables

Principal Component Analysis

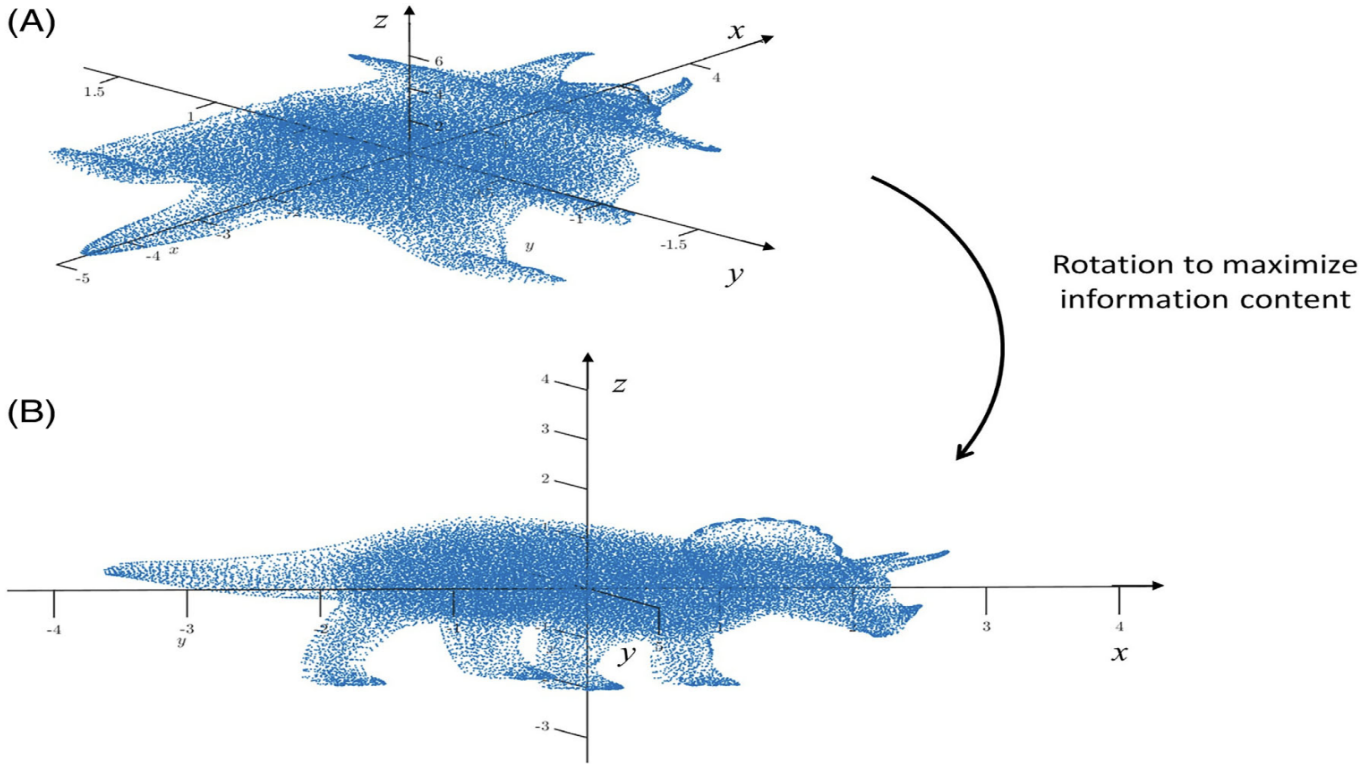
- Intuitive illustration of PCA
- Theory of PCA
- PCA in practice

Motivating Examples (Saccenti 2023)



- The left figure is based on a set of 4608 observations of two variables.
- Right figure is a 65° rotation of the left figure.
- What is the *structure* of the data set?

Motivating Examples (Saccenti 2023)



- The figure is based on a set of 36876 observations of three variables.
- How can we visualize/detect the structure of high-dimensional data?

Intuition: Rotation

- Let (x, y) be a point on the x-y coordinate system.
- Rotating the x-y plane by angle θ yields a new coordinate system, say x'-y' plane, with the relationship

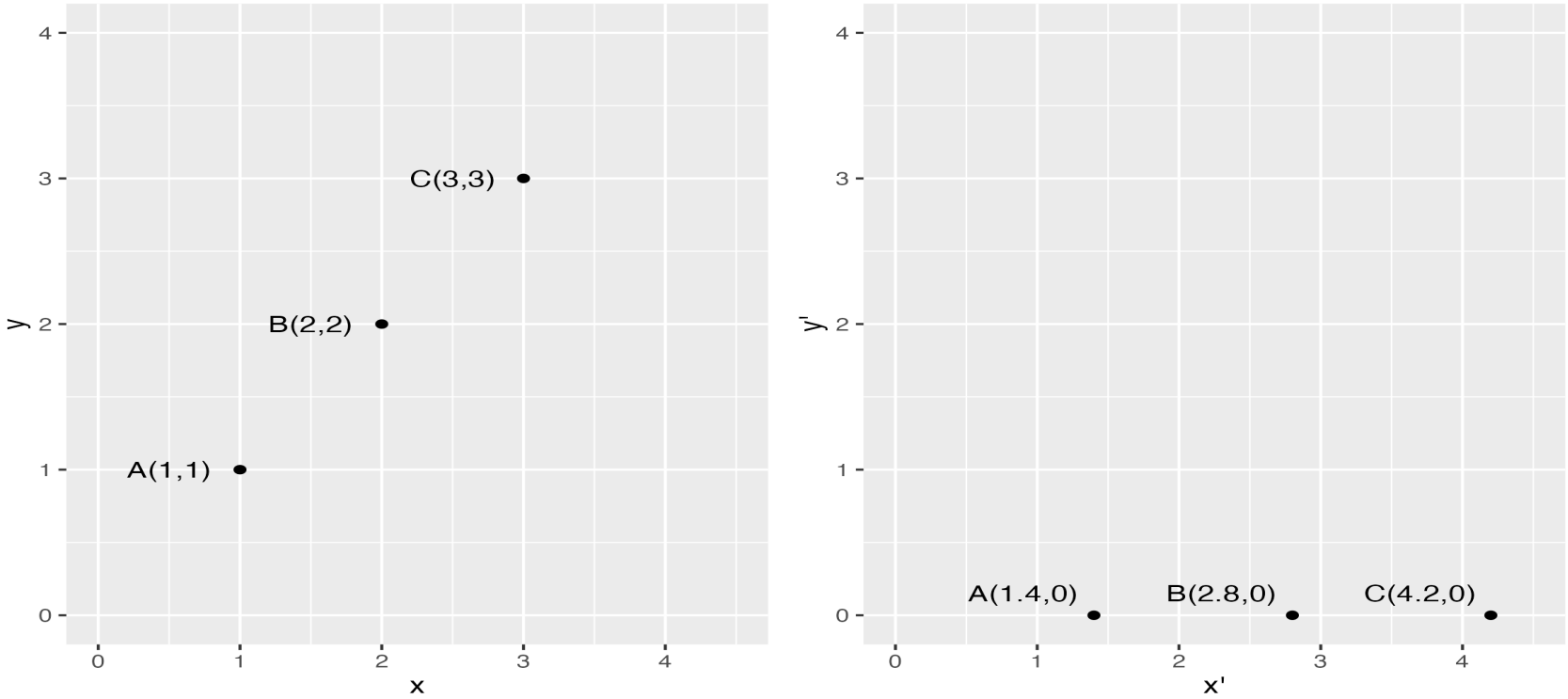
$$x' = x\cos\theta - y\sin\theta \quad \text{and} \quad y' = x\sin\theta + y\cos\theta;$$

  or in matrix notation,

$$\begin{pmatrix} x' & y' \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

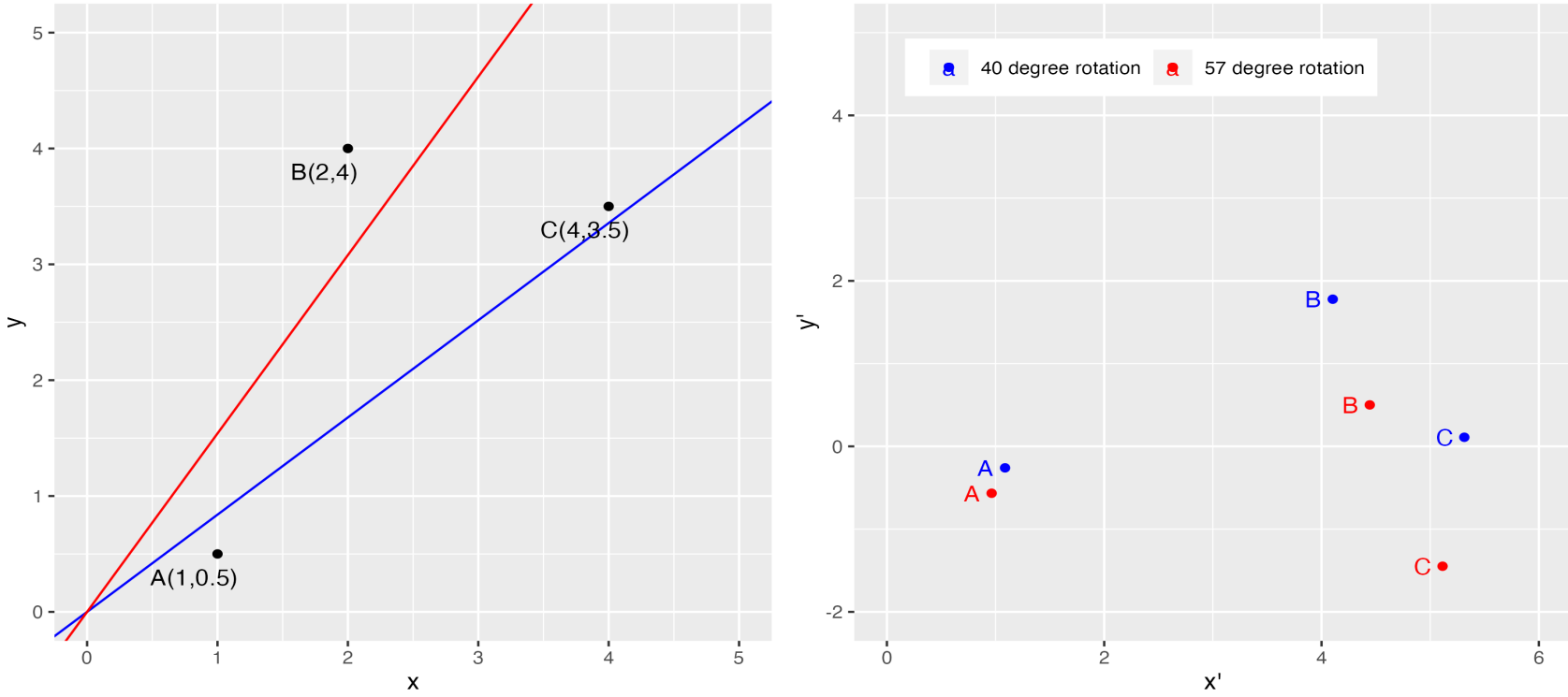
- Given three points $A = (1, 1), B = (2, 2), C = (3, 3)$, rotating 45° (counter clockwise) yields new coordinates $A' = (\sqrt{2} \approx 1.4, 0), B' = (2\sqrt{2} \approx 2.8, 0), C' = (3\sqrt{2} \approx 4.2, 0)$.

Intuition: Rotation



- 45° line is the direction with the maximum variance in the data.
- Total variance, variances explained?

Intuition: Rotation



- Rotating 40°: 78.3% of total variance explained along the x' direction.
- Rotating 57°: 83.3% of total variance explained.
- How to determine the direction of maximum variability in high-dimensional setting?

Population Principal Components

- Let the random vector $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ have covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

- Consider p linear combinations of the variables

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

- Note that

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

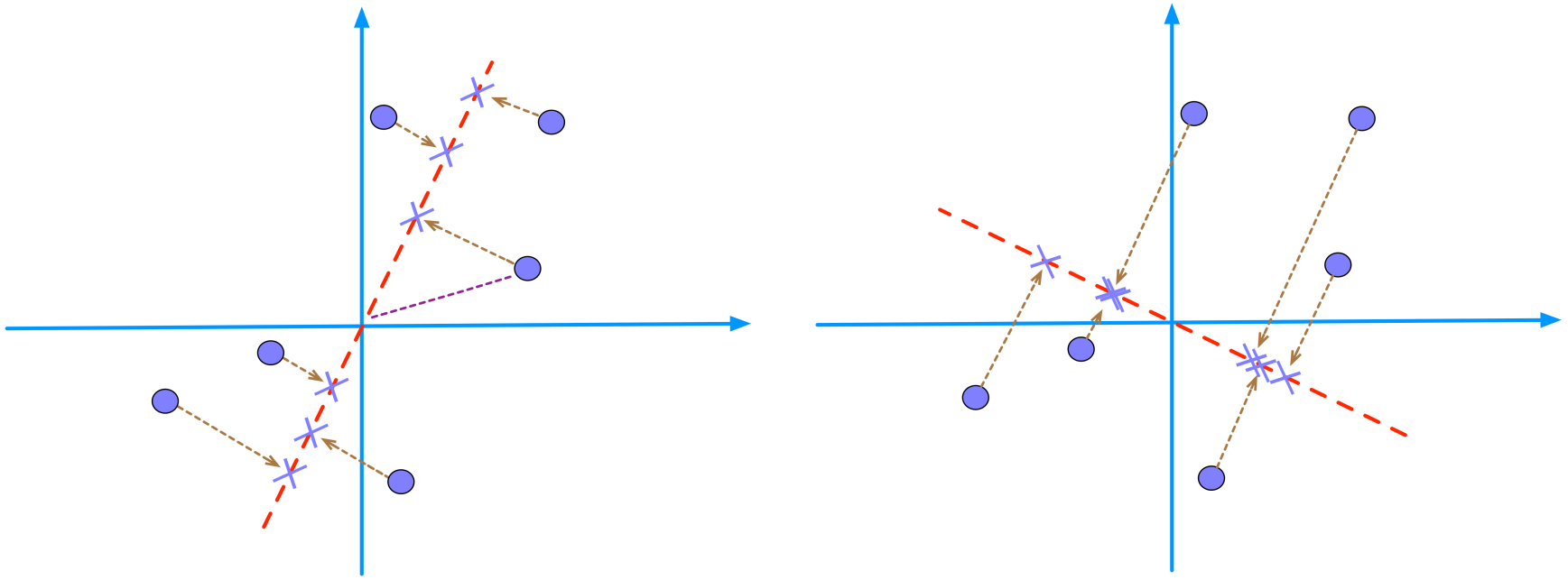
Population Principal Components

- *Principal components* (PC) are the linear combinations determined sequentially as follows:
  - The first PC is $Y_1 := \mathbf{a}_1'\mathbf{X}$ that maximizes $\mathrm{Var}(Y_1) = \mathbf{a}_1'\Sigma\mathbf{a}_1$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$.
  - The second PC is $Y_2 := \mathbf{a}_2'\mathbf{X}$ that maximizes $\mathrm{Var}(Y_2) = \mathbf{a}_2'\Sigma\mathbf{a}_2$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $\mathrm{Cov}(Y_1, Y_2) = \mathbf{a}_1'\Sigma\mathbf{a}_2 = 0$.
    ⋮
  - The ith PC $\mathbf{a}_i$ is $Y_i := \mathbf{a}_i'\mathbf{X}$ that maximizes $\mathrm{Var}(Y_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i$ subject to $\mathbf{a}_i'\mathbf{a}_i = 1$ and $\mathrm{Cov}(Y_i, Y_k) = \mathbf{a}_i'\Sigma\mathbf{a}_k = 0$ for $k < i$.
    ⋮

Some terminologies

- Sometimes the weight vectors $\mathbf{a}_i = (a_{i1}, \cdots, a_{ip})'$ are referred to as "principal components." (Chapter 1 of Jolliffe 2002)
  - We will not use this naming convention in this course.
- The weight vector $\mathbf{a}_i$ corresponding to the ith PC (say 1st PC) will be called *loadings*, which indicate how the old variables are weighted to get the new ones.
  - a_{ik} is the kth component of the ith loading for the ith PC that measures the variable importance.

Geometric Interpretation



- PCA minimizes the distances from the data points to the (red) line.
- Equivalently, PCA maximizes the distances from the projected points to the origin.
- The average of the sum of squared distances is called the *eigenvalue* for the PC.
- The square-root of the sum of squared distances is called the *singular value* for the PC.
- Interpretation of the PC (eigenvector, singular vector)?

Population Principal Components

- Let Σ have eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then, the ith principal component is given by

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

- Then,

$$\text{Var}(Y_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i = \mathbf{e}_i'\lambda_i\mathbf{e}_i = \lambda_i\mathbf{e}_i'\mathbf{e}_i = \lambda_i, \text{ since } \mathbf{e}_i'\mathbf{e}_i = 1$$

$$\text{Cov}(Y_i, Y_k) = \mathbf{e}_i'\Sigma\mathbf{e}_k = \mathbf{e}_i'\lambda_k\mathbf{e}_k = 0, \text{ since } \mathbf{e}_i'\mathbf{e}_k = 0$$

- Using properties of the trace of Σ we have

$$\sigma_{11}+\sigma_{22}+\cdots+\sigma_{pp} = \sum_i \text{Var}(X_i) = \lambda_1+\lambda_2+\cdots+\lambda_p = \sum_i \text{Var}(Y_i).$$

Population Principal Components

- Because the total population variance, $trace(\Sigma)$, is equal to the sum of the variances of the principal components, $\sum_i \lambda_i$, we say that

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

  is the proportion of the total variance associated with (or explained by) the kth principal component.

- If a large proportion of the total variance (say 80% or 90%) is explained by the first k PCs, then we can ignore the original p variables and restrict attention to the first k PCs without much loss of information about variation among members of the population.

Spectral Decomposition of a Covariance Matrix

Mathematically any covariance (or correlation) matrix can be expressed as

$$\Sigma = E\Lambda E'$$

where

$$E = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{p1} \\ e_{12} & e_{22} & \cdots & e_{p2} \\ \vdots & \vdots & & \vdots \\ e_{1p} & e_{2p} & \cdots & e_{pp} \end{bmatrix} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \ldots \ \mathbf{e}_p]$$

is the matrix with eigenvectors as the columns and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}$$

is a diagonal matrix of eigenvalues.

Spectral Decomposition of a Covariance Matrix

The spectral decomposition is also expressed as

$$\Sigma = E\Lambda E' = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^{'} + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^{'} + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p^{'}$$

If the first k eigenvalues account for a large portion of the total variance, the covariance matrix (or correlation matrix) can be well approximated by first k terms in the decomposition, i.e.,

$$\Sigma = E\Lambda E' \approx \lambda_1 \mathbf{e}_1 \mathbf{e}_1^{'} + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^{'} + \cdots + \lambda_k \mathbf{e}_k \mathbf{e}_k^{'}$$

It is desirable to have "k' much smaller than the original number of variables p.

Principal Component Scores

- Suppose that we have $n \times p$ random matrix $\mathbf{X} := [\mathbf{X}_1, \ldots, \mathbf{X}_n]'$, or

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & & X_{np} \end{bmatrix}.$$

  In practice, the data matrix is a realization of the random matrix $\mathbf{X}$.

- Each $\mathbf{X}_i$ has covariance matrix Σ and mean $\boldsymbol{\mu} := (\mu_1, \mu_2, \cdots, \mu_p)'$.
- The representation of the original data in the PC space (i.e., the rotated data) are called *scores*.
- Principal component scores are generally centered at zero.

Principal Component Scores

- The centered score of the k-th principal component for the i-th member of the population is

$$Y_{ik} = e_{k1}(X_{i1} - \mu_1) + e_{k2}(X_{i2} - \mu_2) + \cdots + e_{kp}(X_{ip} - \mu_p)$$

- The expected value of this centered principal component score is zero.

- The population variance of the scores for the i-th principal component is λ_i, the i-th largest eigenvalue.

- The principal component scores are interpreted by understanding what low and high scores represent. This is determined by looking at the signs and relative sizes of the coefficients.

Principal Component Scores

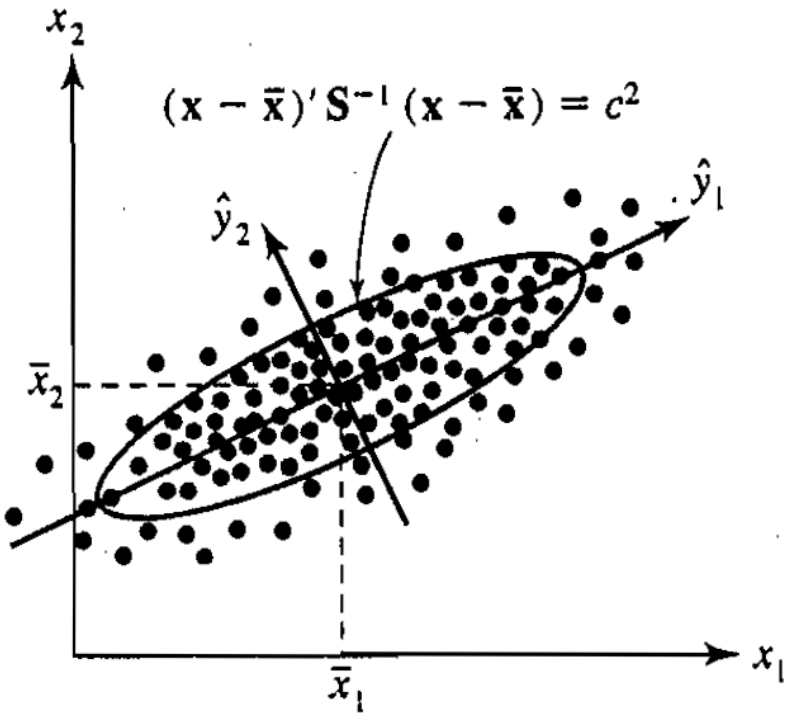
- The correlation between the ith principal component and the kth original variable,

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

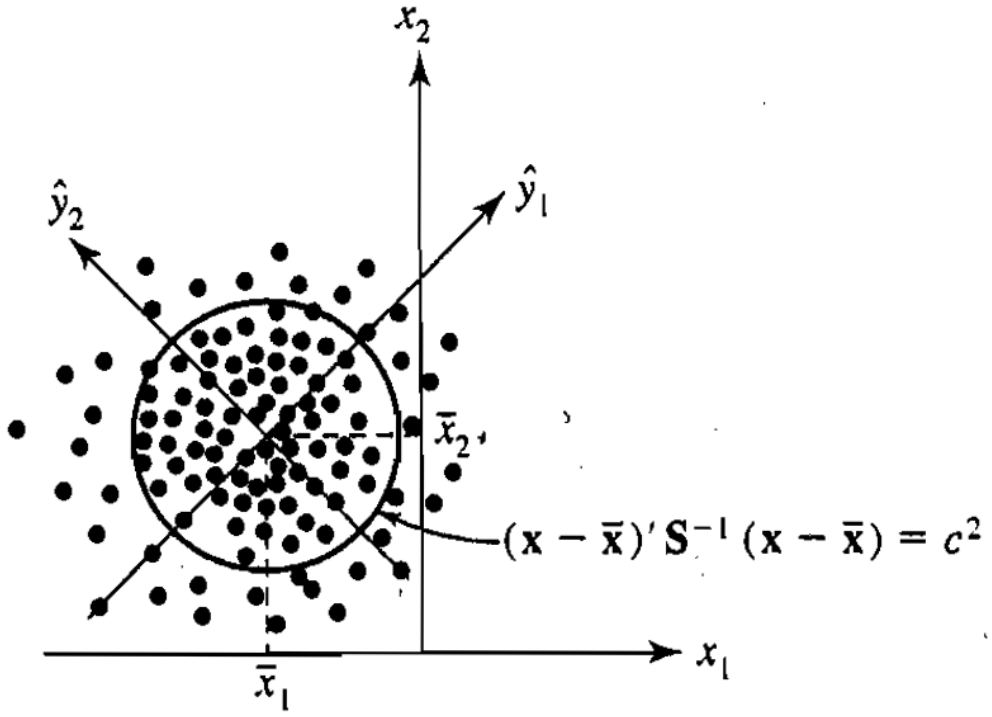
  is a measure of the contribution of the kth variable to the variation of the ith principal component scores.

- If you extract principal components from standardized variables (eigenvectors of the correlation matrix), the k-th element of the i-th eigenvector, e_{ik}, directly determines how much the k-th standardized variable contributes to the score of the i-th principal component.

Two Examples of PCs from MVN Data



(a) $\hat{\lambda}_1 > \hat{\lambda}_2$



(b) $\hat{\lambda}_1 \doteq \hat{\lambda}_2$

PCs from Standardized Variables

- When variables are measured on different scales, it is useful to standardize the variables before extracting the PCs, i.e., compute z-scores:

$$Z_i = \frac{(X_i - \mu_i)}{\sqrt{\sigma_{ii}}},$$

- Note that $\text{Cov}(\mathbf{Z}) = \text{Corr}(\mathbf{X})$, the correlation matrix of the original variables.
- Let $(\lambda_k, \mathbf{e}_k)$ denote the k-th eigenvalue-eigenvector pair of $\text{Corr}(\mathbf{X})$. Then, the score of the k-th principal component is

$$Y_{ik} = e_{k1}Z_{i1} + e_{k2}Z_{i2} + \cdots + e_{kp}Z_{ip}$$

PCs from Standardized Variables

Proceeding as before,

$$\text{trace(correlation matrix)} = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \text{Var}(Z_i) = p$$

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

Further,

$$\left(\begin{array}{c} \text{Proportion of standardized} \\ \text{population variance due} \\ \text{to the } i\text{th principal component} \end{array} \right) = \frac{\lambda_i}{p}, \quad i = 1,...,p,$$

where the λ_i are eigenvalues of the correlation matrix.

PCs from Standardized Variables

- In general, the PCs extracted from Σ=Cov($\mathbf{Z}$) and from Corr($\mathbf{X}$) will not be the same.

- Standardizing variables has important consequences.

- When should one standardize the variables before computing PCs?

PCs from Standardized Variables

- If variables are measured on very different scales (e.g. patient weights in kg vary from 40 to 100, protein concentration in ppm varying between 1 and 10), then the variables with the larger variances will dominate.
- When one variable has a much larger variance than any of the other variables, we will end up with a single PC that is essentially proportional to the dominating variable.
- Consider standardizing the variables when
  - different variables have greatly different variances (this makes all of the variables equally important)
  - you do not want changes in measurement scales to affect the results
  - you want to give more emphasis to describing correlations and less emphasis to describing variances of variables

PCs from Uncorrelated Variables

- If x_1, x_2, ..., x_p are uncorrelated random variables, then Σ is a diagonal matrix with elements $\sigma_{11} = Var(x_1)$, $\sigma_{22} = Var(x_2)$, ..., $\sigma_{pp} = Var(x_p)$. (Suppose $\sigma_{11} \geq \sigma_{22} \geq \cdots \geq \sigma_{pp}$).

- The eigenvalues in this case are $\lambda_i = \sigma_{ii}$ and one choice for the corresponding eigenvector is

$$\mathbf{e}_i = \begin{bmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots 0 \end{bmatrix}'.$$

- Since $\mathbf{e}_i'\mathbf{X} = x_i$ we note that the PCs are just the original variables. Thus, we gain nothing by trying to extract the PCs when the x_i's are uncorrelated.

Principal Components in Practice

- Denote the $n \times p$ data matrix by

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix}.$$

- Each data vector $\mathbf{x}_i := (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is **assumed to be** a random sample from a distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

Principal Components in Practice

- Summary statistics:

$$\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \quad S := \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'$$

$$R := D^{-1/2} S D^{-1/2}$$

  where $D^{-1/2}$ is a diagonal matrix with (j,j) entry $1/\sqrt{s_{jj}}$.

- Principal components: PCs are derived using the same idea as in population principal components, with the true covariance Σ replaced by the sample covariance S.

- Number of PCs is commonly determined by the percentage of sample variations explained by the first few PCs, together with the scree plot.

Sample Principal Components

- Eigenvalue-eigenvector pairs of S are denoted $(\widehat{\lambda}_i, \widehat{\mathbf{e}}_i)$ and the kth sample PC is given by the linear combination

$$\widehat{y}_k = \widehat{\mathbf{e}}_k' \mathbf{x} = \widehat{e}_{k1} x_1 + \widehat{e}_{k2} x_2 + \cdots + \widehat{e}_{kp} x_p, \quad k = 1, ..., p.$$

  where $\mathbf{x} := (x_1, x_2, \cdots, x_p)'$

- Estimated PC scores are centered by subtracting the sample mean vector from each data vector: $(\mathbf{x}_i - \bar{\mathbf{x}})$.

- The estimated score for the i-th subject on the k-th centered PC is $\widehat{y}_{ki} = \widehat{\mathbf{e}}_k'(\mathbf{x}_i - \bar{\mathbf{x}})$.

Sample Principal Components

- The sample mean of the scores for the kth sample PC is zero:

$$\bar{\widehat{y}}_k := \frac{1}{n}\sum_{j=1}^{n} \widehat{\mathbf{e}}_k'(\mathbf{x}_j - \bar{\mathbf{x}}) = \frac{1}{n}\widehat{\mathbf{e}}_k' \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}}) = 0.$$

- The sample variance of the kth sample PC is $\widehat{\lambda}_k$, the k-th largest eigenvalue of S.
- Sample principal components are uncorrelated.
- The total sample variance $s_{11} + s_{22} + ... + s_{pp}$ is equal to $\widehat{\lambda}_1 + ... + \widehat{\lambda}_p$
- The relative contribution of the jth variable to the kth sample PC is given by $r_{\widehat{y}_k, x_j}$.

Carapace Measurements for Female Turtles

- Data on three dimensions of female turtle carapaces (shells):
  - X_1=log(carapace length)
  - X_2=log(carapace width)
  - X_3=log(carapace height)
- Since the measurements are all on the same scale (mm), the PCs may be extracted from the sample covariance matrix S
- R code demonstration

Five Socioeconomic Variables

- Data on socioeconomic variables for $n = 14$ census tracks in Madison, Wisconsin:
  - X_1: population (in thousands)
  - X_2: percentage with professional degrees
  - X_3: percentage employed (over age 16)
  - X_4: government employment (percent)
  - X_5: median home value (in hundreds of thousands of dollars)
- We extracted the PCs using both the covariance matrix S and the correlation matrix R.
- R code demonstration

Bartlett's Test for Equal Eigenvalues for Σ

- Test the null hypothesis $H_0 : \lambda_{q+1} = \lambda_{q+2} = \cdots = \lambda_{q+r}$ that the r smallest population eigenvalues are equal. $(p = q + r)$
- This may correspond to a situation in which the first q principal components account for essentially all of the correlations among the measured attributes and most of the variances. What they do not capture is small random variation with essentially no correlation pattern.
- Large sample chi-square test rejects the null hypothesis if

$$\chi^2 = (v)(r) \ln \left[\frac{1}{r} \sum_{i=q+1}^{q+r} \hat{\lambda}_i \right] - v \sum_{i=q+1}^{q+r} \ln(\hat{\lambda}_i) > \chi^2_{r(r+1)/2-1}$$

  where $v =$ (df for S) $- (2p + 5)/6$.

- R code demonstration

Sample PCA via Singular Value Decomposition

- A more stable method to compute PCA is via singular value decomposition (SVD)
- SVD for PCA requires columns centered data matrix.
- A SVD of the $n \times p$ matrix A yields the composition $A = UDV'$, where
  - U is $n \times n$ orthogonal matrix with $U'U = I_{n \times n}$.
  - D is $n \times p$ diagonal matrix with the diagonal entries called *singular values*.
  - V is $p \times p$ orthogonal matrix with $V'V = I_{p \times p}$.
- If X has zero mean, then $X'X = (n-1)S$.

Sample PCA via Singular Value Decomposition

- Compute $Y := \frac{1}{\sqrt{n-1}}(X - \bar{\mathbf{x}}\mathbf{1}'_{n\times 1})$.
- Perform SVD for $YY' = UDV$.
- The columns of V are the loadings (in our definition) for the principal components of X.
  - Using the other naming convention (that is not adopted in this course), people call the columns of V the PCs of X.
- The data matrix UD is a matrix of scores.

PCA for Other Applications

- Principal component regression:
  - One first identifies d PCs with loadings $\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$
  - Then perform regression of certain response Y on $(\mathbf{a}_1'\mathbf{X}, \ldots, \mathbf{a}_d'\mathbf{X})$
  - Model: $Y_i = \alpha_0 + (\mathbf{a}_1'\mathbf{x}_i)\alpha_1 + \cdots + (\mathbf{a}_d'\mathbf{x}_i)\alpha_d + \epsilon_i$
  - Prediction for Y_0.
  - Some issues:
    - How to select d?
    - In regression, Y is not involved in PCA?
    - Can we perform dimension reduction for X using information from Y?
- PCA can be used for image compression.

Final Remarks

- In many bioinformatics applications, p is much larger than the observations.
- In many applications, X are not real-valued but integer or zero-one valued.
- How can we apply PCA in these settings?