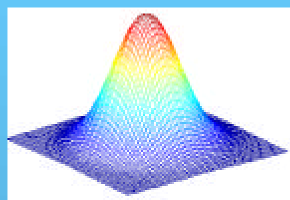
Methods and Techniques

of investigating user behavior



Lecture 4

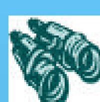
Principal Component Analysis

elementary matrix algebra

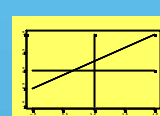
Principal Component Analysis

elementary matrix algebra

Outline

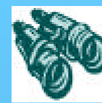


- short flashback
- matrices, elementary operations
- covariance matrices under linear transformations



- principal component analysis: when to use and how to do it
- the biplot; application of PCA

Looking back: Multivariate data



visualization

- univariate *[histogram, boxplot]*
- bivariate *[scatter plot]*
- trivariate *[scatter plot (matrix), bubble plot, coplot]*
- multivariate *[scatter plot matrix, Chernoff faces, stars]*

representation

- data matrix *[n rows, p columns]*
- mean (vector) *[p rows, 1 column]*
- (co)variance (matrix) *[p rows, p columns]*
- correlation (matrix) *[p rows, p columns]*

Numerical summaries

Basic Matrix algebra

A *matrix* is an $n \times p$ array of numbers

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

$n \times 1$ and $1 \times p$ matrices are called *vectors*

$$u = \begin{pmatrix} u_1 & u_2 & \cdots & u_p \end{pmatrix} \qquad v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

Basic Matrix algebra

Multiplying a matrix A by a number (scalar): multiply its elements

Transpose A´ of a matrix A: interchange elements *ij* and *ji*

$$A = \begin{pmatrix} 2 & 0.5 \\ 1.5 & 3 \\ -1 & 6 \end{pmatrix} \Rightarrow 3A = \begin{pmatrix} 6 & 1.5 \\ 4.5 & 9 \\ -3 & 18 \end{pmatrix}, A' = \begin{pmatrix} 2 & 1.5 & -1 \\ 0.5 & 3 & 6 \end{pmatrix}$$

➡ A is $n \times p$ matrix, A´ is a ? × ? matrix

Basic Matrix algebra

Multiplying $n \times p$ matrix A and $p \times m$ matrix B:

equal!

$$C = AB, c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

$$A = \begin{pmatrix} 2 & 0.5 \\ 1.5 & 3 \\ -1 & 6 \end{pmatrix}, B = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 1 & 3 & 5 & 7 \end{pmatrix}, C = AB \Rightarrow$$

$$\Rightarrow C = \begin{pmatrix} 8.5 & 7.5 & 6.5 & 5.5 \\ 9 & 13.5 & 18 & 22.5 \\ 2 & 15 & 28 & 41 \end{pmatrix}$$

➡ A is 3 × 2 matrix, does AA exist? And A´A?

Covariance under linear transformations

X: $n \times p$ data matrix

row i corresponds to measurements of p variables on individual i

S: covariance matrix of X $\quad s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$

mean of j-th column

Consider two 'new' variables, y and z (for each individual):

$$y_i = \sum_{j=1}^{p} a_j x_{ij} \quad \text{and} \quad z_i = \sum_{j=1}^{p} b_j x_{ij}$$

Covariance under linear transformations

Consider two 'new' variables, y and z (for each individual):

$$y_i = \sum_{j=1}^{p} a_j x_{ij} \quad \text{and} \quad z_i = \sum_{j=1}^{p} b_j x_{ij}$$

Then the variance of y, the variance of z and the covariance of y and z are given by:

$$s_{yy} = a'Sa, \quad s_{zz} = b'Sb \quad \text{and} \quad s_{yz} = a'Sb$$

(a *and* b *are considered as column vectors*)

Principal Component Analysis

how?

Use new variables (components) which are linear combinations of the old variables, as follows:

First PC: $y_i = \sum_{j=1}^{p} a_j x_{ij}$

where α is chosen such that:

$\sum_{j=1}^{p} a_j^{\,2} = 1$ and the variance of y is maximized

Mathematical optimization problem:

maximize $f(a) = a'Sa$ such that $a'a = 1$

Maximal discrimination between individuals!

Principal Component Analysis

Optimization problem can be solved explicitly $\Rightarrow a^{(1)}$

First PC: $y_i^{(1)} = \sum_{j=1}^{p} a_j^{(1)} x_{ij}$

Second PC: $y_i^{(2)} = \sum_{j=1}^{p} a_j x_{ij}$

where α is chosen such that:

$\sum_{j=1}^{p} a_j^{\,2} = 1$, the covariance between $y^{(2)}$ and $y^{(1)}$ is zero

and the variance of $y^{(2)}$ is maximized

Mathematical optimization problem:

maximize $f(a) = a'Sa$ such that $a'a = 1$ and $a'Sa^{(1)} = 0$

Principal Component Analysis

Optimization problem can be solved explicitly

Second PC: $y_i^{(2)} = \sum_{j=1}^{p} a_j^{(2)} x_{ij} \Rightarrow a^{(2)}$

Third PC: $y_i^{(3)} = \sum_{j=1}^{p} a_j x_{ij}$

where α is chosen such that:

$\sum_{j=1}^{p} a_j^{2} = 1$, the covariance between $y^{(3)}$ and $y^{(k)}$ is zero (k=1,2)

and the variance of $y^{(3)}$ is maximized

Mathematical optimization problem:

maximize $f(a) = a'Sa$ such that $a'a = 1$ and $a'Sa^{(k)} = 0$ (k=1,2)

and so forth........until p-th component

Principal Component Analysis

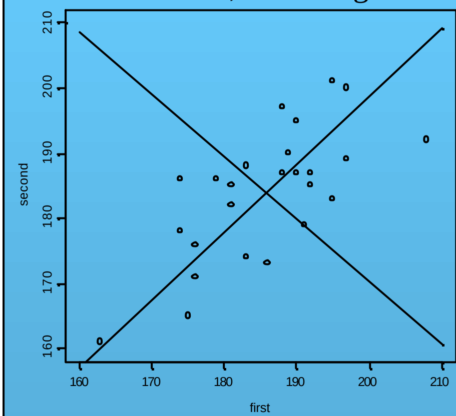
why?

- First PC gives maximal discrimination among individuals
- Reduction in dimensionality if higher PC's have small variance
- Most informative 2- or 3D projection of the data
- As input for further analysis of the data (e.g. regression analysis)
- Detection of outliers

two examples
- head size of brothers
- decathlon

Head size of brothers

In 25 families, head length of two oldest sons in mm.



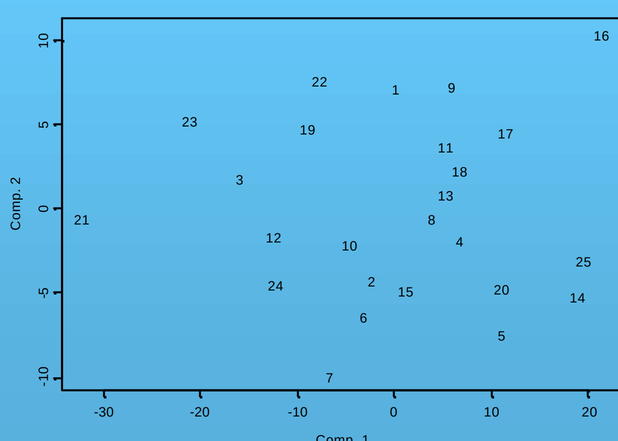
$$\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 185.72 \\ 183.88 \end{pmatrix}$$

$$S = \begin{pmatrix} 95.29 & 69.17 \\ 69.17 & 100.94 \end{pmatrix}$$

$$\begin{cases} a^{(1)} = (0.693, 0.721)' \\ a^{(2)} = (0.721, -0.693)' \end{cases}$$

Head size of brothers

Plot of scores on two components



Olympic decathlon

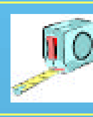
revisited

10 disciplines, 34 competitors

and many other data sets!

Important issue concerning these data:

Units of measurement incomparable among different disciplines (e.g. 100m in seconds and high jump in meters)



Use correlation matrix instead of covariance matrix to extract the principal components (equivalently: rescale original variables in data matrix to zero mean and unit variance)

Olympic decathlon

- Compute principal component loadings (α- vectors). E.g.:

$$a^{(1)} = (0.36\ \ 0.36\ \ 0.32\ \ 0.27\ \ 0.29\ \ 0.37\ \ 0.31\ \ 0.39\ \ 0.29\ \ 0.08)'$$

$$a^{(2)} = \underbrace{(-0.20\ \ -0.20\ \ 0.39\ \ -0.01\ \ -0.43\ \ -0.13\ \ 0.42\ \ 0.06\ \ 0.30\ \ -0.55)'}_{p=10}$$

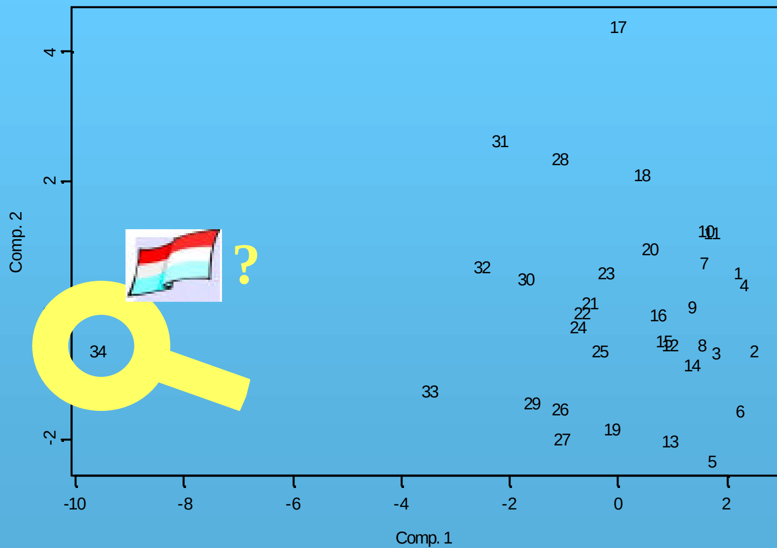
- Compute scores (what do individuals *score* on the various PC's?)

$$y^{(1)} = (2.21\ \ 2.52\ \ 1.79\ \ 2.33\ \ 1.73\ \cdots\ -2.15\ -2.50\ -3.45\ -9.59)'$$

$$y^{(2)} = (\underbrace{0.60\ -0.61\ -0.64\ \ 0.43\ -2.31\ \cdots\ 2.64\ \ 0.71\ -1.24\ -0.62}_{n=34})'$$

- Plot scores on first two PC's

Olympic decathlon



Olympic decathlon

analysis without <u>outlier</u> "34"

- Compute principal component loadings (α- vectors). E.g.:

$$a^{(1)} = (0.42\ \ 0.39\ \ 0.27\ \ 0.21\ \ 0.36\ \ 0.43\ \ 0.18\ \ 0.38\ \ 0.18\ \ 0.17)'$$

$$a^{(2)} = (\underbrace{-0.15\ \ -0.15\ \ 0.48\ \ 0.03\ \ -0.35\ \ -0.07\ \ 0.50\ \ 0.15\ \ 0.37\ \ -0.42}_{p=10})'$$

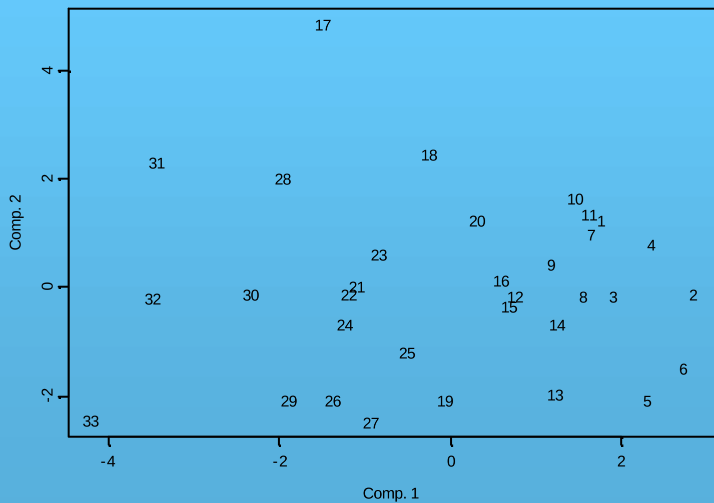
- Compute scores (what do individuals *score* on the various PC's?)

$$y^{(1)} = (1.76\ \ 2.83\ \ 1.91\ \ 2.35\ \ 2.30\ \cdots\ -2.32\ -3.43\ -3.47\ -4.19)'$$

$$y^{(2)} = (\underbrace{1.25\ -0.10\ -0.14\ \ 0.81\ -2.05\ \cdots\ -0.12\ \ 2.31\ -0.19\ -2.43}_{n=33})'$$

- Plot scores on first two PC's

Olympic decathlon



Olympic decathlon

interpretation of the PC's

Consider first PC:

$$a^{(1)} = (0.42\ \ 0.39\ \ 0.27\ \ 0.21\ \ 0.36\ \ \ 0.43\ \ 0.18\ \ 0.38\ \ 0.18\ \ 0.17)'$$

Positive loadings of all variables; measure of general performance; (kind of mean); 100m and 110m hurdles have highest loadings.

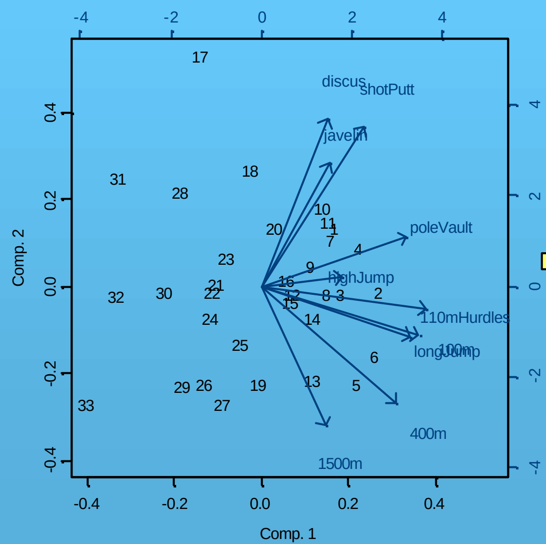
The second PC:

$$a^{(2)} = (-0.15\ \ -0.15\ \ 0.48\ \ 0.03 - 0.35\ \ -0.07\ \ 0.50\ \ 0.15\ \ 0.37\ -0.42)'$$

High positive loadings on shot putt, discus and javelin. High negative loadings on 400m and 1500m;
'contrast between power and endurance'

Olympic decathlon

biplot

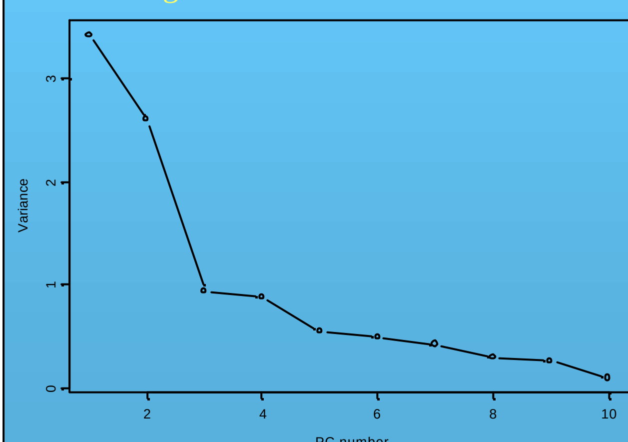


Visualizes the scores on the two first PC's *and* the loadings of the original variables on these two PC's

Olympic decathlon

how many PC's to use?

Scree diagram:



Cut off if variance is sufficient.

E.g.:

- sum ≥ 0.8 total
- variance PC ≤ 1
- elbow in scree diagram

Principal Component Analysis

Given: $n \times p$ data matrix containing measurements on p (correlated) variables on n objects / individuals / experimental units

Aim: lower dimensional representation and visualization of the data matrix in terms of scores on (uncorrelated) principal components (new variables; linear combinations of the original variables) $\Rightarrow$ score plot and biplot

Choices to be made:

- covariance or correlation matrix (raw or standardized data)
- number of PC's to use