Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Principal Component Analysis

Applied Multivariate Statistics – Spring 2013

Overview

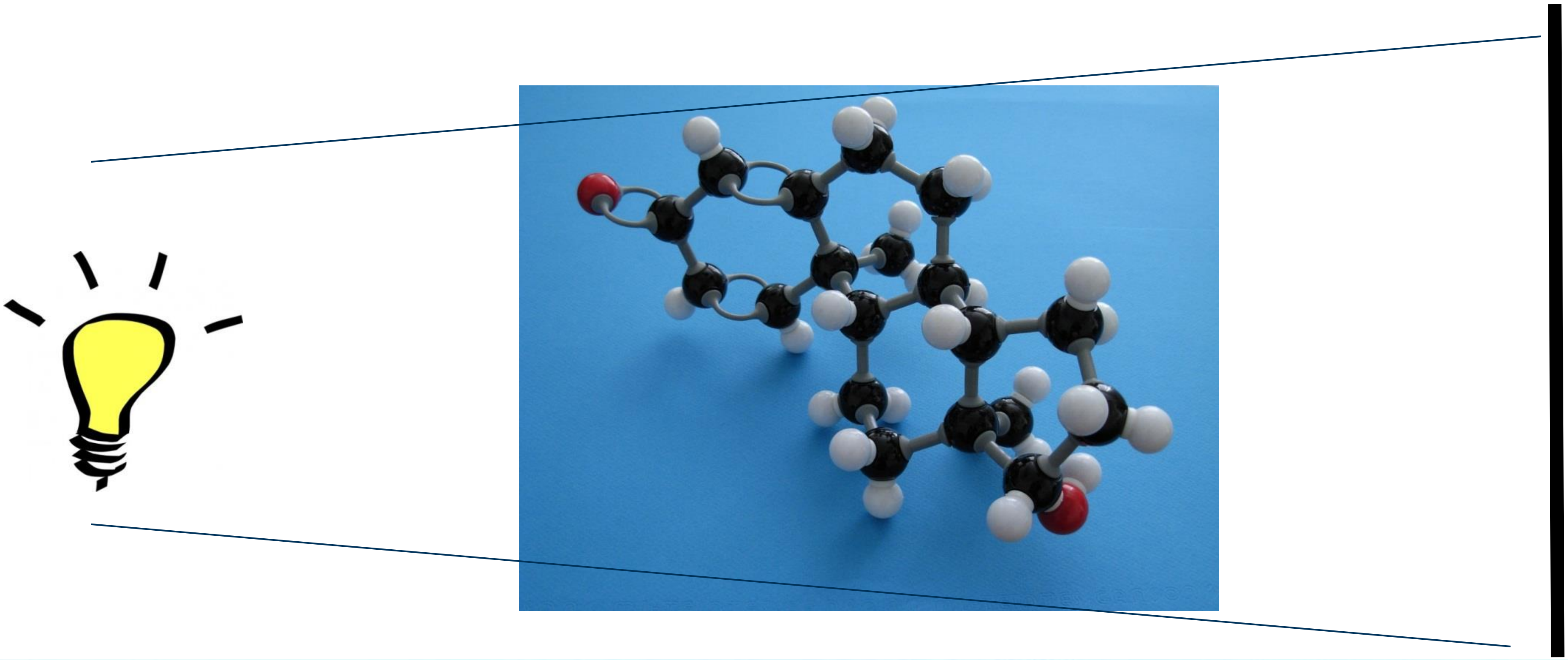
- Intuition
- Four definitions
- Practical examples
- Mathematical example
- Case study

PCA: Goals

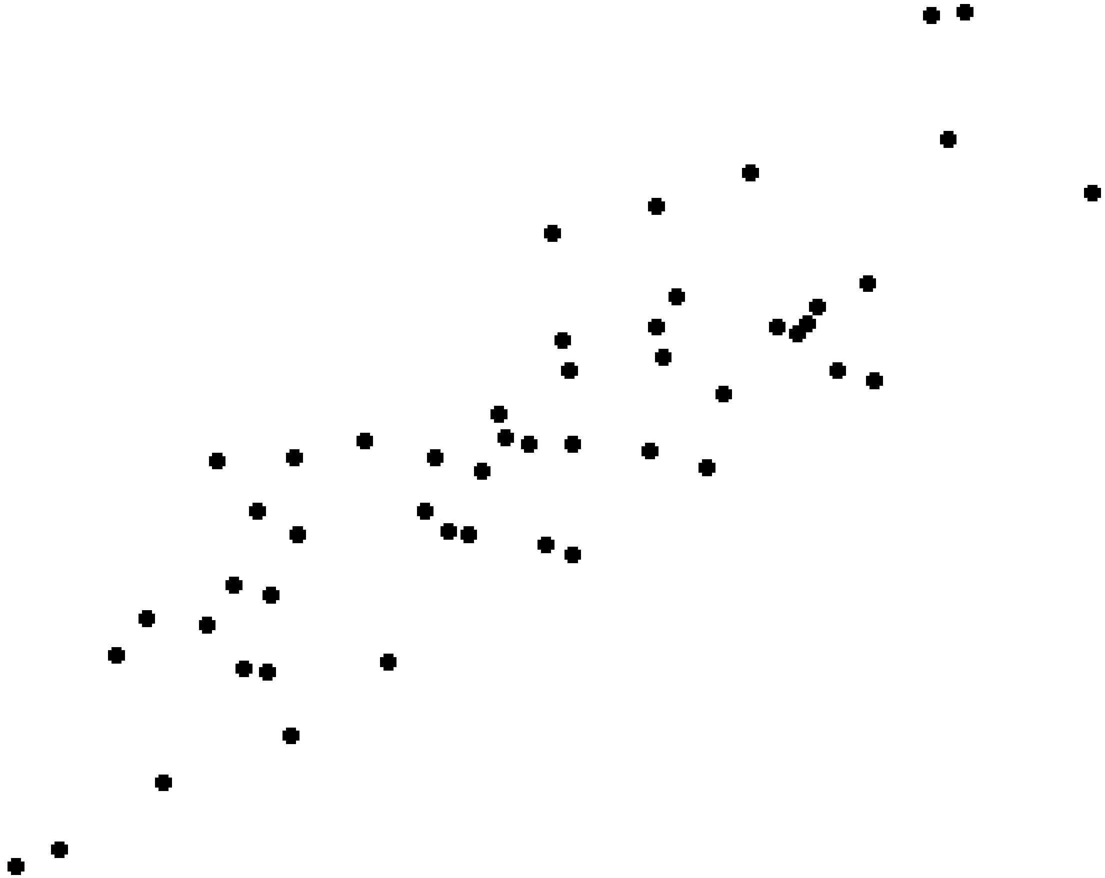
- Goal 1: Dimension reduction to a few dimensions while explaining most of the variance
(use first few PC's)
- Goal 2: Find one-dimensional index that separates objects best
(use first PC)

PCA: Intuition

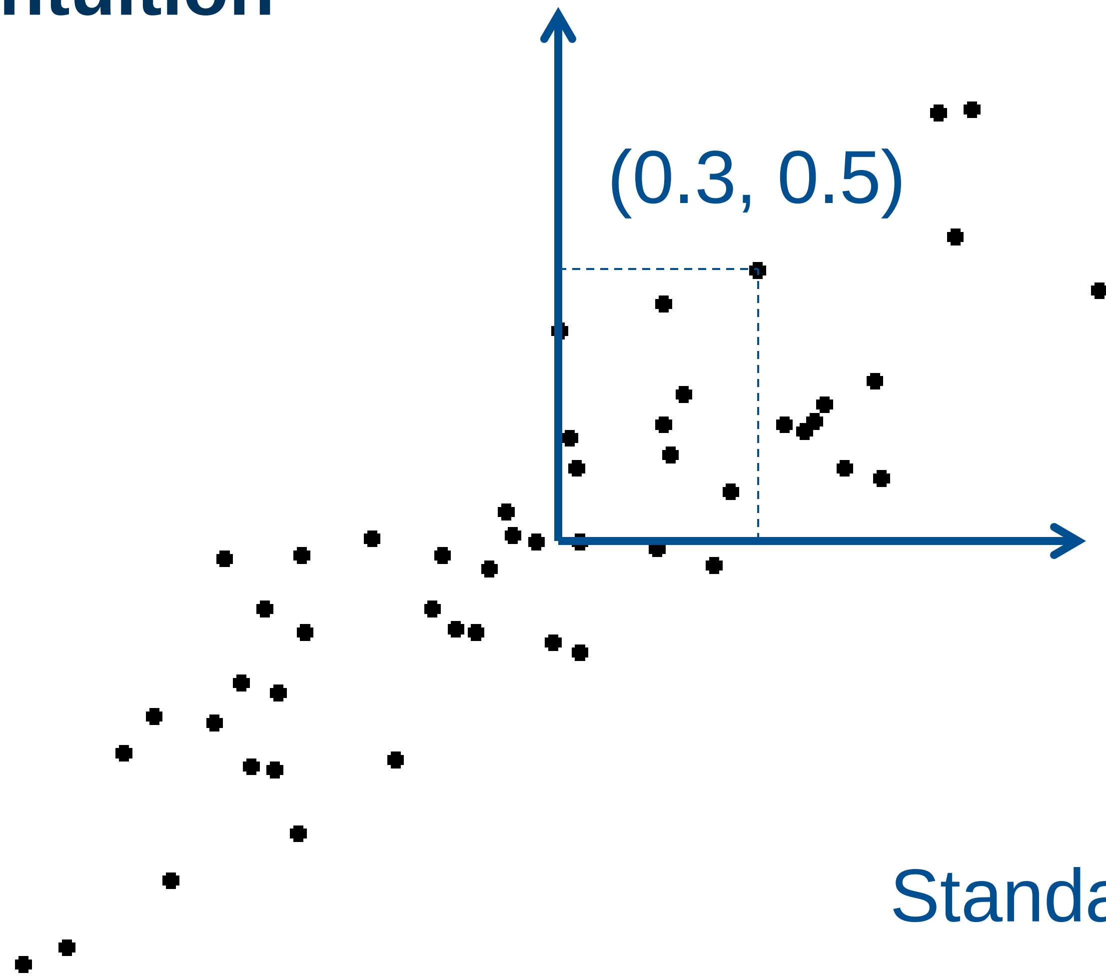
- Find low-dimensional projection with largest spread



PCA: Intuition



PCA: Intuition



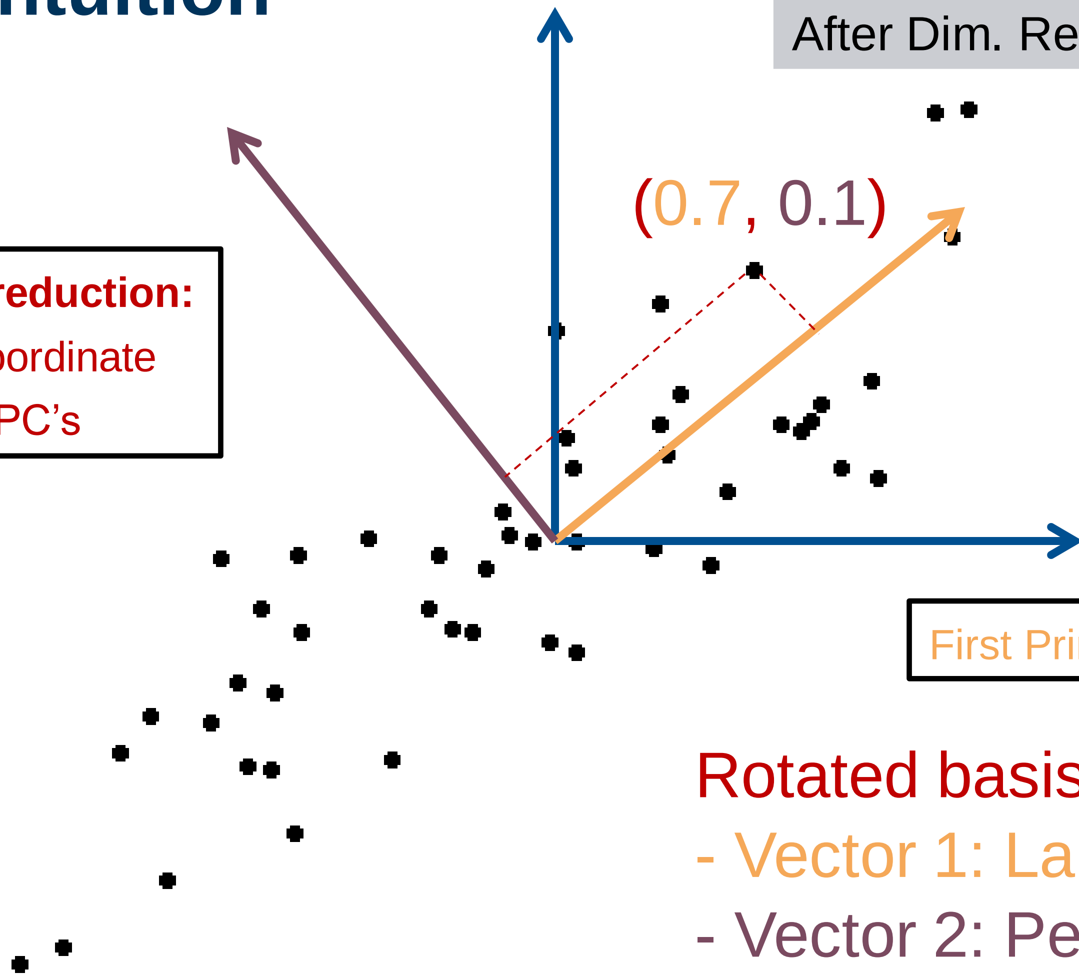
Standard basis

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PCA: Intuition

	X_1	X_2
Std. Basis	0.3	0.5
PC Basis	0.7	0.1
After Dim. Reduction	0.7	-

Dimension reduction:
Only keep coordinate of first (few) PC's

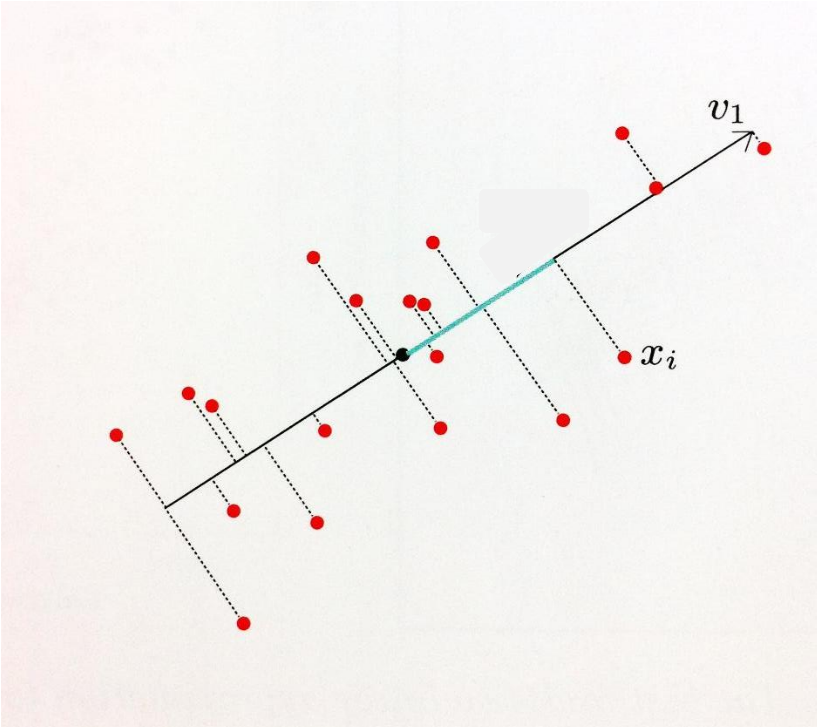


First Principal Component (1.PC)

Rotated basis:
- Vector 1: Largest variance
- Vector 2: Perpendicular

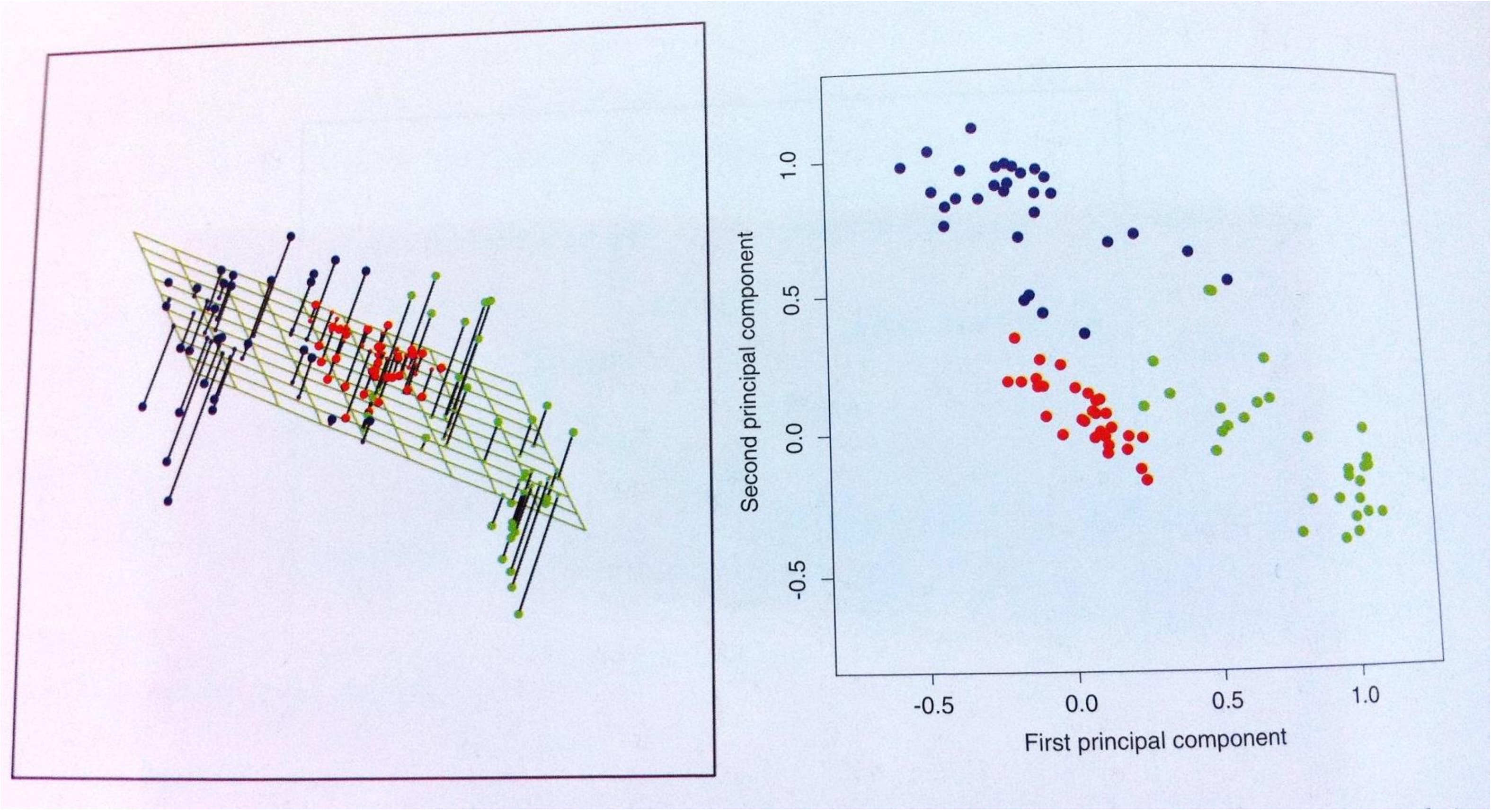
Second Principal Component (2.PC)

PCA: Intuition in 1d



Taken from “The Elements of Stat. Learning”, T. Hastie et.al.

PCA: Intuition in 2d



Taken from “The Elements of Stat. Learning”, T. Hastie et.al.

Overview

- ~~Intuition~~
- Four definitions
- Practical examples
- Mathematical example
- Case study

PCA: Four equivalent definitions

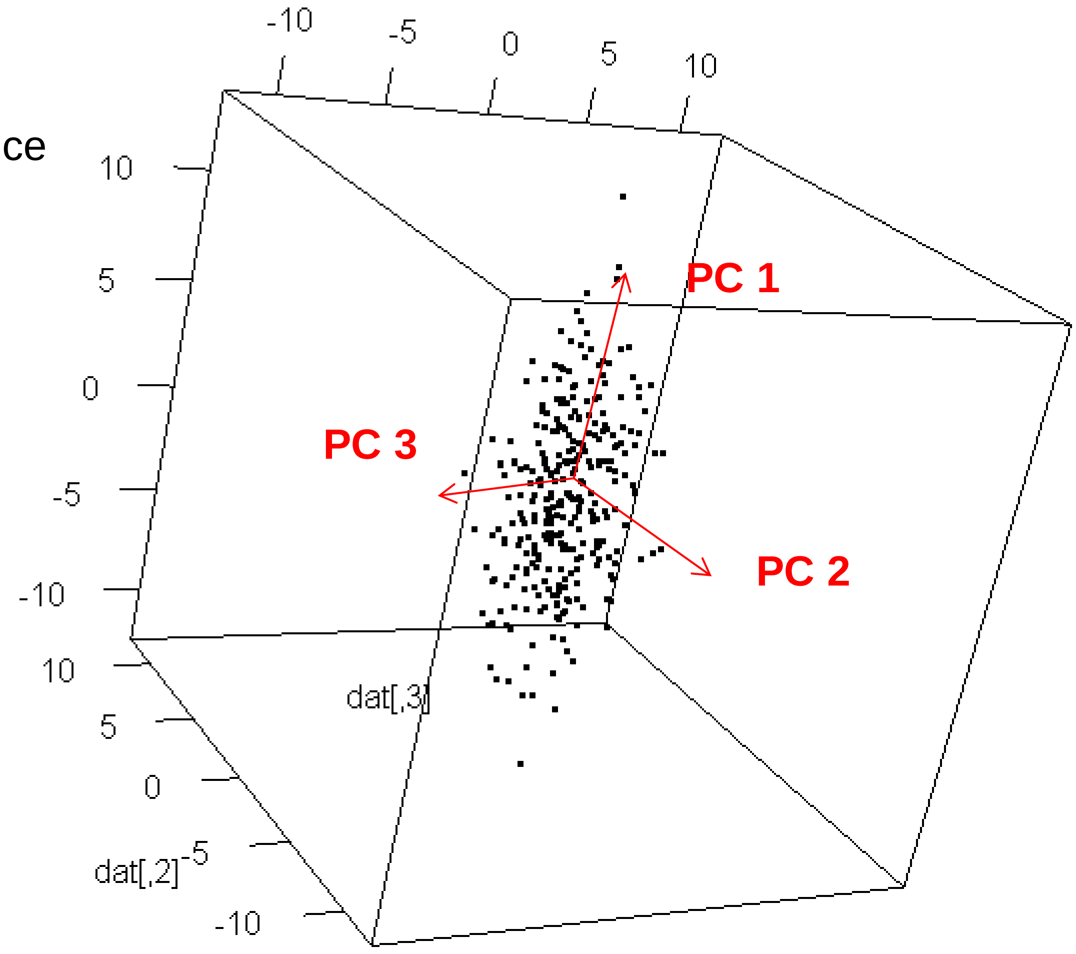
Good for intuition

- (Always center data first !)
- Orthogonal directions with largest variance
- Linear subspace (straight line, plane, etc.) with minimal squared residuals
- Using Spectraldecomposition (=Eigendecomposition)
- Using Singular Value Decomposition (SVD)

Good for computing

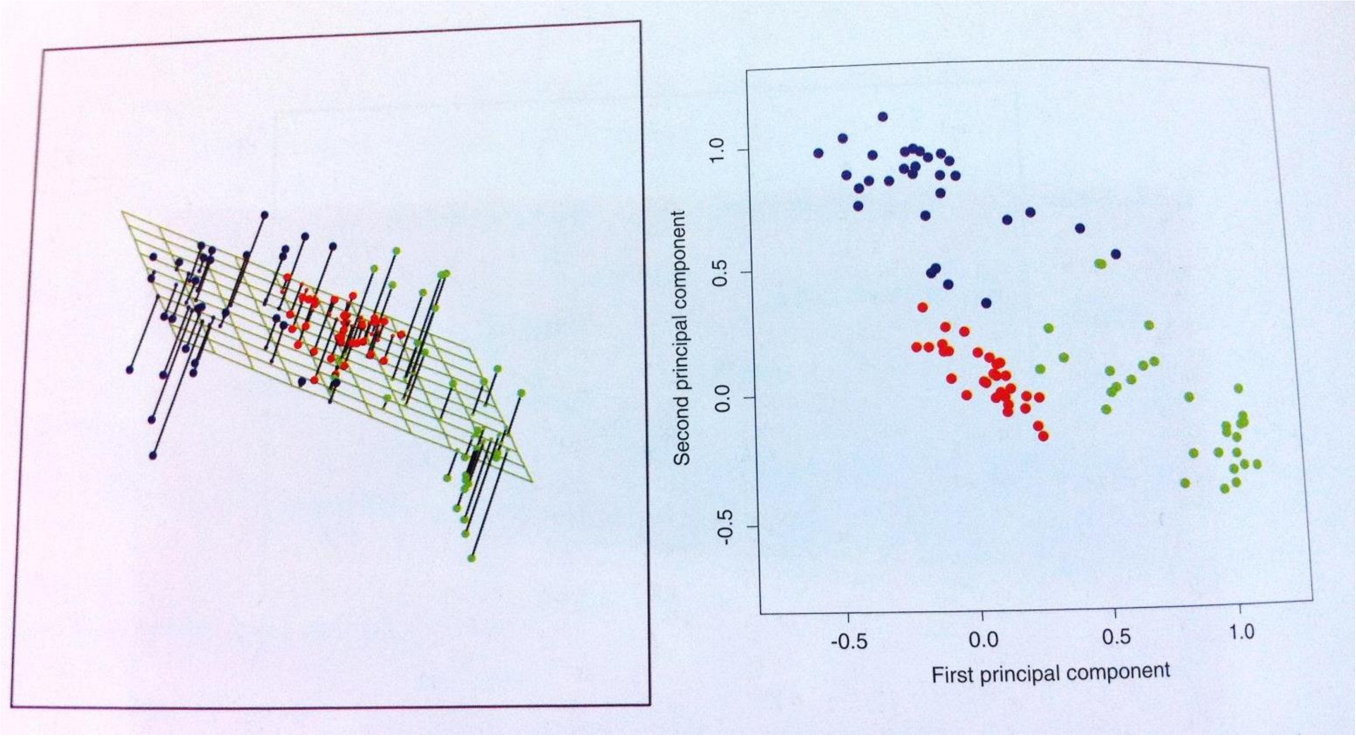
PCA (Version 1): Orthogonal directions

- PC 1 is direction of largest variance
- PC 2 is
  - perpendicular to PC 1
  - again largest variance
- PC 3 is
  - perpendicular to PC 1, PC 2
  - again largest variance
- etc.



PCA (Version 2): Best linear subspace

- PC 1: Straight line with smallest orthogonal distance to all points
- PC 1 & PC 2: Plane with smallest orthogonal distance to all points
- etc.



PCA (Version 3): Eigendecomposition

- **Spectral Decomposition Theorem**:
  Every symmetric, positive semidefinite Matrix R can be rewritten as
  $$R = A\ D\ A^T$$
  where D is diagonal and A is orthogonal.
- Eigenvectors of Covariance/Correlation matrix are PC's
  Columns of A are PC's
- Diagonal entries of D (=eigenvalues) are variances along PC's (usually sorted in decreasing order)

- R: Function "princomp"

PCA (Version 4): Singular Value Decomposition

- **Singular Value Decomposition**:
  Every matrix R can be rewritten as
  $$R = U\ D\ V^T$$
  where D is diagonal and U, V are orthogonal.
- Columns of V are PC's
- Diagonal entries of D are "singular values"; related to standard deviation along PC's (usually sorted in decreasing order)
- UD contains samples measured in PC coordinates

- R: Function "prcomp"

Overview

- ~~Intuition~~
- ~~Four definitions~~
- Practical examples
- Mathematical example
- Case study

Example: Headsize of sons

Standard deviation in direction of 1.PC,

Var = 12.69^2 = 167.77

```
> summary(head_pca, loadings = TRUE)
Importance of components:
                          Comp.1 Comp.2
Standard deviation        12.69   5.22
Proportion of Variance     0.86   0.14
Cumulative Proportion      0.86   1.00

Loadings:
      Comp.1 Comp.2
head1  0.69  -0.72
head2  0.72   0.69
```

Standard deviation in direction of 2.PC,

Var = 5.22^2 = 28.33

Total Variance = 167.77 + 28.33 = 196.1

1.PC contains

167.77/196.1 = 0.86

of total variance

2.PC contains

28.33/196.1 = 0.14

of total variance

y_1 = 0.69*x1 + 0.72*x2

y_2 = -0.72*x1 + 0.69*x2

Computing PC scores

- Substract mean of all variables
- Output of princomp: $scores
  First column corresponds to coordinate in direction of 1.PC,
  Second col. corresponds to coordinate in direction of 2.PC,
  etc.
- Manually (e.g. for new observations):
  Scalar product of loading of i^{th} PC gives coordinate in direction of i^{th} PC
- Predict new scores: Use function “predict”
  (see ?predict.princomp)
- Example: Headsize of sons

Interpretation of PCs

- Oftentimes hard
- Look at loadings and try to interpret:

```
Loadings:
       Comp.1 Comp.2
head1   0.69  -0.72
head2   0.72   0.69
```

Difference in head sizes of both sons

Average head size of both sons

To scale or not to scale…

- R: In princomp, option “cor = TRUE” scales variables
  Alternatively: Use correlation matrix instead of covariance matrix
- Use correlation, if different units are compared
- Using covariance will find the variable with largest spread as 1. PC
- Example: Blood Measurement

How many PC's?

- No clear cut rules, only rules of thumb
- Rule of thumb 1: Cumulative proportion should be at least 0.8 (i.e. 80% of variance is captured)
- Rule of thumb 2: Keep only PC's with above-average variance
  (if correlation matrix / scaled data was used, this implies: keep only PC's with eigenvalues at least one)
- Rule of thumb 3: Look at scree plot; keep only PC's before the "elbow" (if there is any…)

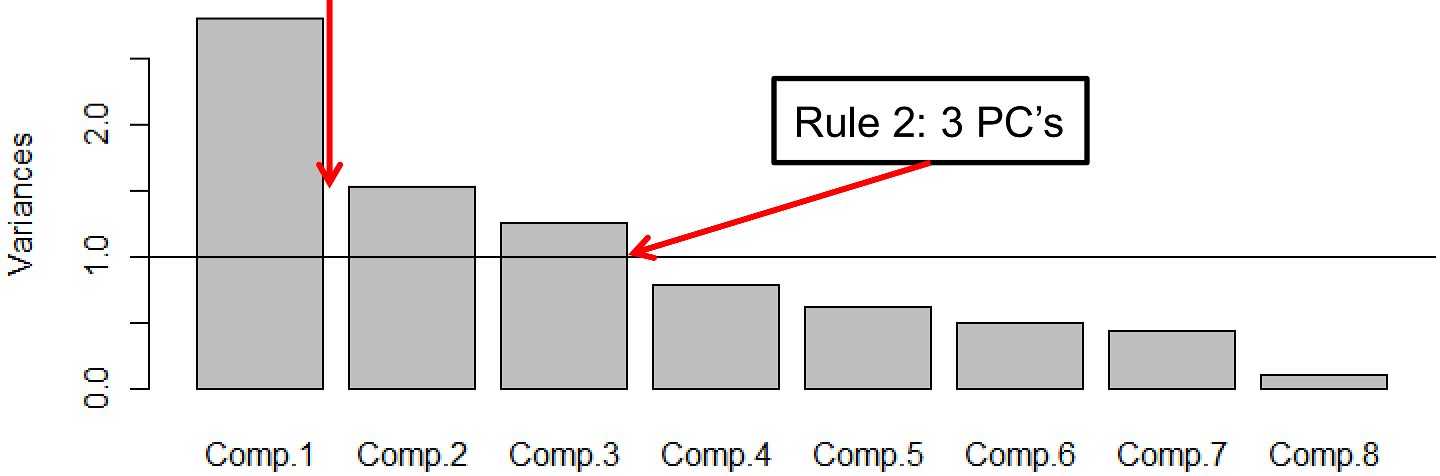
How many PC’s: Blood Example

Rule 1: 5 PC’s

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.6710100 1.2375848 1.1177138 0.88227419 0.78829505 0.69917350
Proportion of Variance 0.3490343 0.1914520 0.1561605 0.09730097 0.07769584 0.06110545
Cumulative Proportion  0.3490343 0.5404863 0.6966468 0.79394778 0.87164363 0.93274908
                           Comp.7     Comp.8
Standard deviation     0.66002394 0.31996216
Proportion of Variance 0.05445395 0.01279697
Cumulative Proportion  0.98720303 1.00000000
```

Rule 3: Ellbow after PC 1 (?)

blood_pcacor



Overview

- ~~Intuition~~
- ~~Four definitions~~
- ~~Practical examples~~
- Mathematical example
- Case study

Mathematical example in detail: Computing eigenvalues and eigenvectors

- Correlation matrix: $R = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$
- Find basis, in which R is diagonal:
  Eigenvectors are these basis vectors
  Eigenvalues are entries in diagonal matrix

Mathematical example in detail: Computing eigenvalues

- $\det(R - \lambda 1) = 0$, solve for λ
- $\det(R - \lambda 1) = \det\left(\begin{pmatrix} 1-\lambda & r \\ r & 1-\lambda \end{pmatrix}\right) = \lambda^2 - 2\lambda + 1 - r^2 = 0$
- Thus the eigenvalues are: $\lambda_{1,2} = 1 \pm r$
  The variance along PC 1 is 1+r, the variance along PC 2 is 1-r
- Thus, there is a basis, in which R looks like:

$$\mathrm{R} = \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix}$$

Mathematical example in detail: Computing eigenvectors

- For each eigenvalue, find a vector v_i so that $Rv_i = \lambda_i v_i$ holds
- Choose vectors that have unit length for convenience
- For 1+r: $Rv_1 = (1+r)v_1 \rightarrow v_1 = (0.71, 0.71)$
  For 1-r: $Rv_2 = (1-r)v_2 \rightarrow v_2 = (-0.71, 0.71)$
- v_1, v_2 are the directions of PC1 and PC2
- New observations can be expressed using coordinates of PC1 and PC2 by the linear algebra technique “change of base”
- That’s what the R function “princomp” does

Overview

- ~~Intuition~~
- ~~Four definitions~~
- ~~Practical examples~~
- ~~Mathematical example~~
- Case study

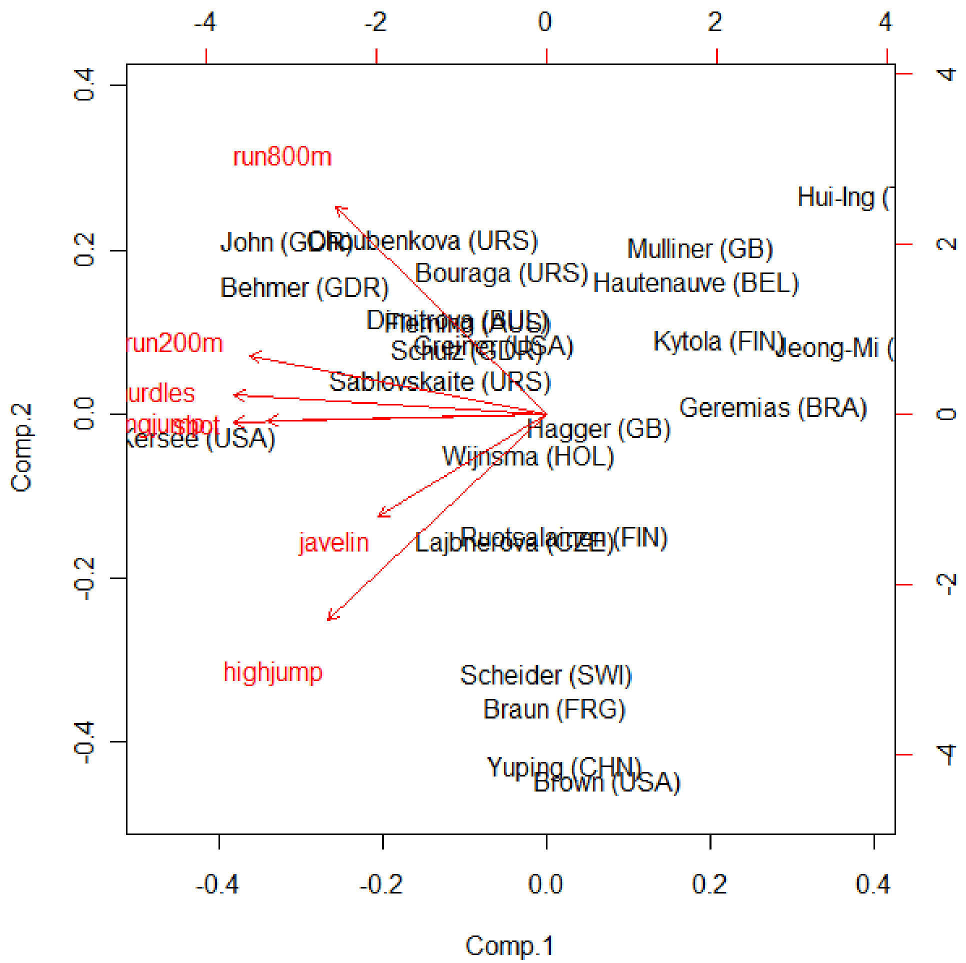
Case study: Heptathlon Seoul 1988

Biplot: Show info on samples AND variables

Approximately true:

- Data points: Projection on first two PCs
  Distance in Biplot ~ True Distance
- Projection of sample onto arrow gives original (scaled) value of that variable
- Arrowlength: Variance of variable
- Angle between Arrows: Correlation

Approximation is often crude;
good for quick overview



PCA: Eigendecomposition vs. SVD

- PCA based on Eigendecomposition: princomp
  + easier to understand mathematical background
  + more convenient summary method
- PCA based on SVD: prcomp
  + numerically more stable
  + still works if more dimensions than samples

- Both methods give same results up to small numerical differences

Concepts to know

- 4 definitions of PCA
- Interpretation: Output of princomp, biplot
- Predict scores for new observations
- How many PC's?
- Scale or not?
- Know advantages of PCA based on SVD

R functions to know

- princomp, biplot
- (prcomp – just know that it exists and that it does the SVD approach)