



Laboratory Techniques and Medical Device Approvals I

Part 02: Multivariate Data Analysis

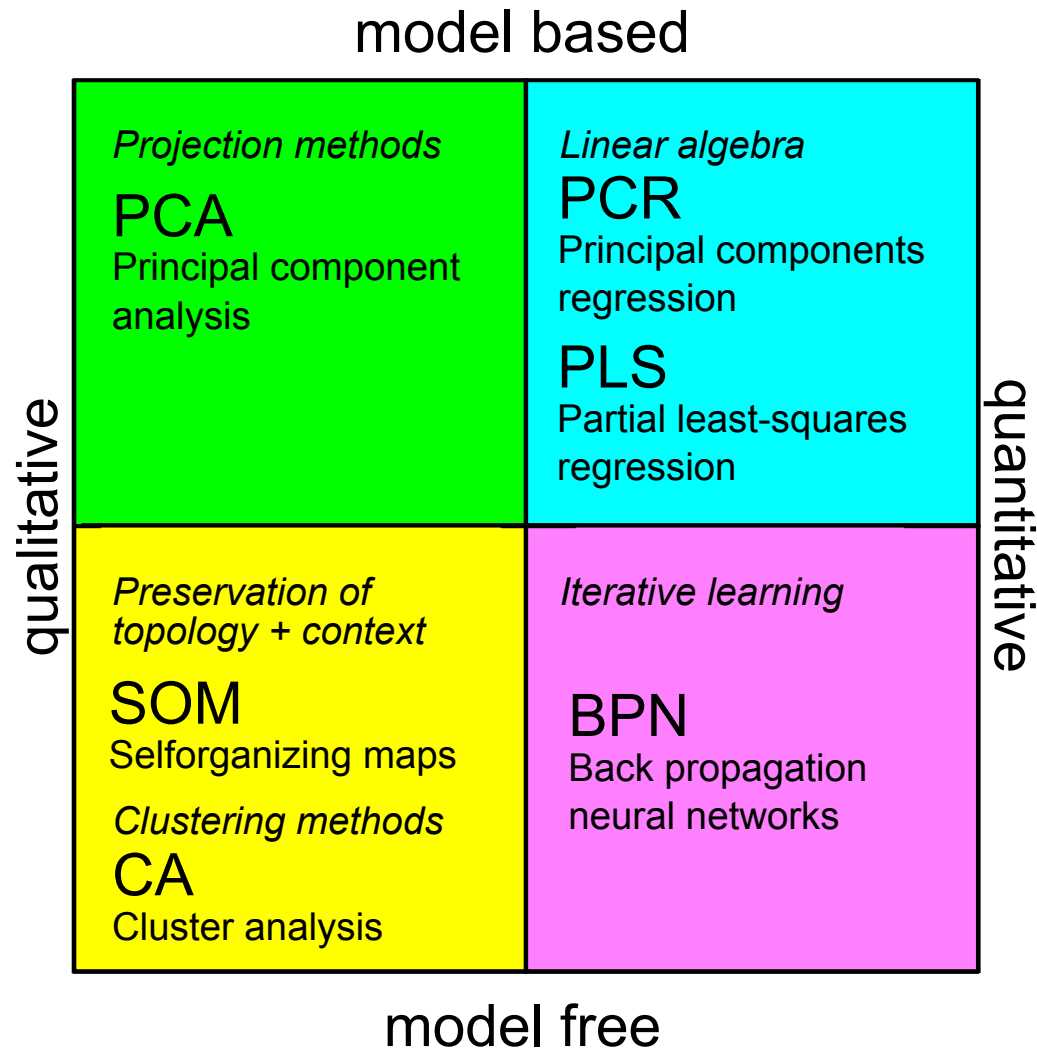


Multivariate Data Analysis

Introduction and basic aspects



Data analysis: Overview over different methods





hard modelling:

The number of variables or parameters is fixed

soft modelling:

- The number of variables or parameters is not fixed. It is determined after running the regression analysis.
- Depending on the number of variables, the data evaluation is either univariate or multivariate.
- In the context of classical algorithms it is also called ***Multiple Regression***
- The terms univariate, multivariate und multiple multivariate are explained in the following with some examples
 - x_i is the independent variable (e.g. concentrations)
 - y_i is the dependent variable (e.g. measurement results)
- These relationships can be solved either
 - classical with a direct solution $y = f(x)$ or
 - with an inverse approach for $x = g(y)$
- In general the samples are called **objects**



univariate

1 independent, 1 dependent variable
(→ 1 component: E.g. measurement
of one wavelength at different points
in time or at different samples)

$$\mathbf{x}, \mathbf{y} \quad o = \text{Number of objects}$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_o \end{pmatrix} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_o \end{pmatrix}$$

multivariate

1 independent variable (component),
multiple dependent variables
(e.g. spectral measurement)

$$\mathbf{x}, \mathbf{Y} \quad v = \text{Number of variables}$$

e. g. number of wavelengths

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_o \end{pmatrix} \quad \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1v} \\ y_{21} & y_{22} & \cdots & y_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ y_{o1} & y_{o2} & \cdots & y_{ov} \end{pmatrix}$$

multiple multivariate

Multiple independent variables
(components), multiple dependent
variables (e.g. spectral
measurements)

$$\mathbf{X}, \mathbf{Y} \quad k = \text{Number of components}$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{o1} & x_{o2} & \cdots & x_{ok} \end{pmatrix} \quad \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1v} \\ y_{21} & y_{22} & \cdots & y_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ y_{o1} & y_{o2} & \cdots & y_{ov} \end{pmatrix}$$



- ***Univariate Regression:***
 - all measured values are depending on **one variable** only.
 - In a Multi Component Analysis you may have many components or its concentrations, but you do not have a multi dimensional array of physico-chemical different variables consisting of combinations of retention times, pressures, eluents etc. like in the chromatography.
 - ***Multivariate Data Analysis***
 - More complex relationships can be described
 - The correlation of different kind of variables can be investigated simultaneously
 - The data set can be evaluated at the same time
 - The intention is to classify the data set and to identify unknown properties
 - Requirement: The methods used are operating on the principle “soft modelling” (**parameterised**) or even model free (**non parameterised**).
 - The aim of the Multivariate Data Analysis is to reduce the data set (***data reduction***).
-



- ***Objects***

Samples, Chromatograms, Operating conditions of a plant,
Chemical structures, Patients, Archaeological features

- ***Pattern***

contains the whole range of measurement results for one object.
This means the pattern is a set of features selected for the
description of the objects.

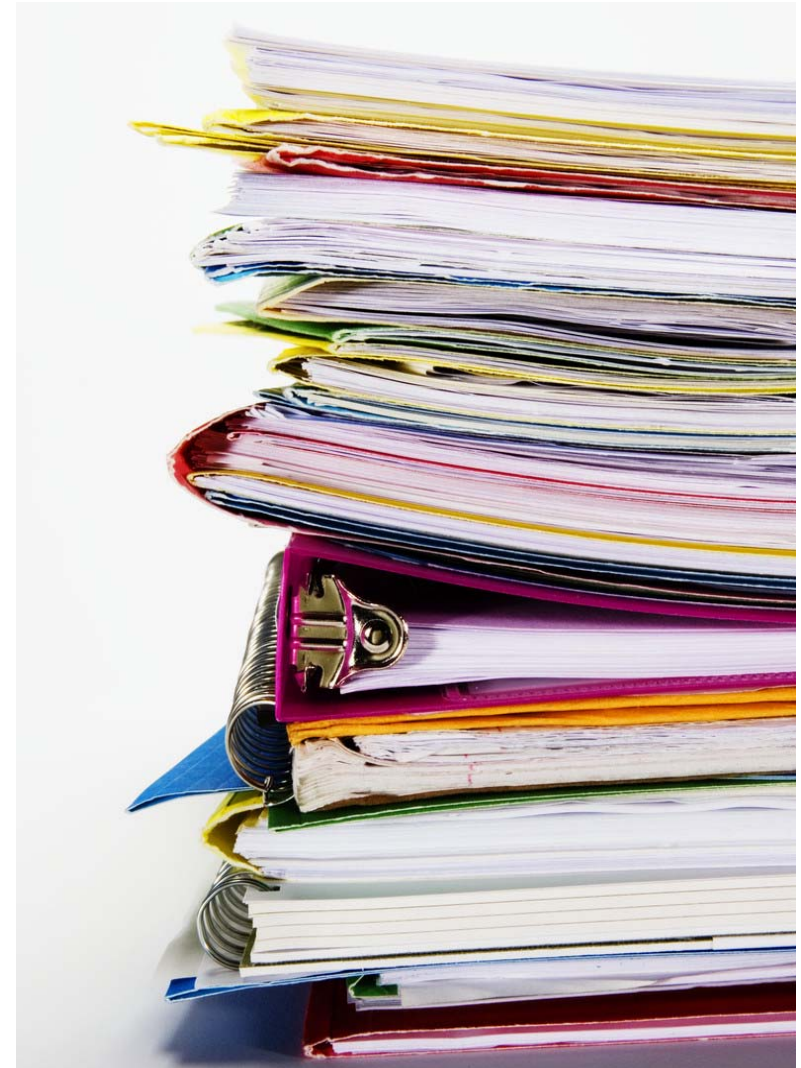
- ***Features***

are variables like e. g. the concentrations of different components
or the peak areas of a chromatogram.



Multivariate data analysis

- Introduction:
Why multivariate data analysis?
- Evaluation methods
 - **Classification:** Cluster analysis, PCA
 - **Regression:** univariate, MLR, PCR





- Interpretation and inspection of analytical data
 - Graphical representation
 - Correlation and identification of significant variables
- Evaluation of analytical data
 - Classification
 - Regression
- In total, there is a huge variety of methods. You will be introduced into the most important methods in the following.



- **A data matrix $\mathbf{X}_{(n,p)}$ contains**

- the objects (= measurements / samples) in the rows (1...n) and
- the features in the columns (1...p).

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

n = number of objects, p = number of features



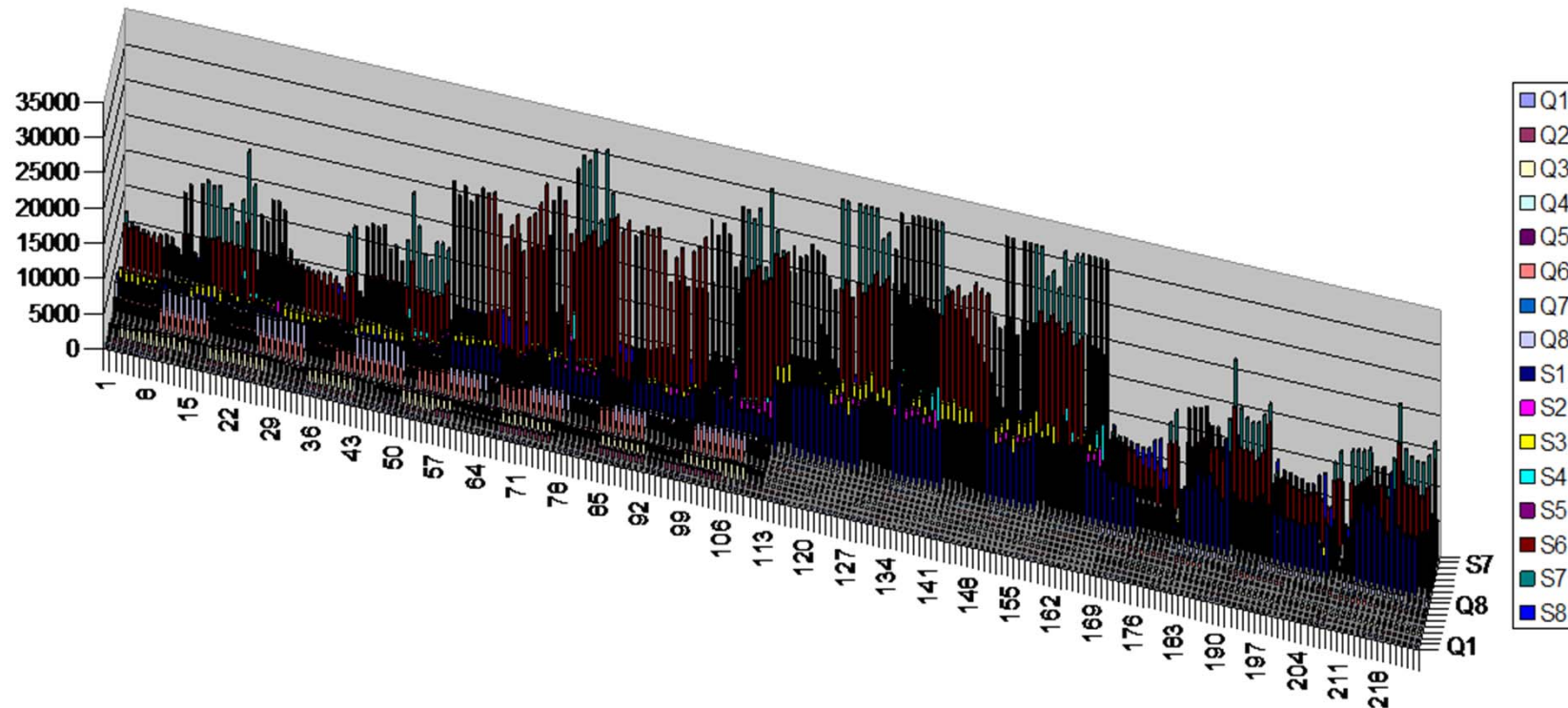
Practical example: MOSES II data

- Data acquired with a MOSES II system containing 8 QMB (Quartz crystal microbalance) and 8 MOS (metal oxide sensor) sensors
- 4 different analytes:
 - Anisol
 - Cyclohexanon
 - Propanol
 - Toluene
- Measurements repeated over a time of 23 months
→ 224 Objects x 16 Features





Practical example: Birds eye perspective



A view on the data from a birds eye perspective.

Where is the information?



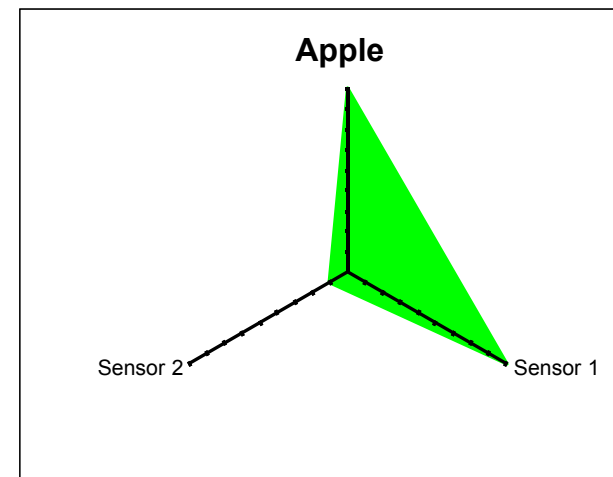
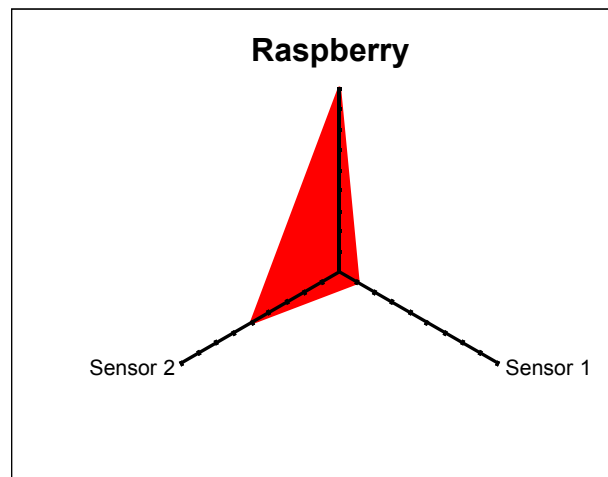
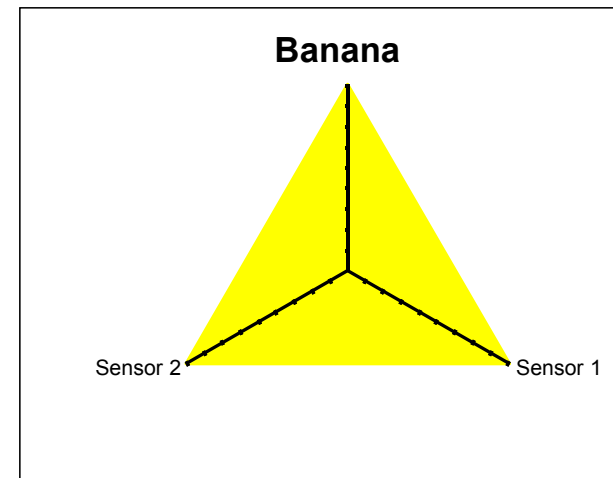
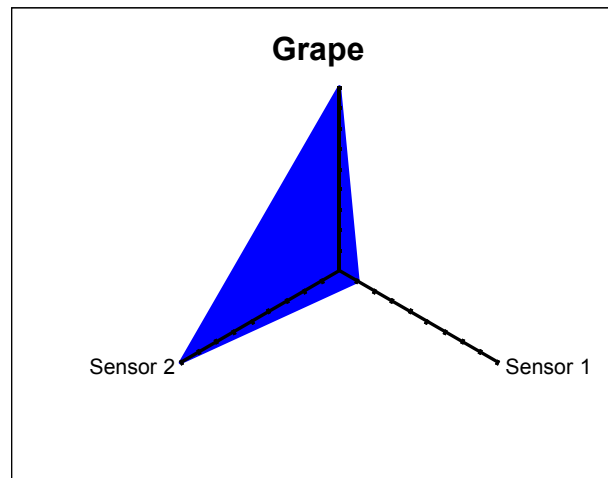
Practical example: numerical data

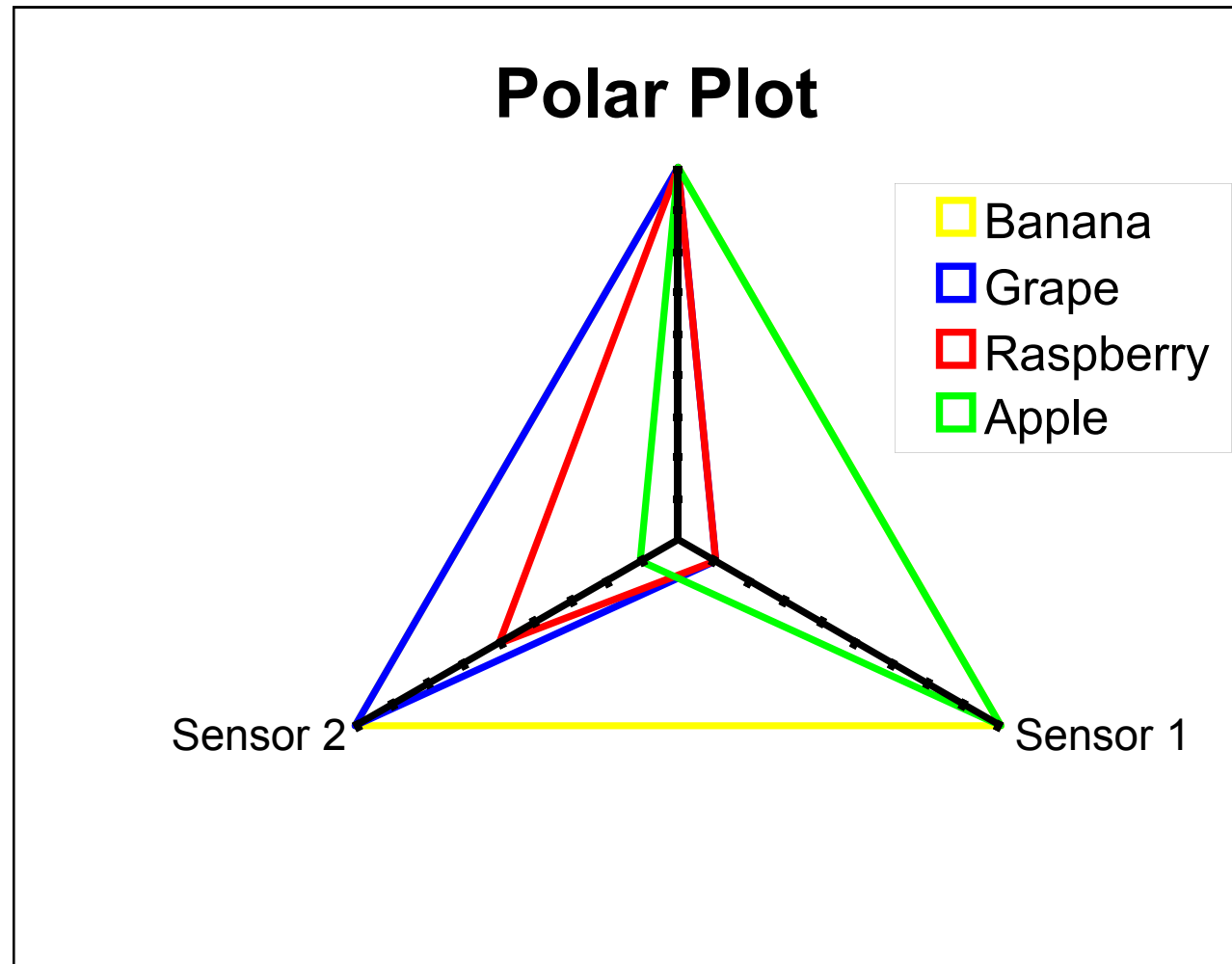
- Looking at the individual values won't bring us additional insights...

	Q1	Q2	Q3	Q4	Q5
Anisol 500 1 #1 (10),mos	319,67	475	1615,67	697	479,67
Anisol 500 1 #1 (11),mos	314	470,33	1581,33	681,67	471,67
Anisol 500 1 #1 (12),mos	317,33	475,33	1602,33	691,33	474,67
Anisol 500 1 #1 (13),mos	312,33	468,33	1562,33	678	468,67
Anisol 500 1 #1 (14),mos	308,67	462	1531,33	664	459,67
Anisol 500 1 #1 (15),mos	311,67	469,33	1661	672	468,67
Anisol 500 1 #1 (16),mos	313,33	473	1662,67	680,67	470,67
Anisol 500 1 #1 (17),mos	336,67	512	1831,67	752,67	515,67
Anisol 500 1 #1 (18),mos	311	469,67	1658,33	680,33	462,67
Anisol 500 1 #1 (19),mos	322,33	507,33	1732,67	757,33	507,67
Anisol 500 1 #1 (2),mos	328,67	482	1900,67	716	511,67
Anisol 500 1 #1 (2),mos	337,33	500	1942	746	531,67
Anisol 500 1 #1 (20),mos	305	483,67	1615,67	730,33	467,33
Anisol 500 1 #1 (3),mos	331	488,33	1872,33	732	507,67
Anisol 500 1 #1 (4),mos	327,67	494	1877,67	726,33	503,67
Anisol 500 1 #1 (5),mos	316,67	484	1817,33	722	505,67
Anisol 500 1 #1 (6),mos	307,67	473	1626,33	698,33	492,33
Anisol 500 1 #1 (7),mos	310,67	473	1622,67	700,67	475,67
Anisol 500 1 #1 (8),mos	326,33	503,33	1760,33	753	509,67
Anisol 500 1 #1 (9),mos	344	518,67	1795	775,33	531,67
Anisol 500 1 #1,mos	327,67	483,67	1885	716,67	512,67
Anisol 500 1 #1,mos	299	364	1846	655,67	458,67
Anisol 500 1 #2 (2),mos	323	467,33	1923,33	703,33	507,67
Anisol 500 1 #2 (3),mos	322	423	1900,67	707,67	501,67
Anisol 500 1 #2 (4),mos	315,33	362	1839,67	691,67	491,33
Anisol 500 1 #2 (5),mos	316	446,67	1839,33	683,33	488,67
Anisol 500 1 #2,mos	312	345,33	1847,67	679,33	480,33



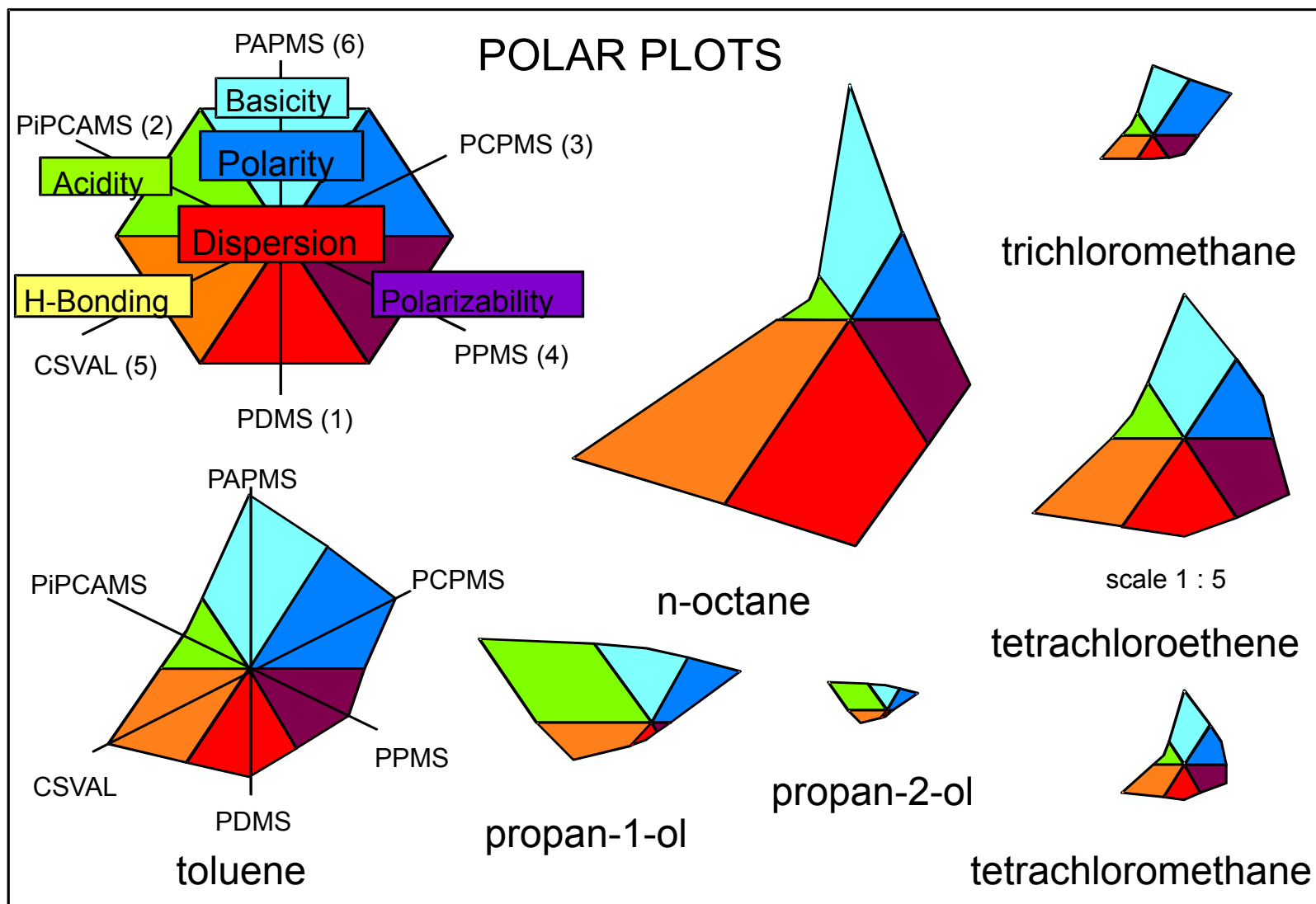
Polar Plot







Polar plots





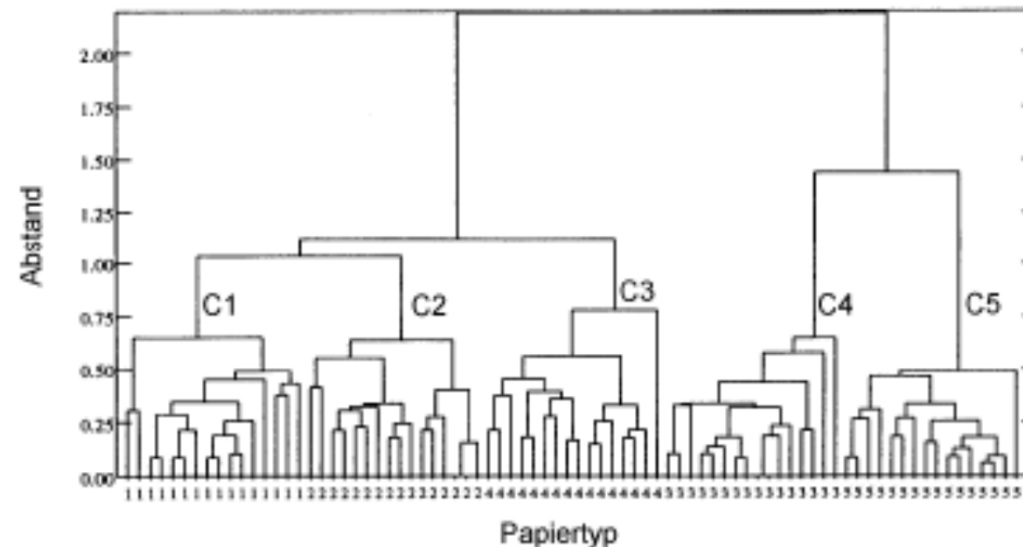
Unsupervised training: Cluster analysis (qualitative Information)

- Cluster analysis is a method which groups objects of a data set into homogeneous classes (**Cluster**).
- Objects with similar variable patterns are grouped into the same classes.
- The groups are unknown beforehand.
 - automatic / unsupervised classification
- Look for objects which are close to each other in the variable space.
- Calculate the distance dimensions in order to group the objects
- Develop a **dendrogram**



Hierarchical, agglomerative

- Each element (measured value) is considered as a separate cluster
- The separate clusters are successively merged together to larger clusters beginning with the closest clusters (non reversible)





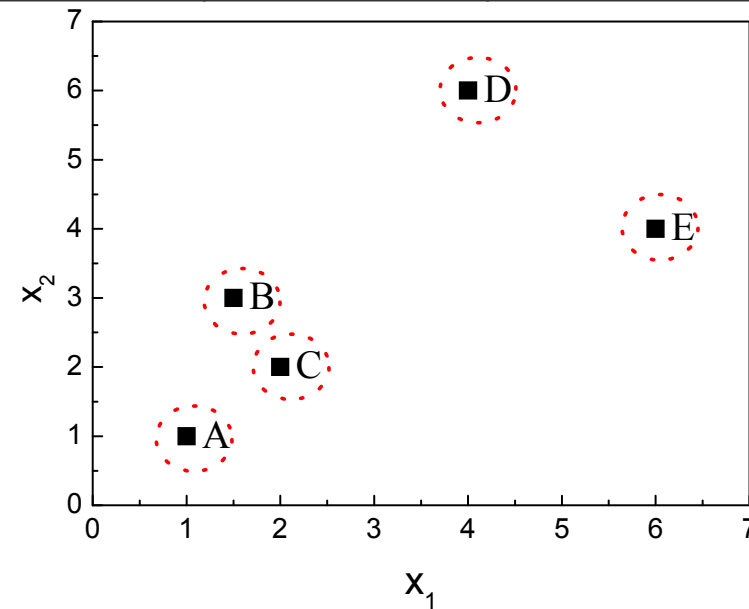
Hierarchical cluster analysis: Example

1. Plot the objects in the variable space.
2. Look for groups / clusters between the objects

Object	x_1	x_2
A	1	1
B	1.5	3
C	2	2
D	4	6
E	6	4

Cluster algorithm:

1. Assume that every object is a “Cluster” with the size “1”.





Hierarchical cluster analysis: Cluster algorithms

2. Compare the distances of the individual clusters with each other and look for the smallest distance d_{ij}

Euclidean distance:

Distance between two objects i and j in the n -dimensional space with the coordinates $(x_{i1}, x_{i2}, \dots, x_{ip})$ and $(x_{j1}, x_{j2}, \dots, x_{jp})$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

3. Group the two closest objects to a new cluster

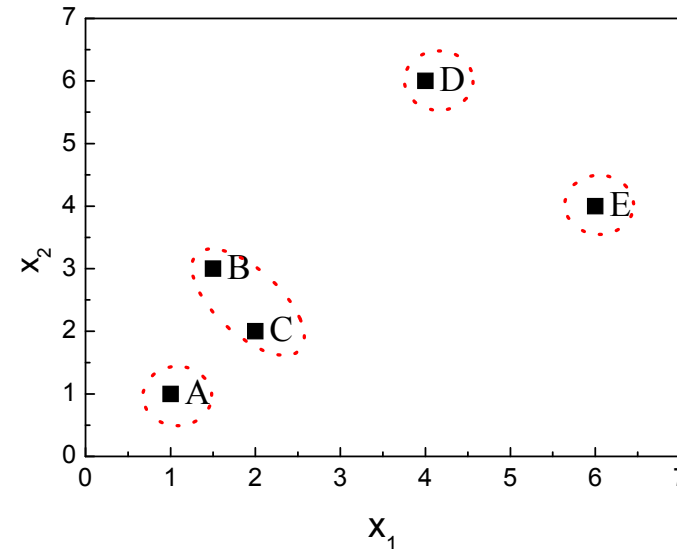
here :

$$d_{AB} = \sqrt{(1 - 1,5)^2 + (1 - 3)^2} = \sqrt{4,25}$$

$$d_{AC} = \sqrt{(1 - 2)^2 + (2 - 1)^2} = \sqrt{2}$$

$$d_{BC} = \sqrt{(1,5 - 2)^2 + (3 - 2)^2} = \sqrt{1,25}$$

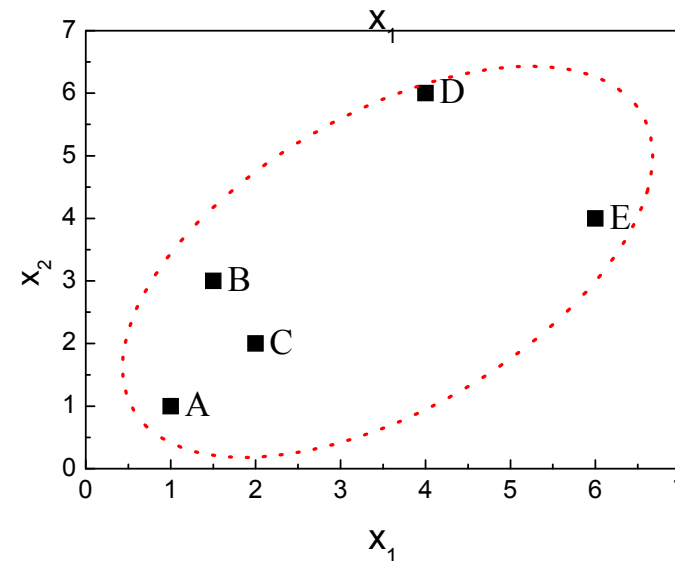
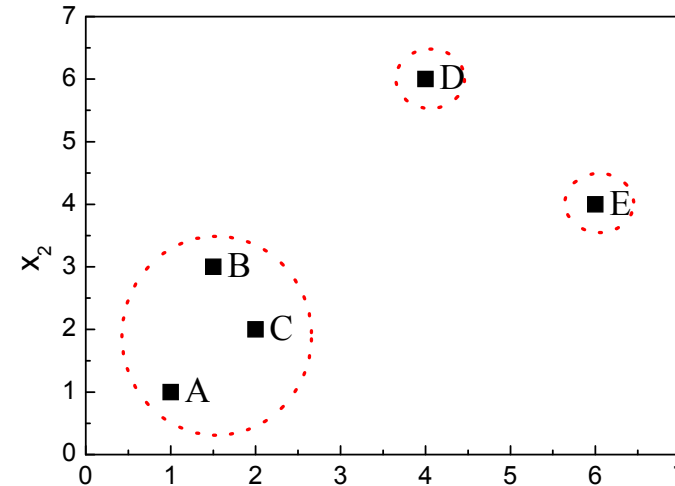
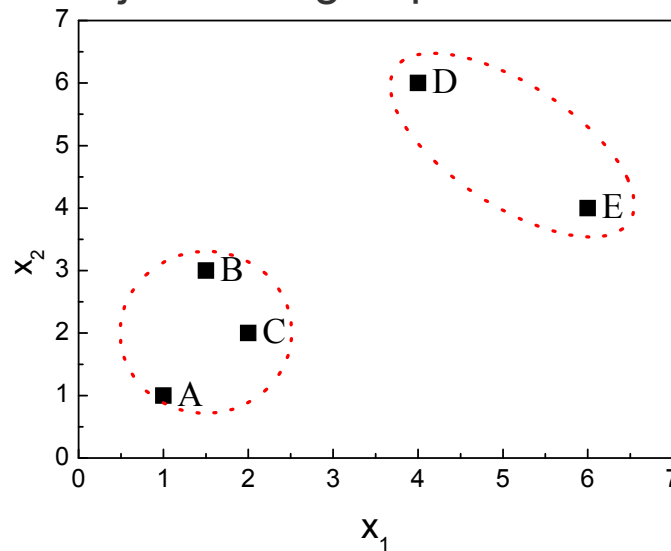
$$d_{DE} = \sqrt{(5 - 6)^2 + (6 - 4)^2} = \sqrt{5}$$





Hierarchical cluster analysis: Cluster algorithms

4. Repeat step 2: Compare the distances of the individual clusters with each other and look for the smallest distance d_{ij}
5. Group the two closest clusters to a new cluster
6. Repeat the procedure until all objects are grouped in a cluster.



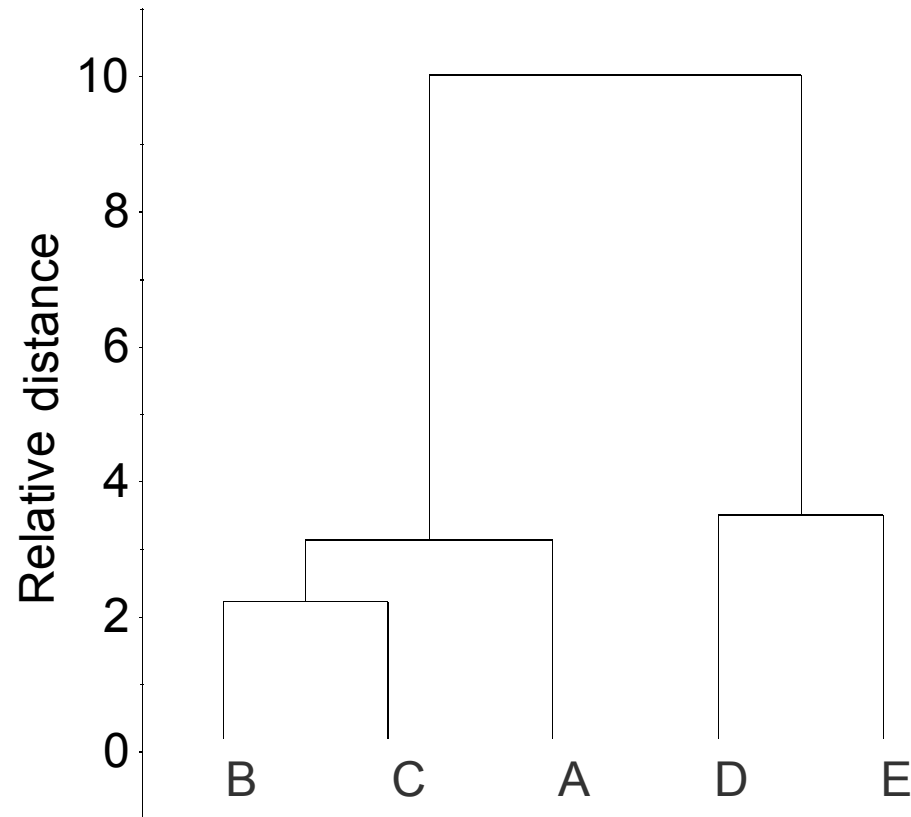


Hierarchical cluster analysis: Dendrogram

- Plot of the successive objects grouping in a dendrogram
- Vertical axis:
 - Relative distance d_{ij} between two objects
 - Similarity s_{ij}

$$s_{ij} = 100 \cdot \left(1 - \frac{d_{ij}}{d_{\max}} \right)$$

$d_{\max}$... maximum distance
between two objects





Some more distance dimensions...

- Minkowski distance
(general distance dimension)

$$d(x_i, x_j) = \left(\sum_{k=1}^p (x_{ik} - x_{jk})^p \right)^{1/p}$$

- City Block / Manhattan distance
(corresponds to the way around a city block;
Special case of the Minkowski distance:
 $p=1$)

$$d(x_i, x_j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

- Euclidean distance
(geometric distance; $p=2$)

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

- Mahalanobis distance

Takes the different dimension of the variables and there correlation into account

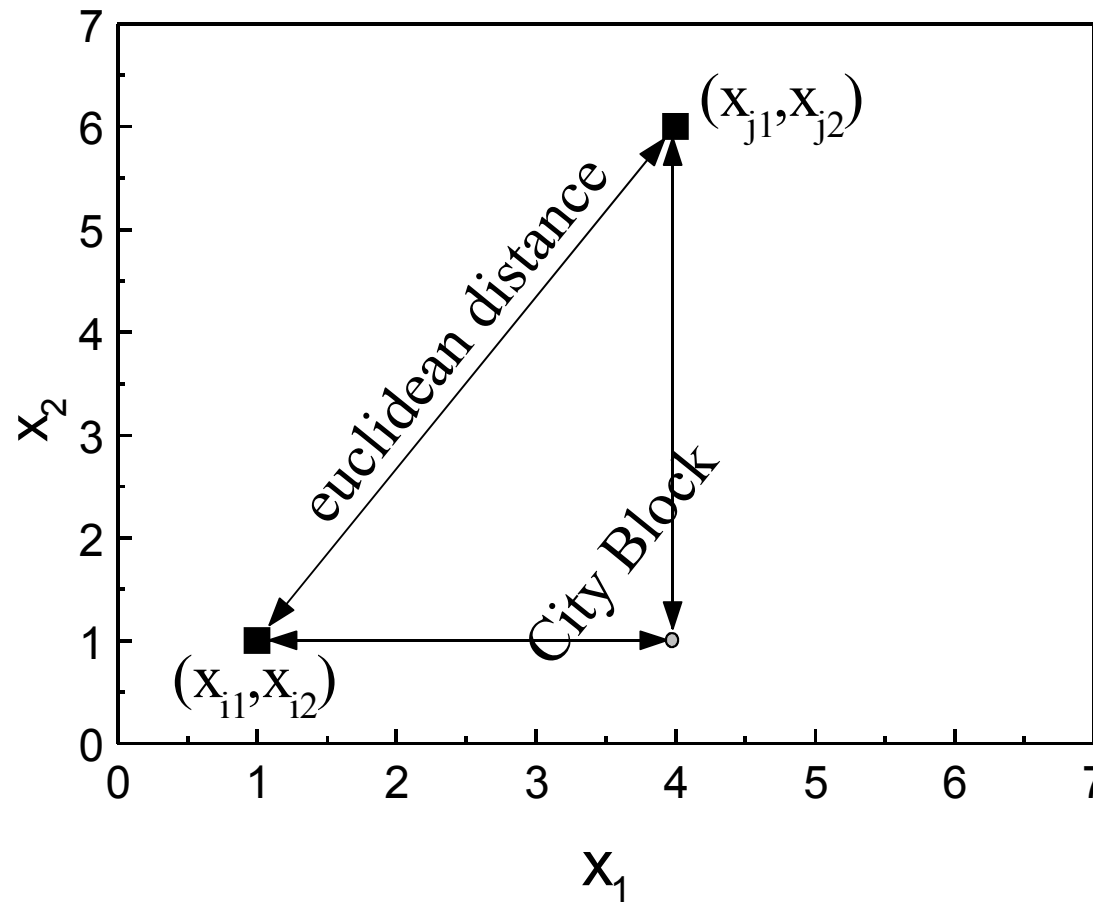
$$D_{ij}^2 = (x_i - x_j)^t C^{-1} (x_i - x_j)$$

C is the covariance matrix

Disadvantage: All distance dimensions are depending on the used dimensional / measuring units.



Some more distance dimensions... graphical representation





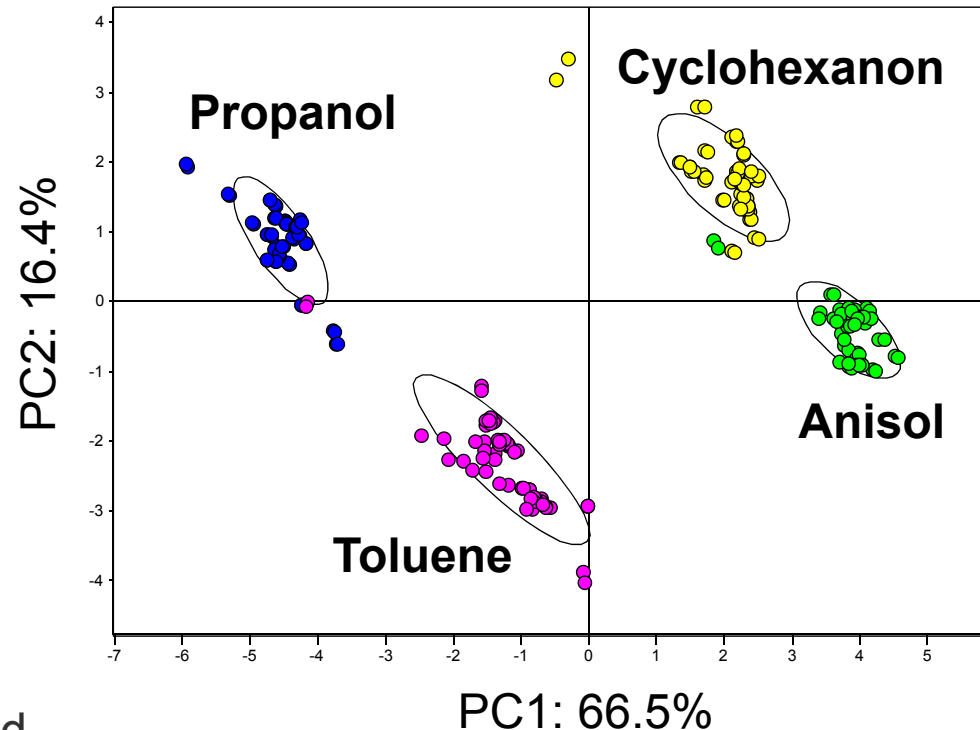
Principal component analysis (PCA): Basic aspects

break down of
more
dimensions in
to 2 or 3
dimensions

covariance show
correlation between
the different
dimensions,
PCA helps to get rid
of noise/ information
that is not needed

The PCA scores plot
represents
a **2-dimensional** view on the
data, in which

- the maximum of the variance is contained
- while preserving the relative distances between the objects.
- Often the amount of information represented by one principal component (PC) is printed in the axis label.



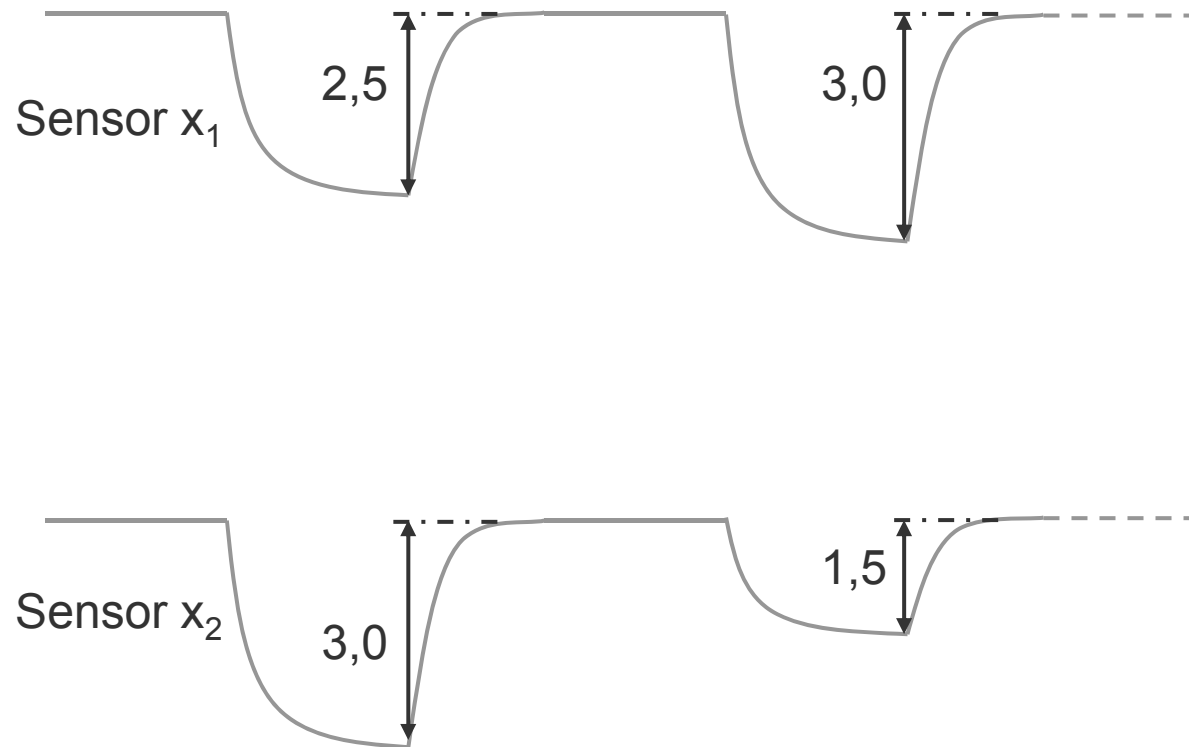
coordination
transformation!!!!

- The principal components are always sorted by their information content
- PC1 is the principal component with most information (**variance**, here: 66.5 %)

variance is good, distribution of the components



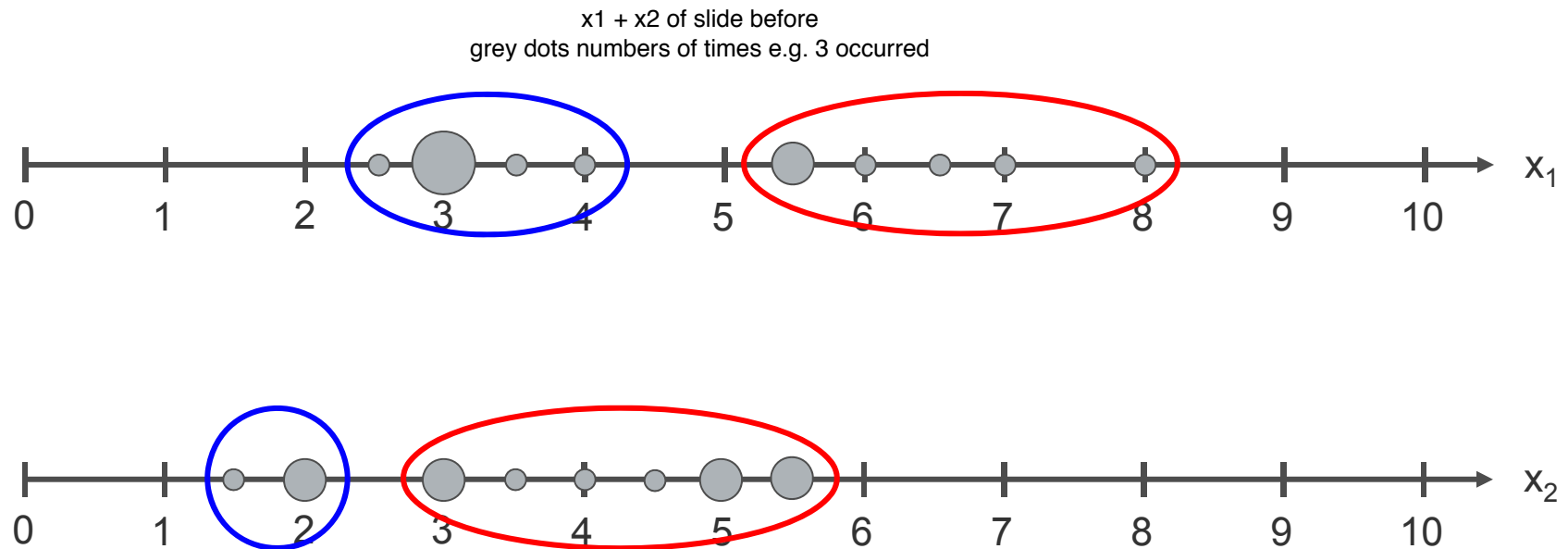
PCA: Procedure



x_1	x_2
2.5	3.0
3.0	1.5
4.0	2.0
5.5	5.5
7.0	4.0
3.0	3.5
6.5	4.0
3.5	2.0
8.0	5.0
5.5	5.0
3.0	3.0
6.0	5.5



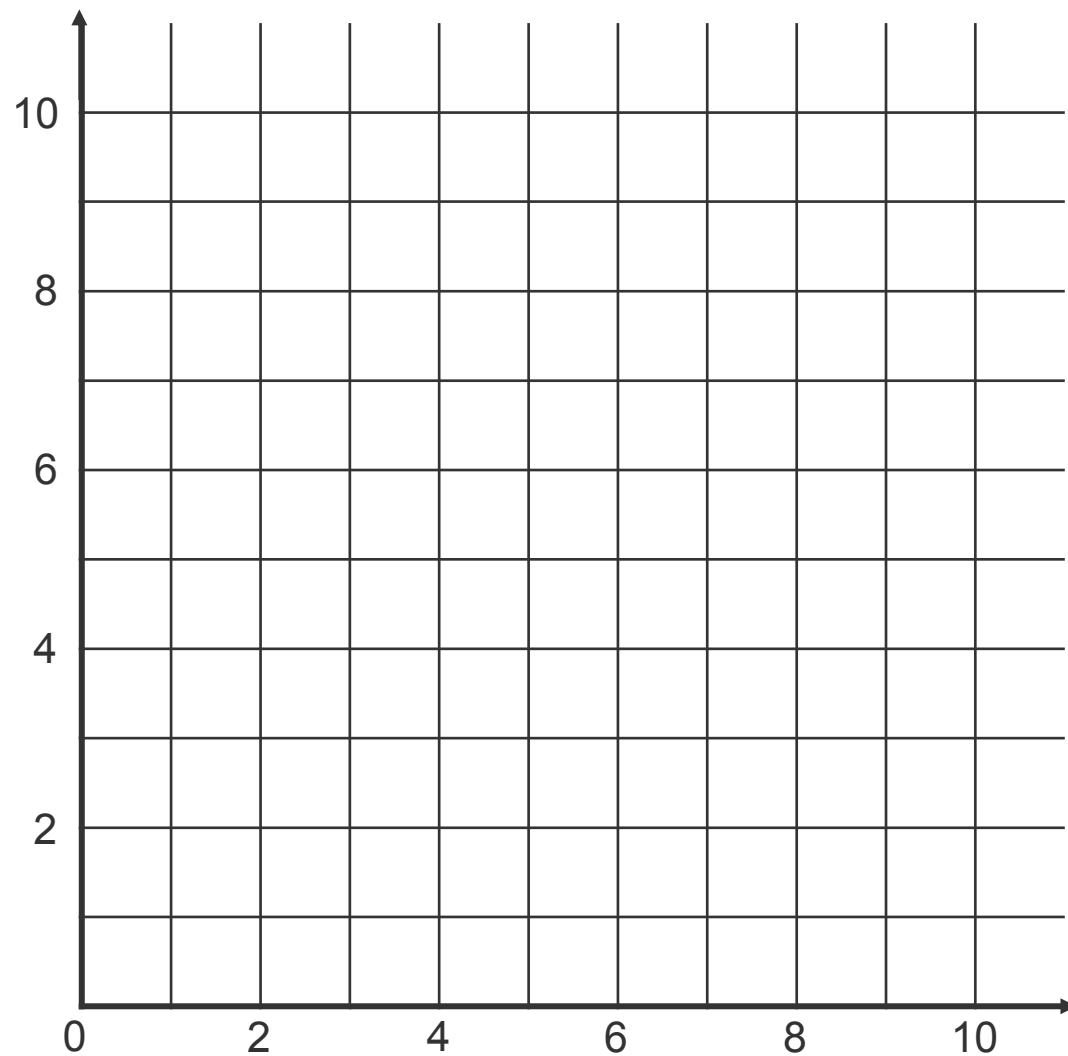
One dimensional mapping





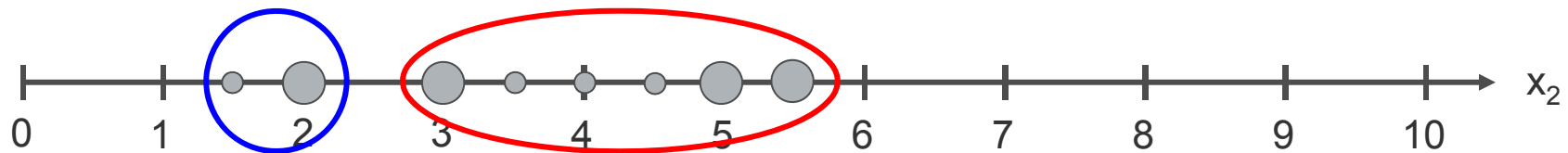
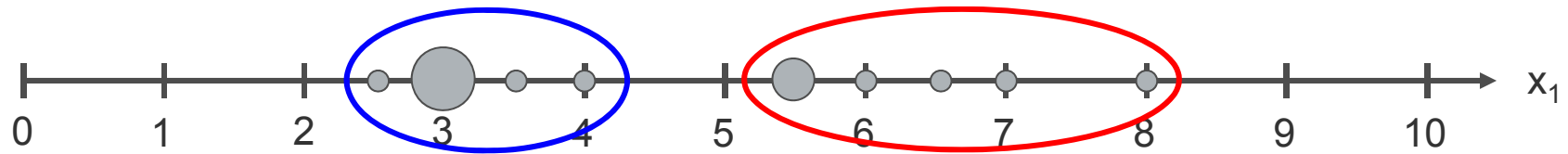
A small exercise

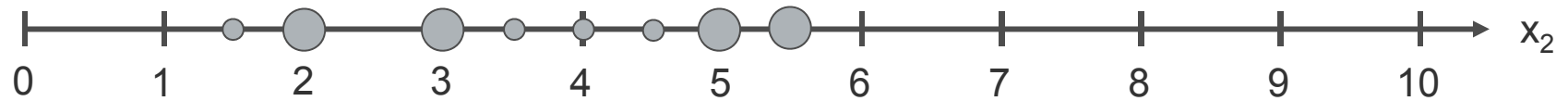
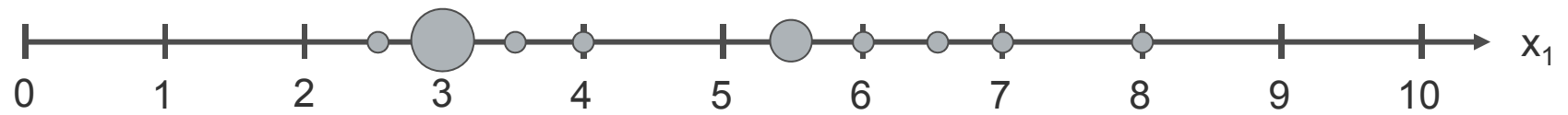
x_1	x_2
2.5	3.0
3.0	1.5
4.0	2.0
5.5	5.5
7.0	4.0
3.0	3.5
6.5	4.0
3.5	2.0
8.0	5.0
5.5	5.0
3.0	3.0
6.0	5.5





One dimensional mapping



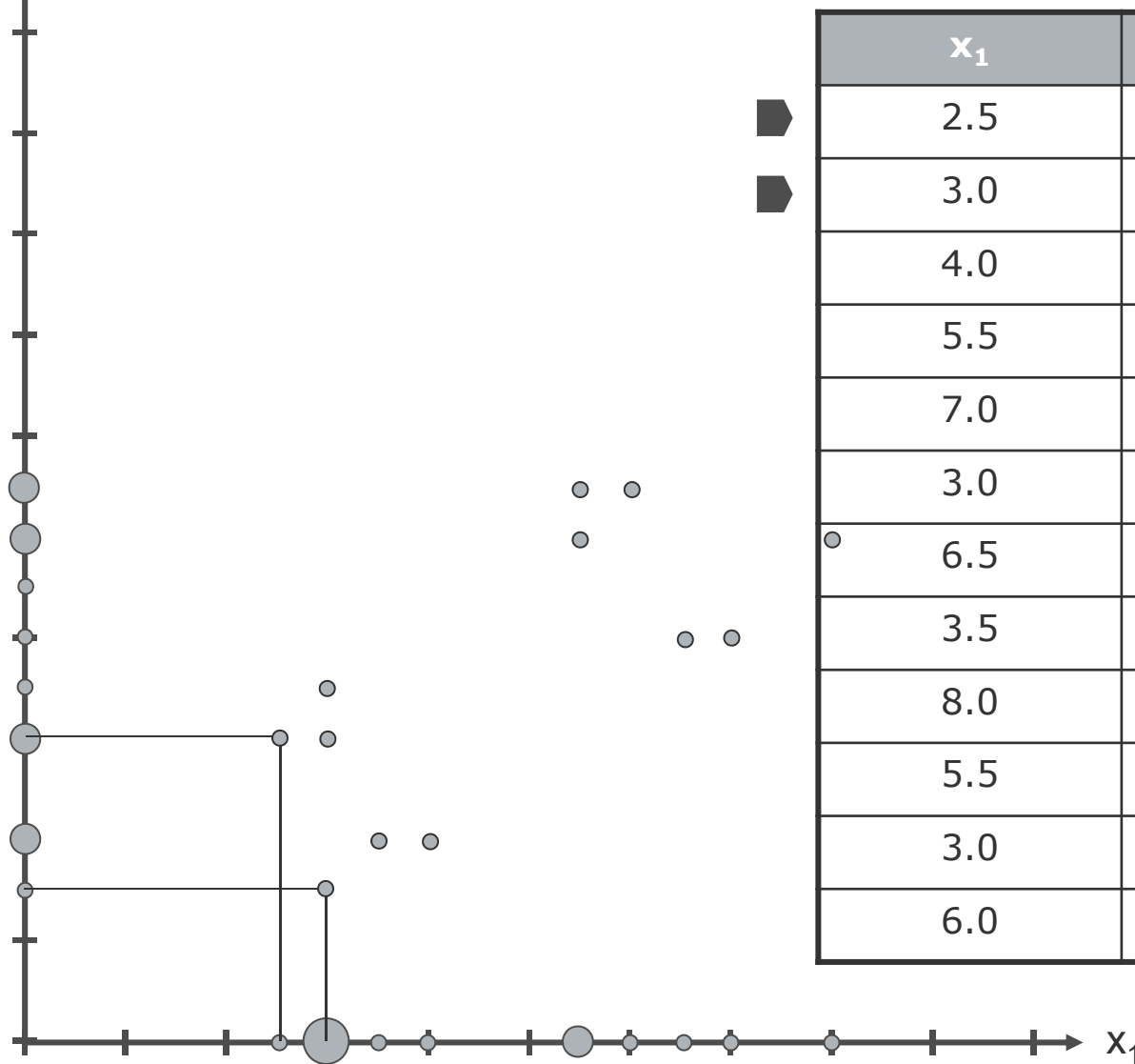


x_1

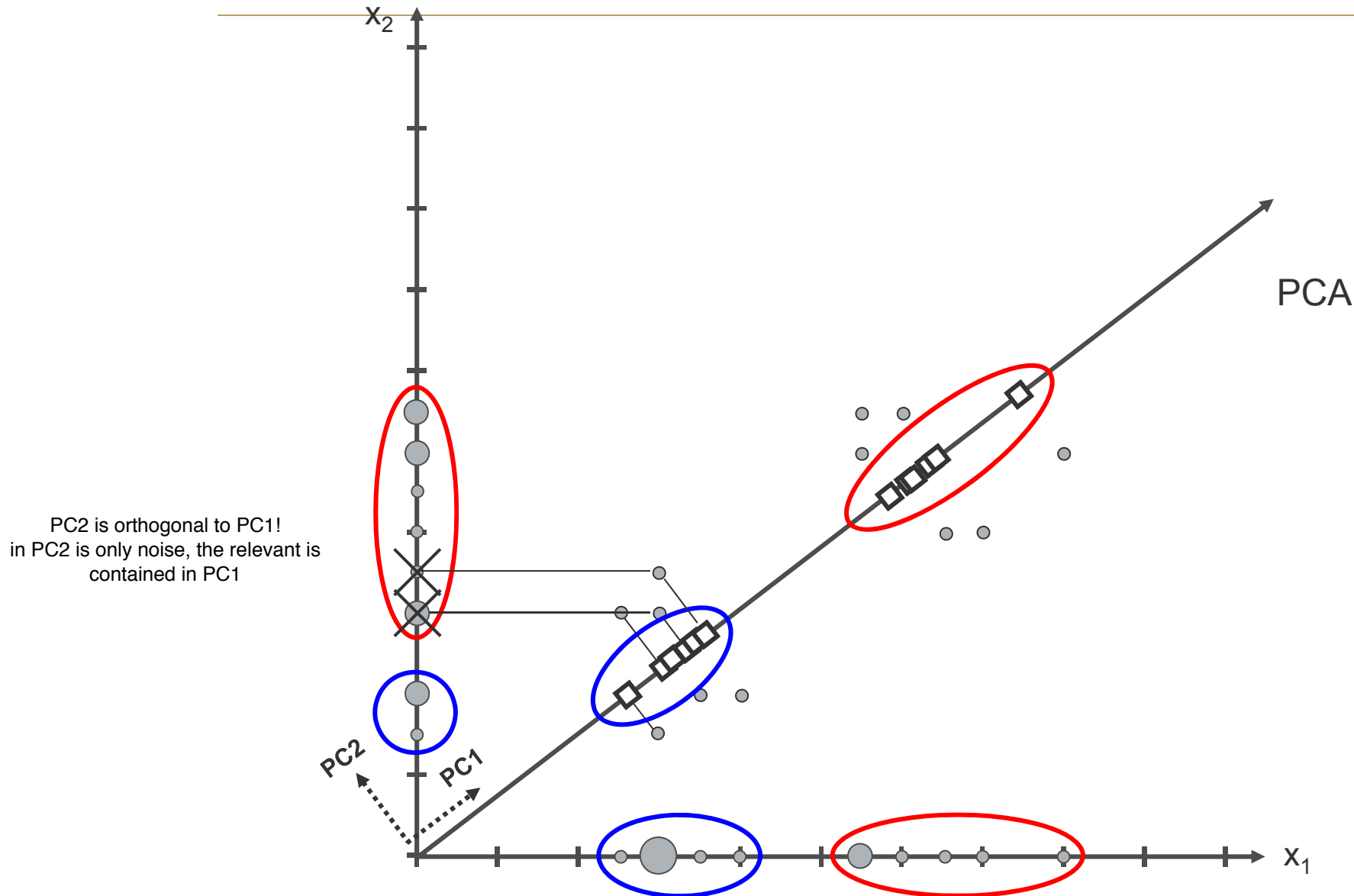


x_2

find new coordinate that is
representing the data the best ->
coordinate transformation! next slide!



x_1	x_2
2.5	3.0
3.0	1.5
4.0	2.0
5.5	5.5
7.0	4.0
3.0	3.5
6.5	4.0
3.5	2.0
8.0	5.0
5.5	5.0
3.0	3.0
6.0	5.5





In PCA a data matrix $\mathbf{X}$ is decomposed (e.g. via SVD or NIPALS) to:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad \mathbf{S} = \mathbf{diag}(\sqrt{\lambda_i})$$

Where:

$\mathbf{U}$ is a matrix containing the Eigenvectors of $\mathbf{X}\mathbf{X}^T$

$\mathbf{V}$ is a matrix (called loadings) containing the Eigenvectors of $\mathbf{X}^T\mathbf{X}$

$\mathbf{S}$ is a diagonal matrix containing the square roots of the Eigenvalues λ .



Please keep in mind:

$\mathbf{X}^T\mathbf{X}$ corresponds to the covariance or correlation matrix of $\mathbf{X}$ (depending on the scaling of the variables).

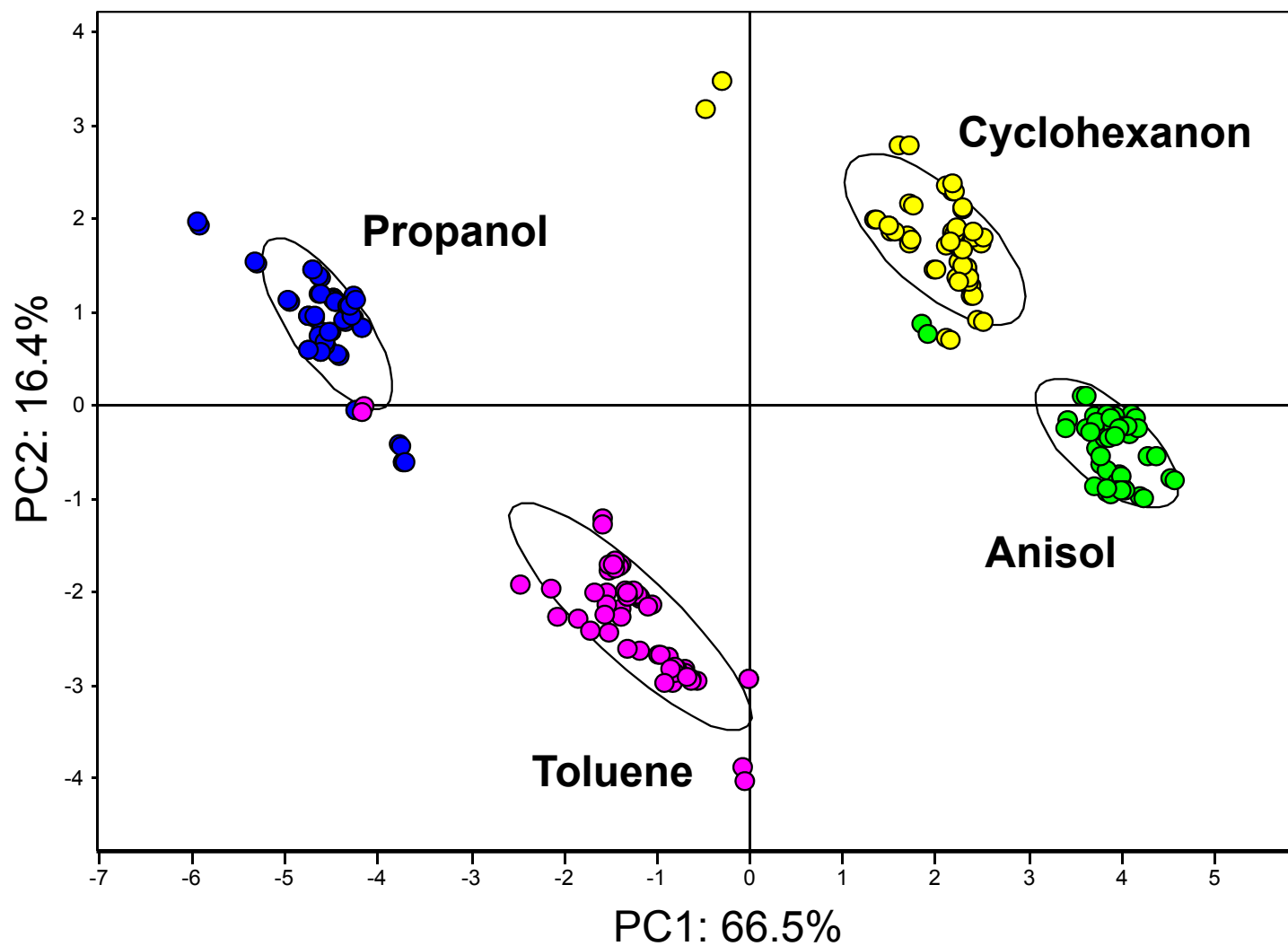
$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad \mathbf{S} = \mathbf{diag}(\sqrt{\lambda_i})$$

- The scores of the objects in $\mathbf{X}$ are simply the eigenvectors in $\mathbf{U}$ multiplied by $\mathbf{S}$. (The eigenvectors in $\mathbf{U}$ und $\mathbf{V}$ are normalized to the length 1).
- To calculate the scores $\mathbf{t}$ for a new object, we multiply the vectors of the elements $\mathbf{x}$ with the loadings $\mathbf{V}$:

$$\mathbf{t} = \mathbf{V}\mathbf{x}$$



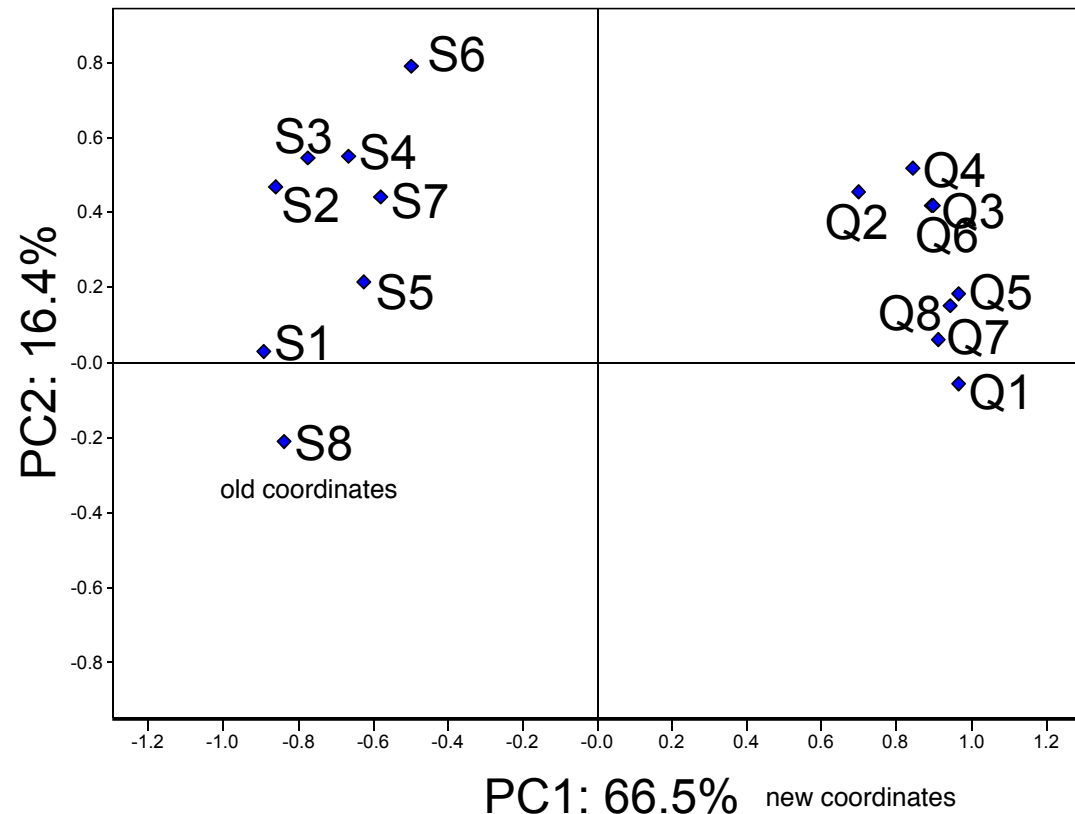
PCA: Scores





PCA: Loadings plot

- The loadings plot shows the influence of each feature onto the principal components(PC's)
- Features not contributing will be near the centre.
- Features with similar information content will be represented next to each other.
- (Compare: Cluster analysis)



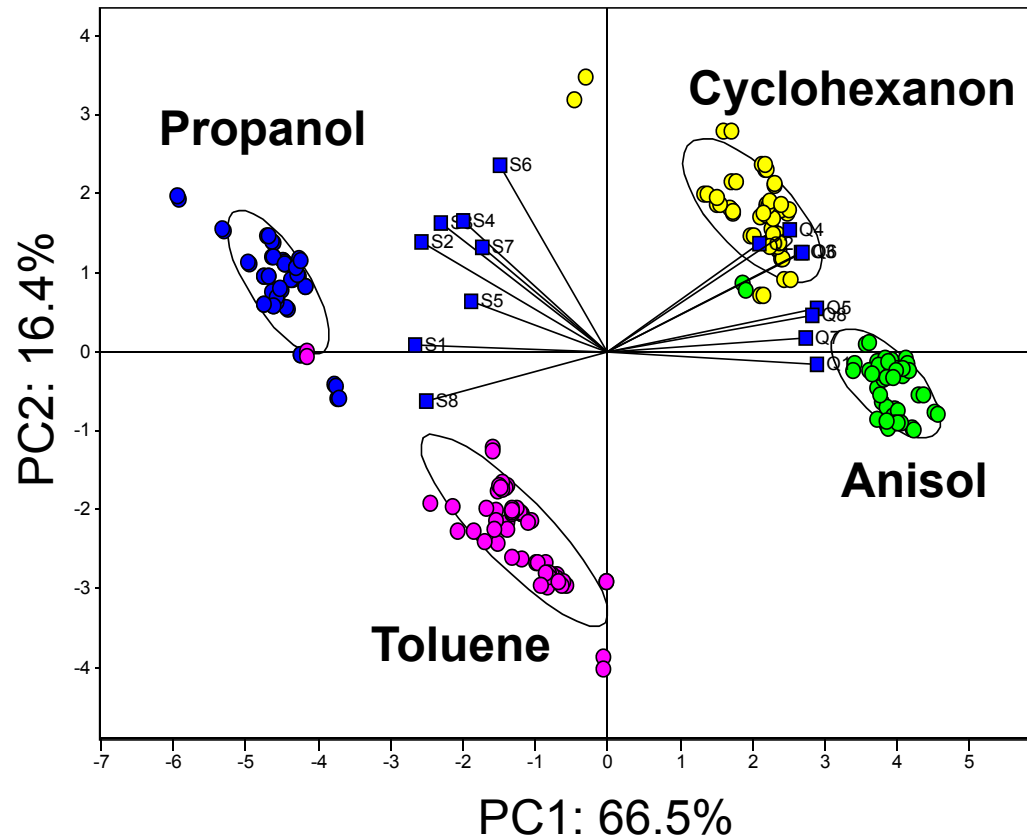
- S = Metal oxide sensor
 - Q = Quartz crystal microbalance sensor
- S1-8
Q1-8



PCA: Biplot of scores and loadings

In a biplot, scores and loadings are shown (with adequate scaling)

By comparing the directions of the loadings with the position of the objects it is possible to draw conclusions on the suitability of individual features.



only 180°C bc the other one can multiplied by -1

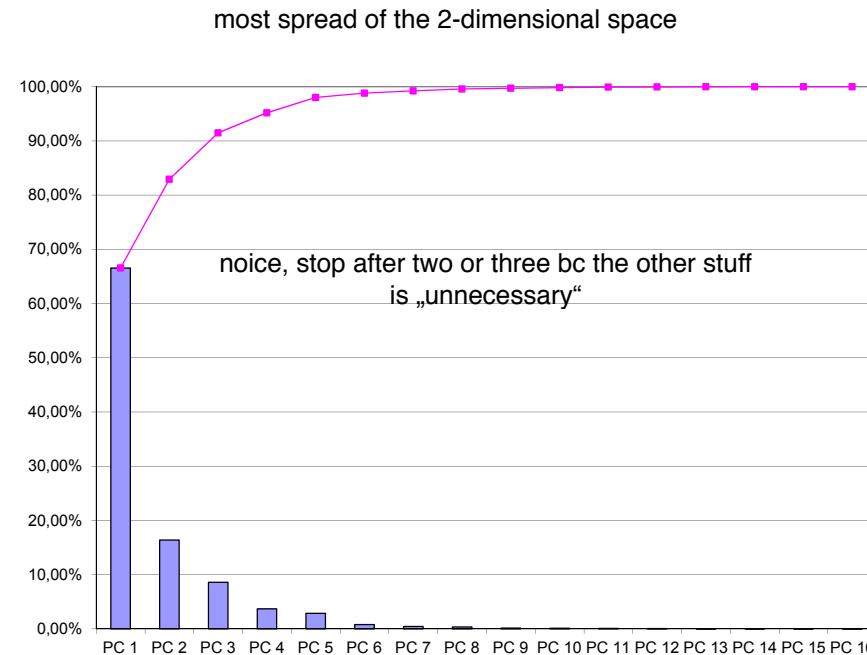


PCA: Scree plot

The scree plot shows the amount of information (variance) represented by each principal component (bars)

It is also possible to show the information content covered by the first x components (line)

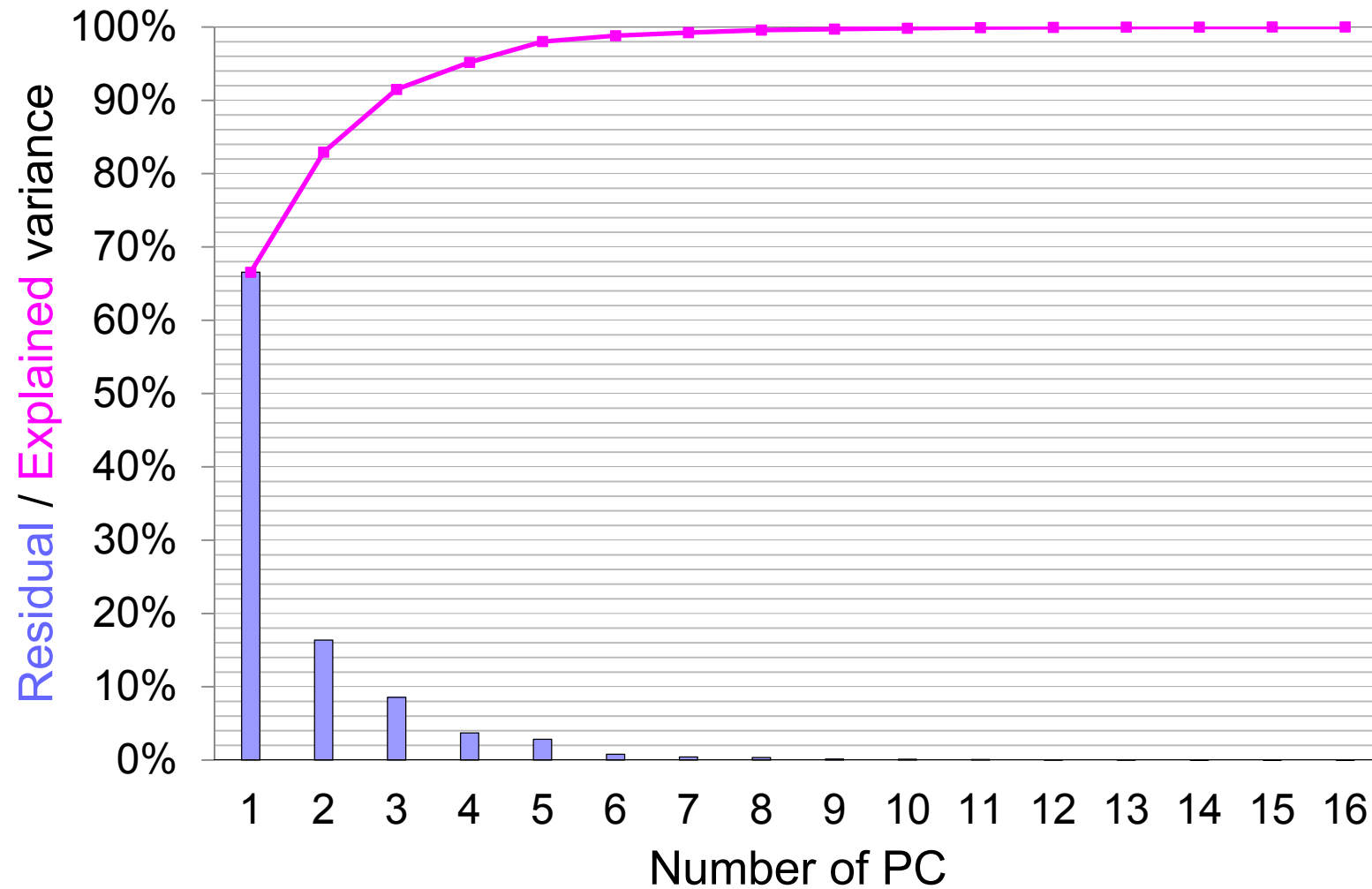
The information content of a principal component is the size of its Eigenvalue λ , scaled in a way, that the sum of all Eigenvalues is 1.



The scree plot helps to identify the number of components, necessary to explain the data. In above plot, the first 3 components amount to more than 90% of all information.



PCA: Scree plot





Use of PCA for:

- Fast inspection of the data
 - Outliers
 - Drift
 - Main grouping
- Unbiased representation of data
- Identification of feature structure
 - Similarities
 - Influences on objects
- Reduction of the dimensionality of the data
- Removing of noise
- Working with small datasets



- PCA is an unsupervised method! It does not take into consideration any knowledge about the objects themselves (class membership, Y-data,...)
- If a large amount of noise / drift is present, PCA will not unveil the desired information.
- PCA does not identify the optimum feature composition for classification or regression tasks.
- PCA requires a careful weighting of the input variables.
- PCA plots may be misleading when the dimensionality of the problem is high.



- Understanding of the terms model based and model free, qualitative and quantitative
- Objects, Patterns and Features
- Application of a simple hierarchical clustering
- Verbal explanation (not mathematical) of PCA
- Understanding of scores and loadings



END of Part 02