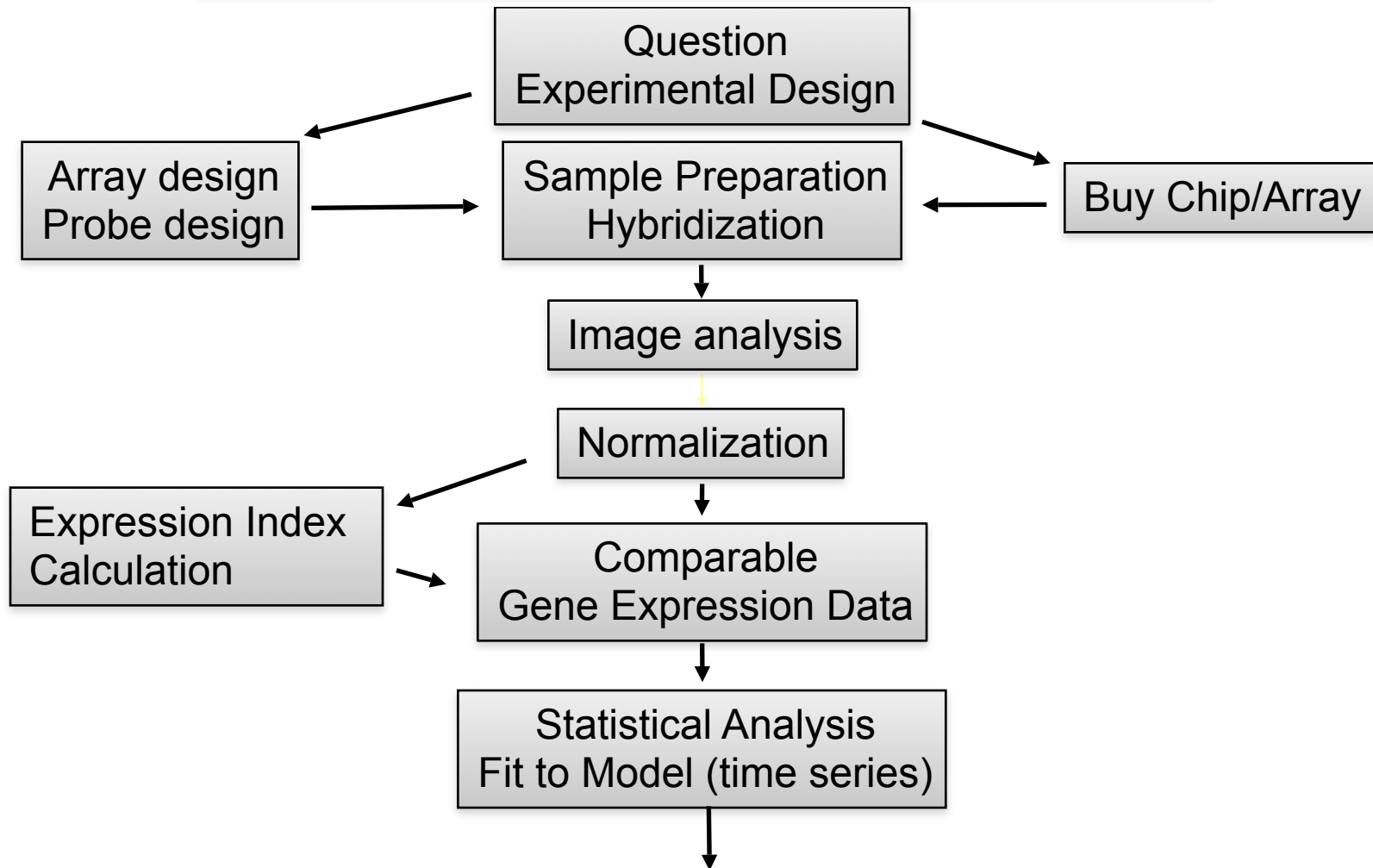


Dimension reduction : PCA and Clustering

By Hanne Jarmer

Slides by Christopher Workman
Center for Biological Sequence Analysis
DTU

The DNA Array Analysis Pipeline



Advanced Data Analysis

Clustering PCA Classification Promoter Analysis
Meta analysis Survival analysis Regulatory Network

CENTER FOR
RIBOBIOTIC
CALCULUS
ENZYME
ANALYSIS **CBS**



What is Principal Component Analysis (PCA)?

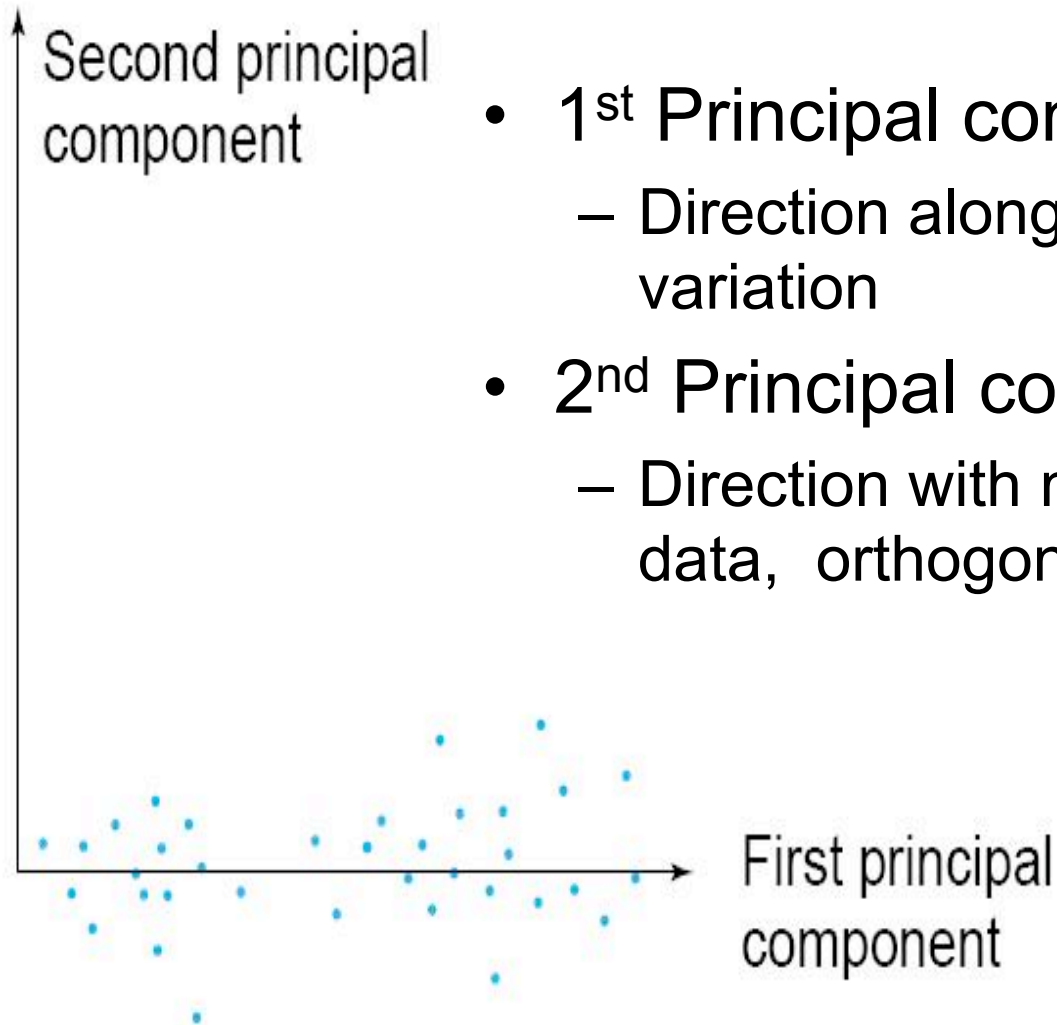


- Numerical method
- Dimensionality reduction technique
- Primarily for visualization of arrays/samples
- "Unsupervised" method used to explore the intrinsic variability of the data

PCA

- Performs a rotation of the data that maximizes the variance in the new axes
- Projects high dimensional data into a low dimensional sub-space (visualized in 2-3 dims)
- Often captures much of the total data variation in a few dimensions (< 5)
- Exact solutions require a fully determined system (matrix with full rank)
 - i.e. A “square” matrix with independent rows

Principal components



- 1st Principal component (PC1)
 - Direction along which there is greatest variation
- 2nd Principal component (PC2)
 - Direction with maximum variation left in data, orthogonal to PC1

Singular Value Decomposition

- An implementation of PCA
- Defined in terms of matrices:

$$X = USV^T$$

X is the expression data matrix

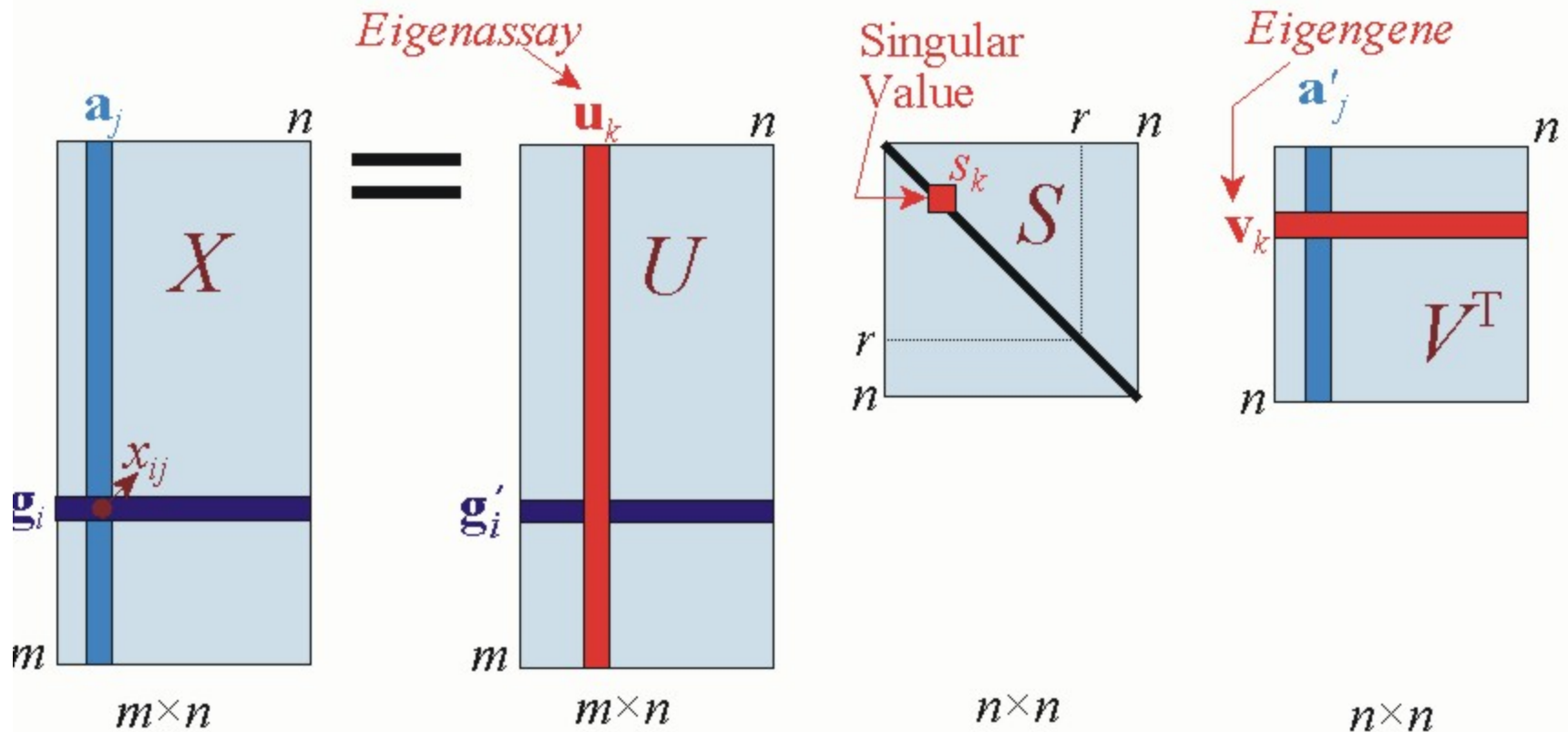
U are the left eigenvectors

V are the right eigenvectors

S are the singular values ($S^2 = \Lambda$)

Singular Value Decomposition

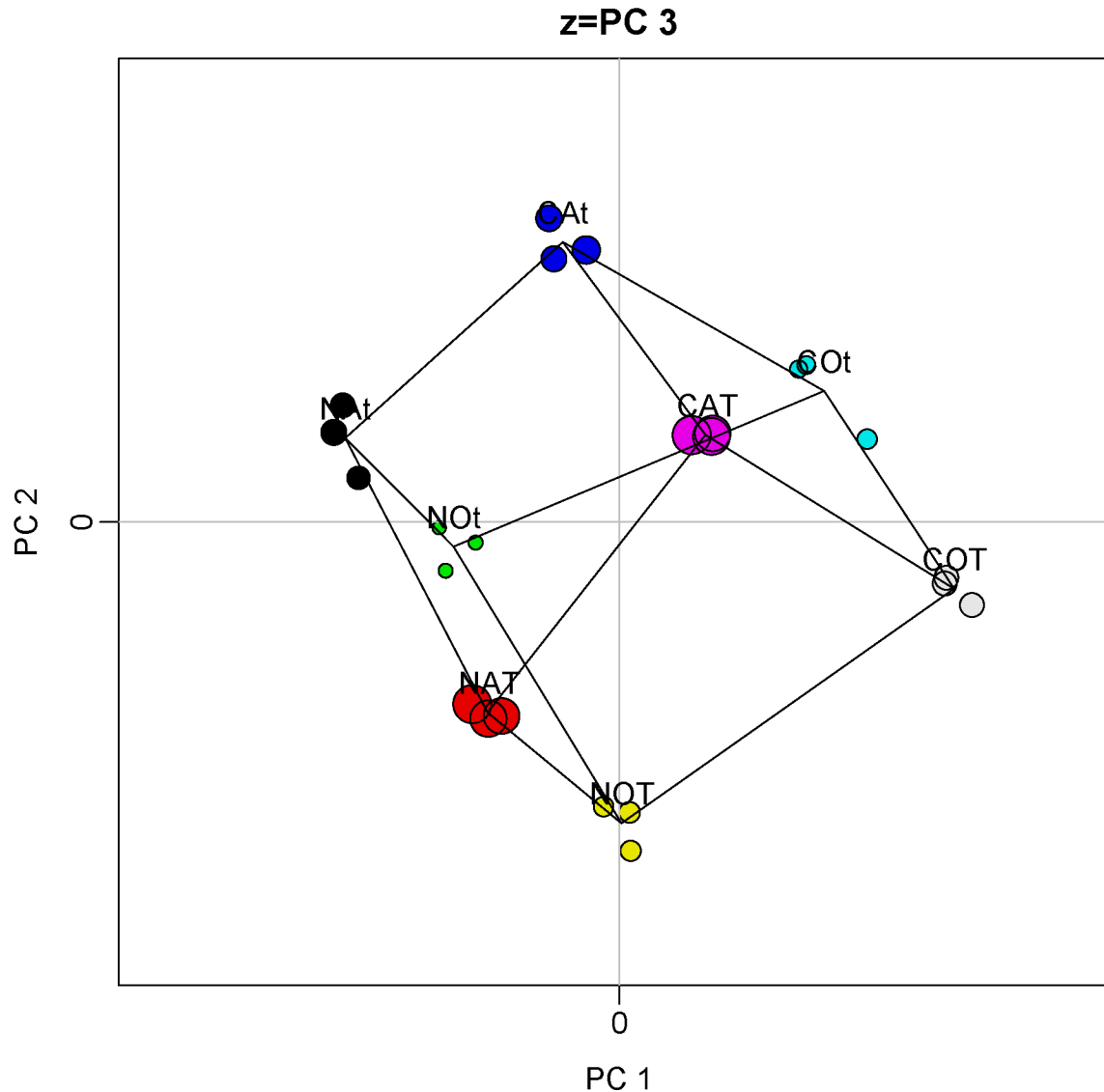
$$X = USV^T$$



Singular Value Decomposition

- Requirements:
 - No missing values
 - “Centered” observations, i.e. normalize data such that each gene has mean = 0

PCA projections (as XY-plot)



Related methods

- Factor Analysis*
- Multidimensional scaling (MDS)
- Generalized multidimensional scaling (GMDS)
- Semantic mapping
- Isomap
- Independent component analysis (ICA)

* Factor analysis is often confused with PCA though the two methods are related but distinct. Factor analysis is equivalent to PCA if the error terms in the factor analysis model are assumed to all have the same variance.

Why do we cluster?

- Organize observed data into meaningful structures
- Summarize large data sets
- Used when we have no *a priori* hypotheses
- Optimization:
 - Minimize within cluster distances
 - Maximize between cluster distances

Many types of clustering methods

- Method:
 - K-class
 - Hierarchical, e.g. UPGMA
 - Agglomerative (bottom-up) ... all alone ... join ...
 - Divisive (top-down) ... all together ... split ...
 - Graph theoretic
- Information used:
 - Supervised vs unsupervised
- Final description of the items:
 - Partitioning vs non-partitioning
 - fuzzy, multi-class

Hierarchical clustering

- Representation of all pair-wise distances
- Parameters: none (distance measure)
- Results:
 - One large cluster
 - Hierarchical tree (dendrogram)
- Deterministic

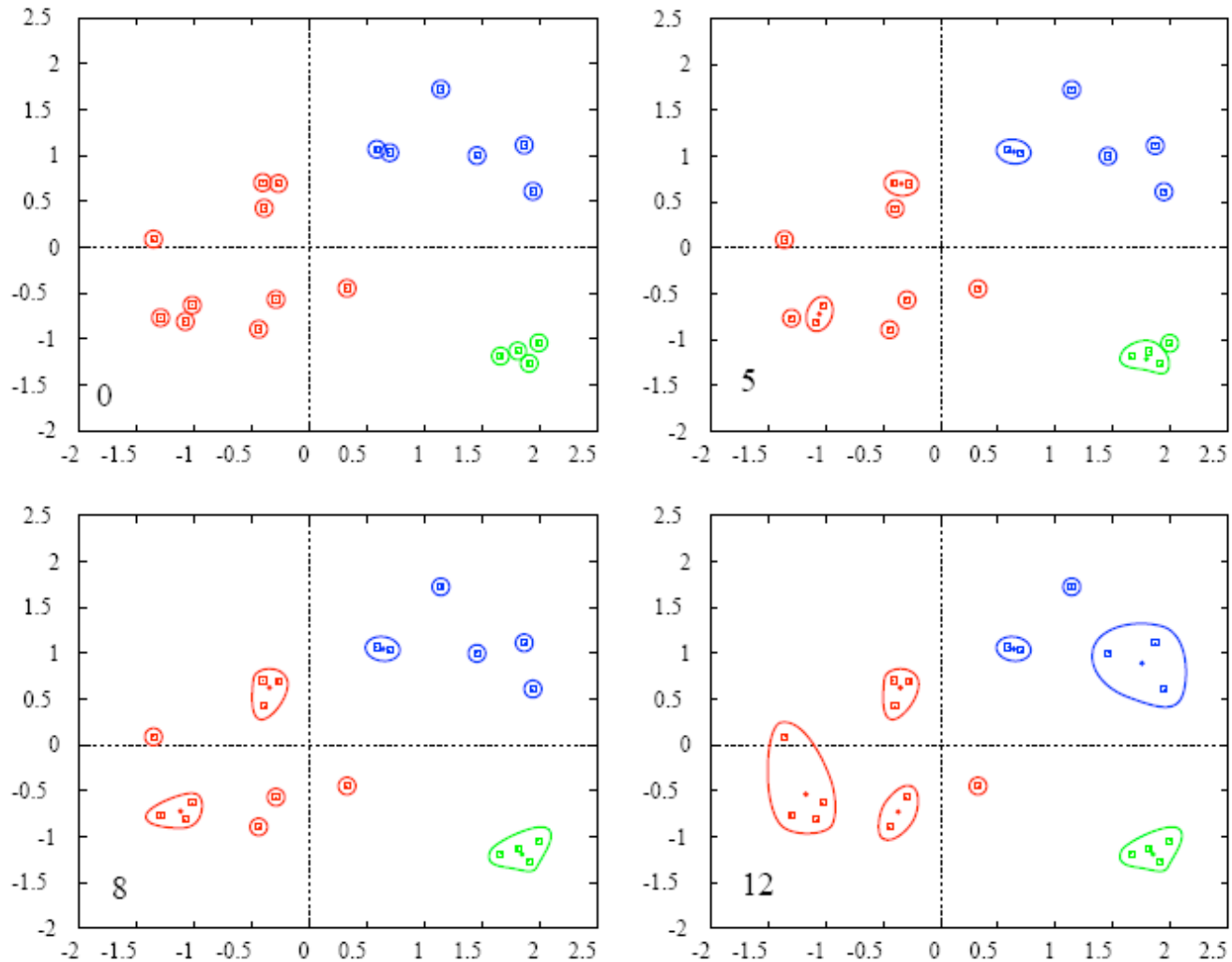
Hierarchical clustering – UPGMA Algorithm

Unweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean

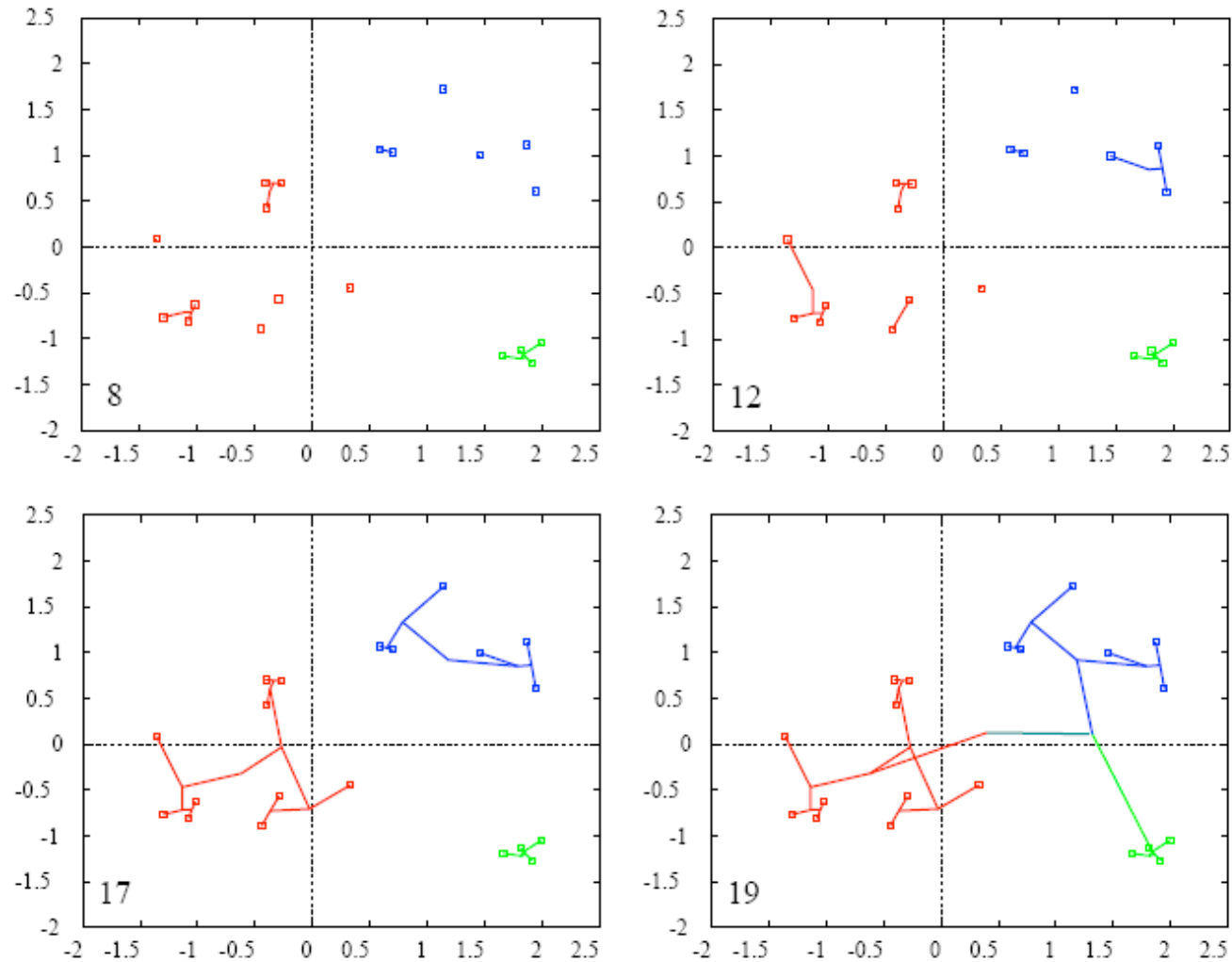


- Assign each item to its own cluster
- Join the nearest clusters
- Re-estimate the distance between clusters
- Repeat for 1 to n

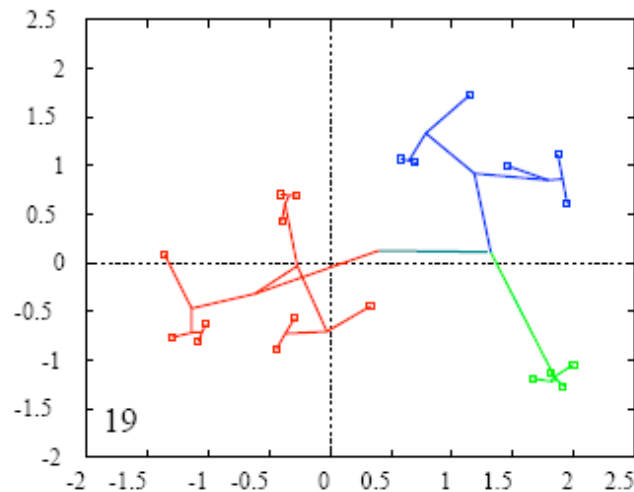
Hierarchical clustering



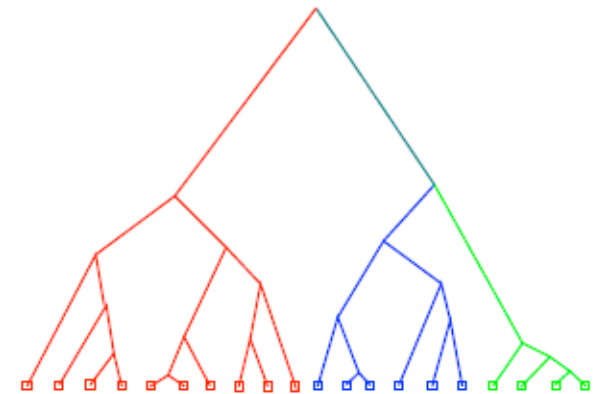
Hierarchical clustering



Hierarchical Clustering



Data with clustering order
and distances

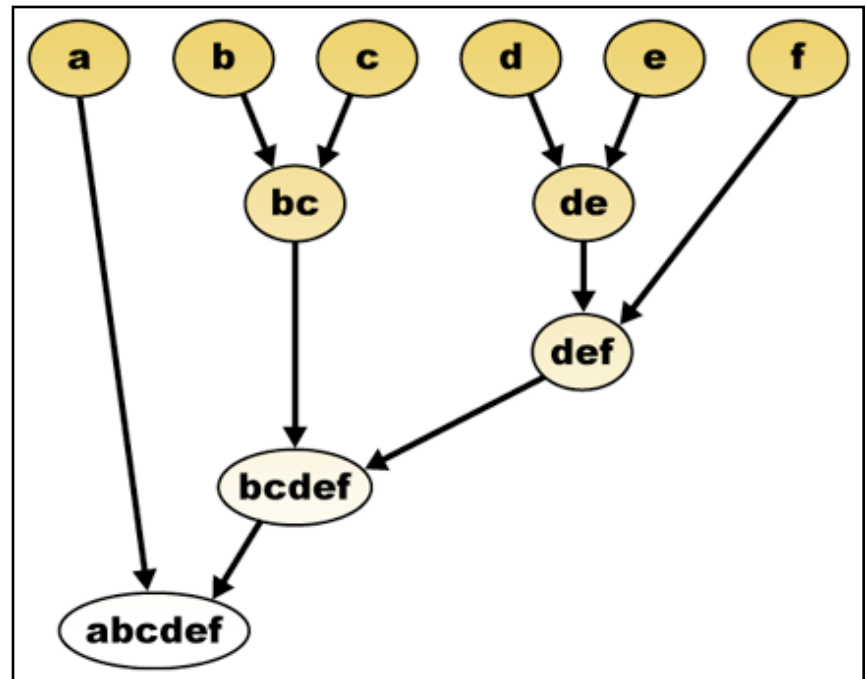
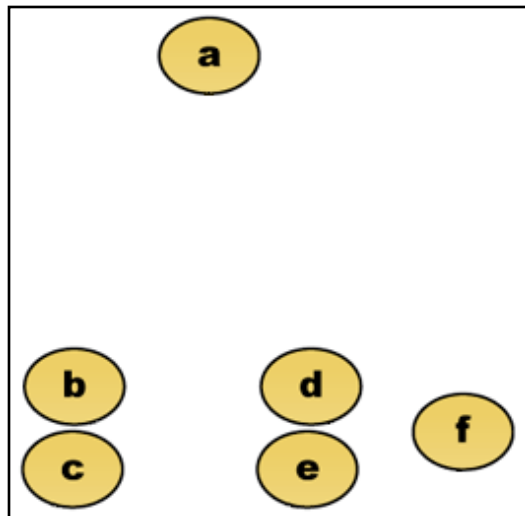


Dendrogram representation

2D data is a special (simple) case!

Hierarchical Clustering

Original data space



Merging steps define a dendrogram

K-means - Algorithm

Begin

Assign each item a class in 1 to K (randomly)

For 1 to max-iteration {

For each class 1 to K {

Calculate centroid (one of the " K means")

Calculate distance from centroid to each item

}

Assign each item the class of the nearest centroid

Exit if no items are re-assigned (convergence)

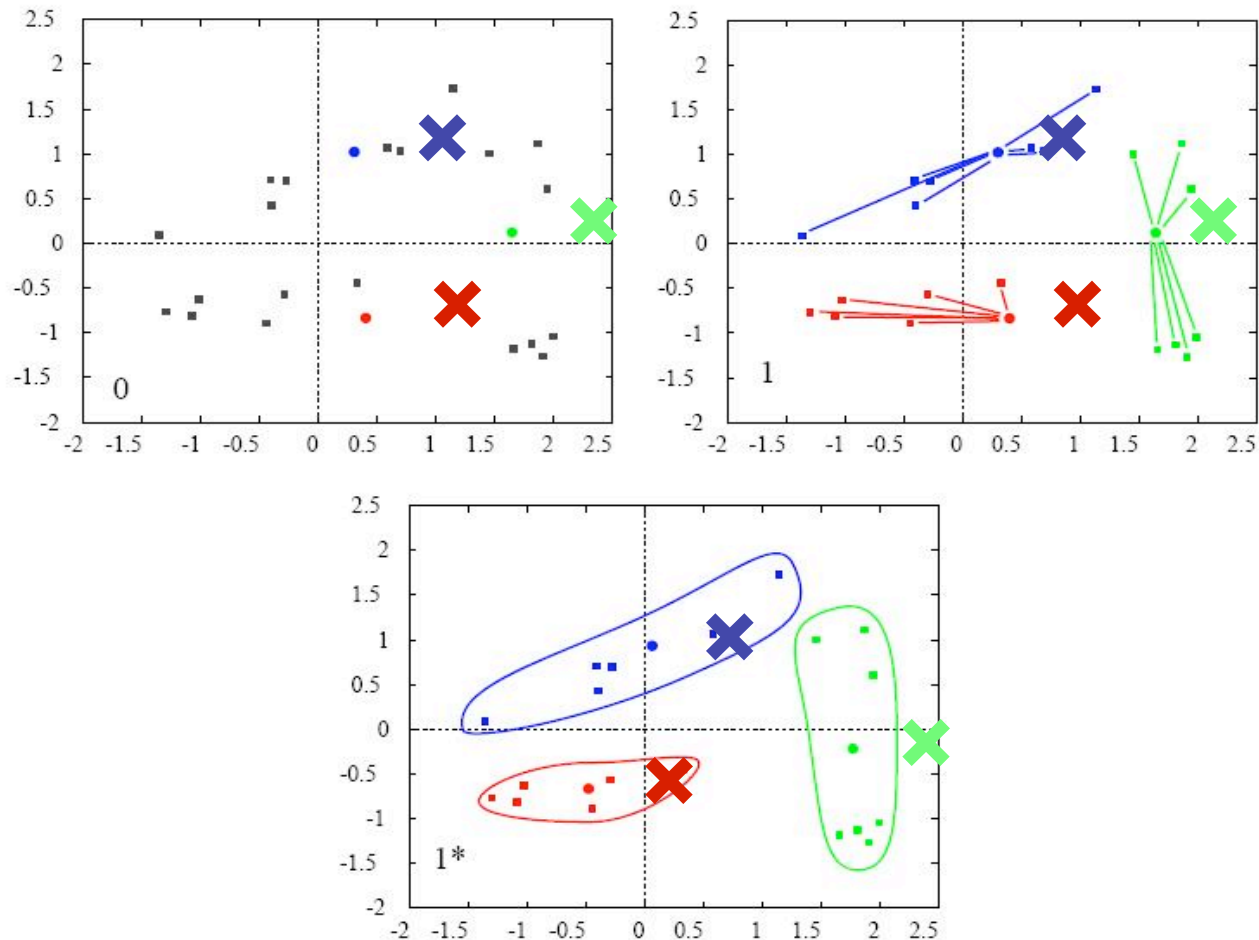
}

End

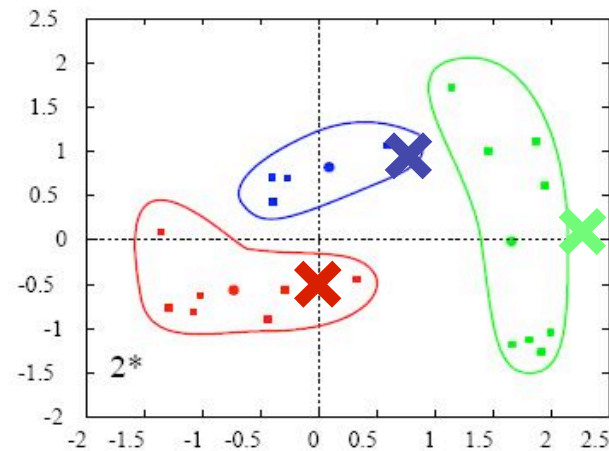
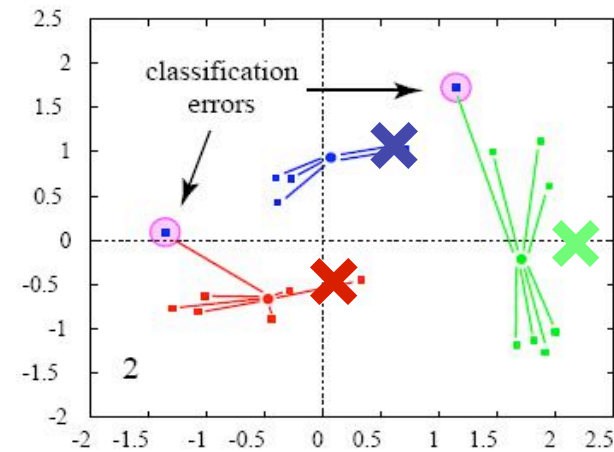
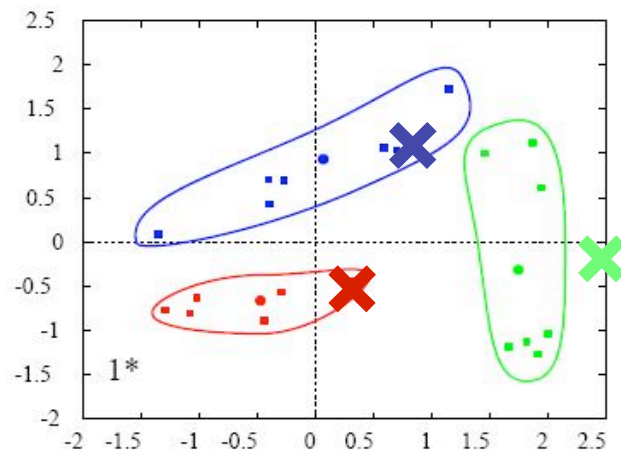
J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations",
Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability,
Berkeley, University of California Press, 1:281-297

K-mean clustering, K=3

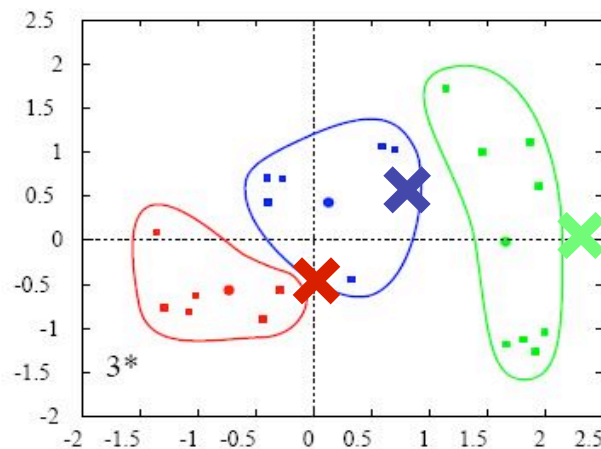
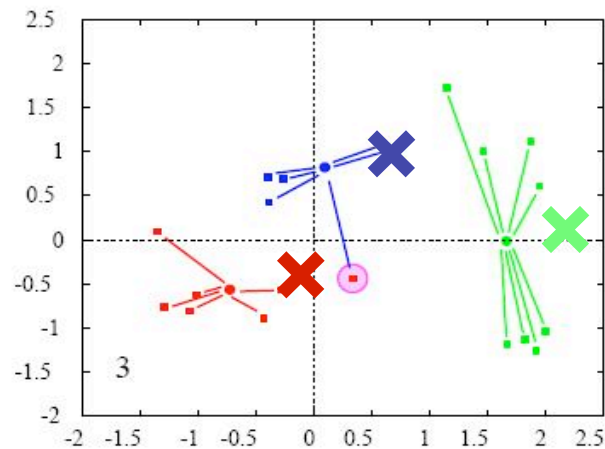
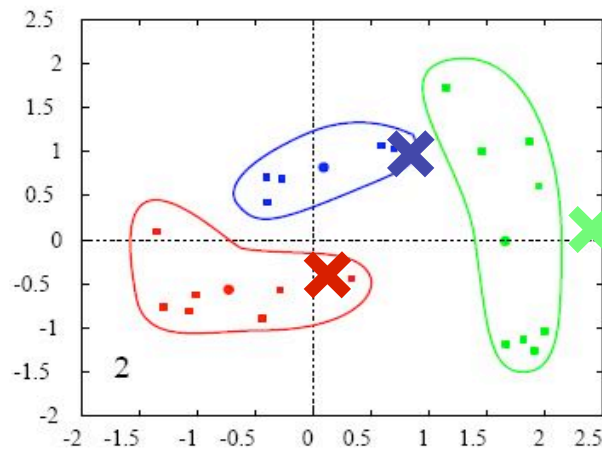
start



K-mean clustering, K=3

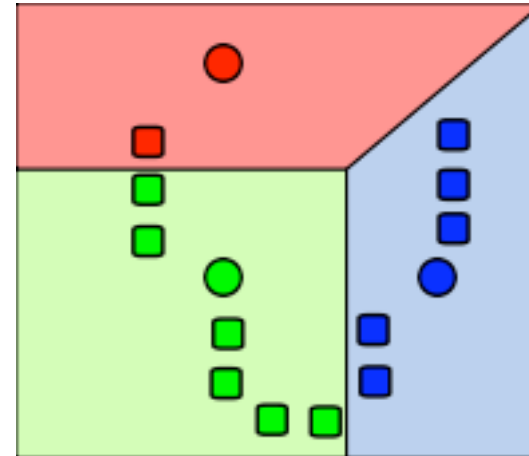
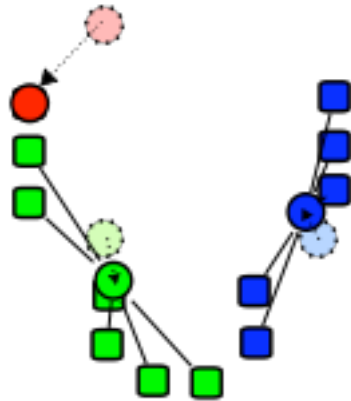
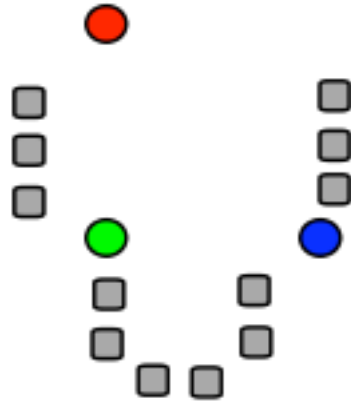


K-mean clustering, K=3

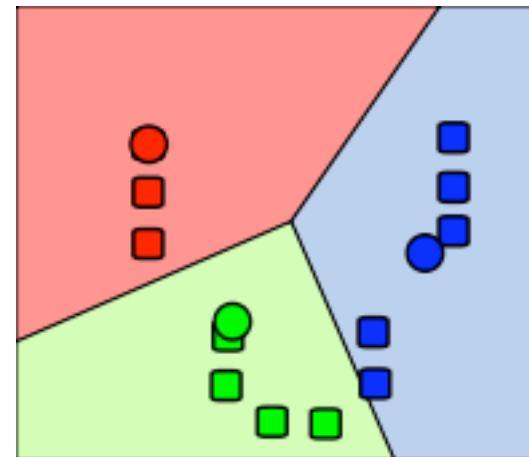


end

K-Means



Iteration i

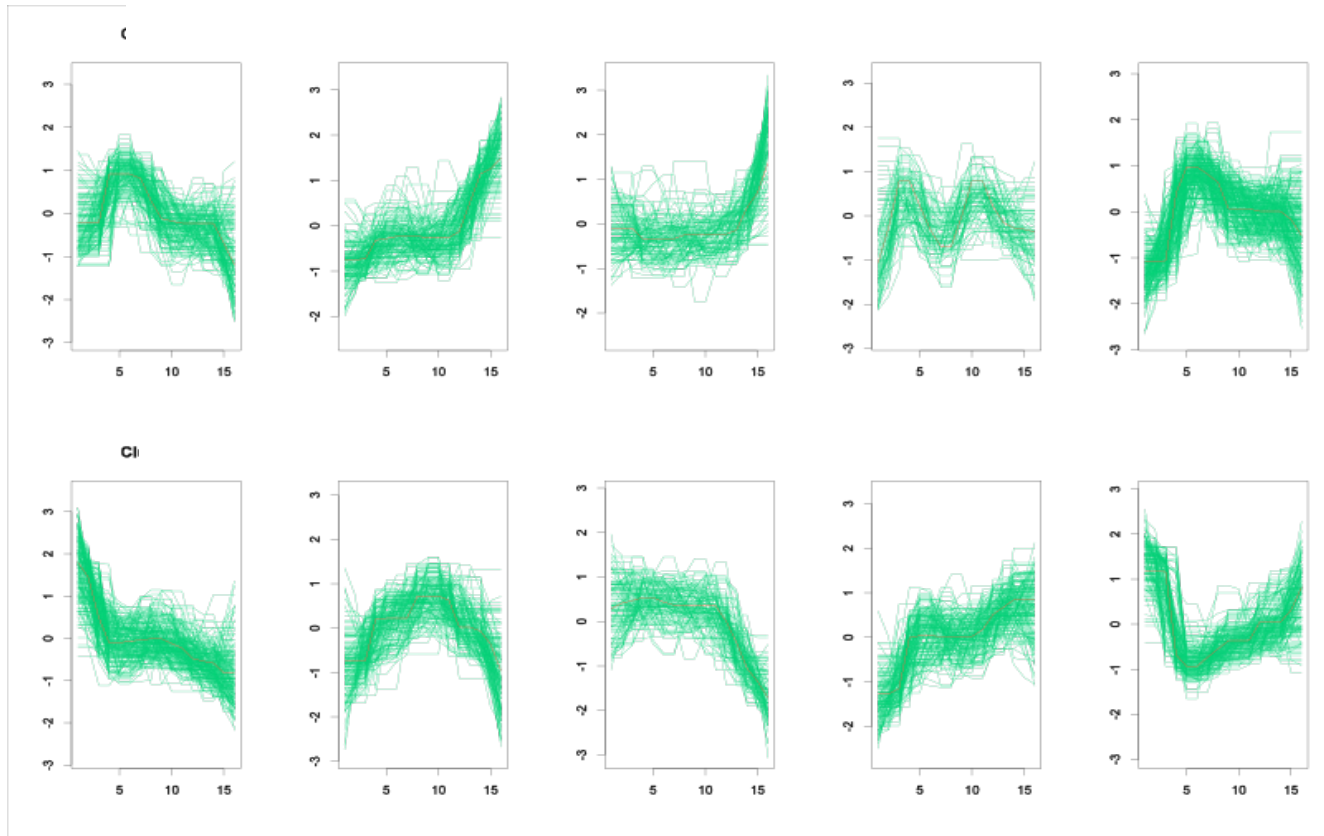


Iteration $i+1$

Circles: “prototypes” (parameters to fit)
Squares: data points

K-means clustering

Cell Cycle data

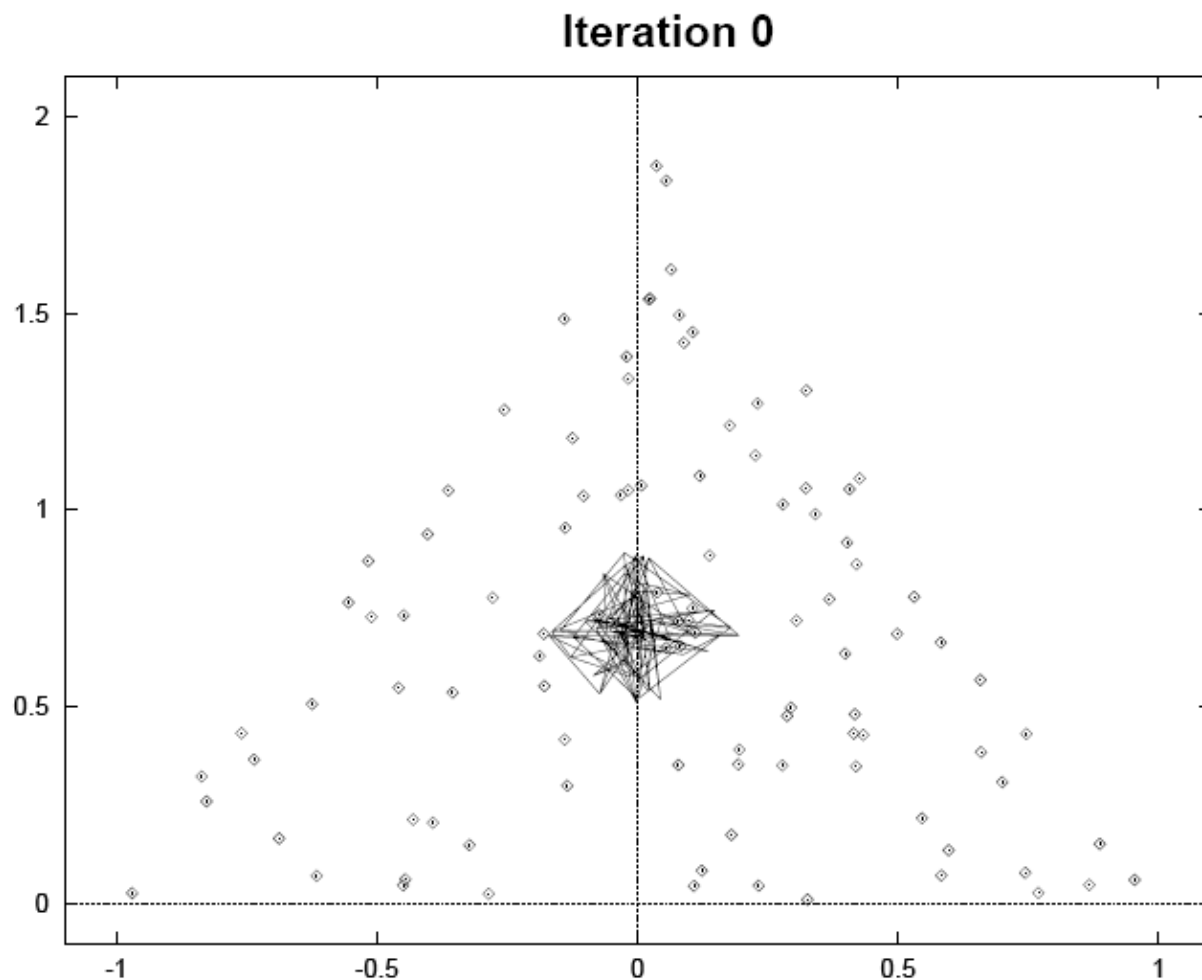


Self Organizing Maps (SOM)

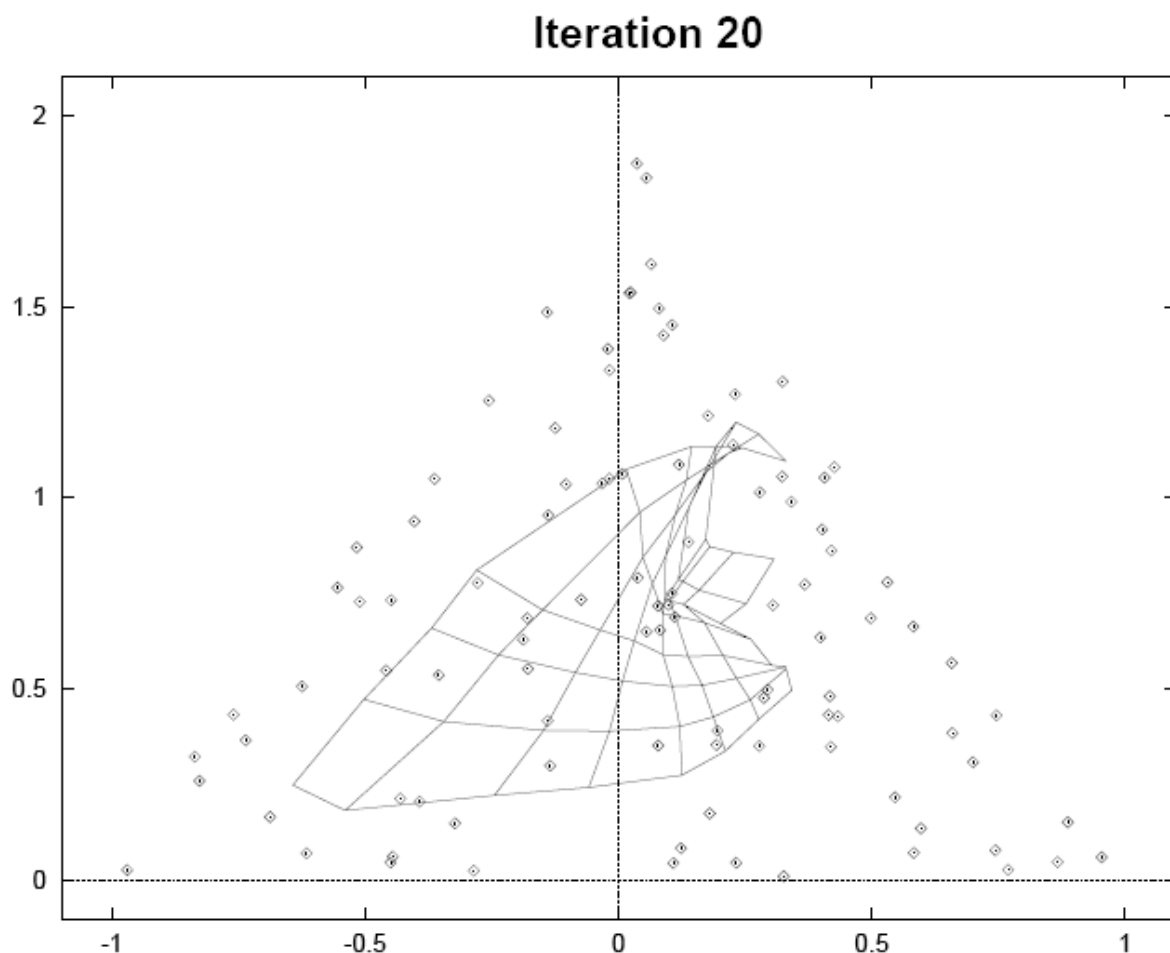


- Partitioning method
(similar to the K-means method)
- Clusters are organized in a two-dimensional grid
- Size of grid must be specified
 - (eg. 2x2 or 3x3)
- SOM algorithm finds the optimal organization of data in the grid

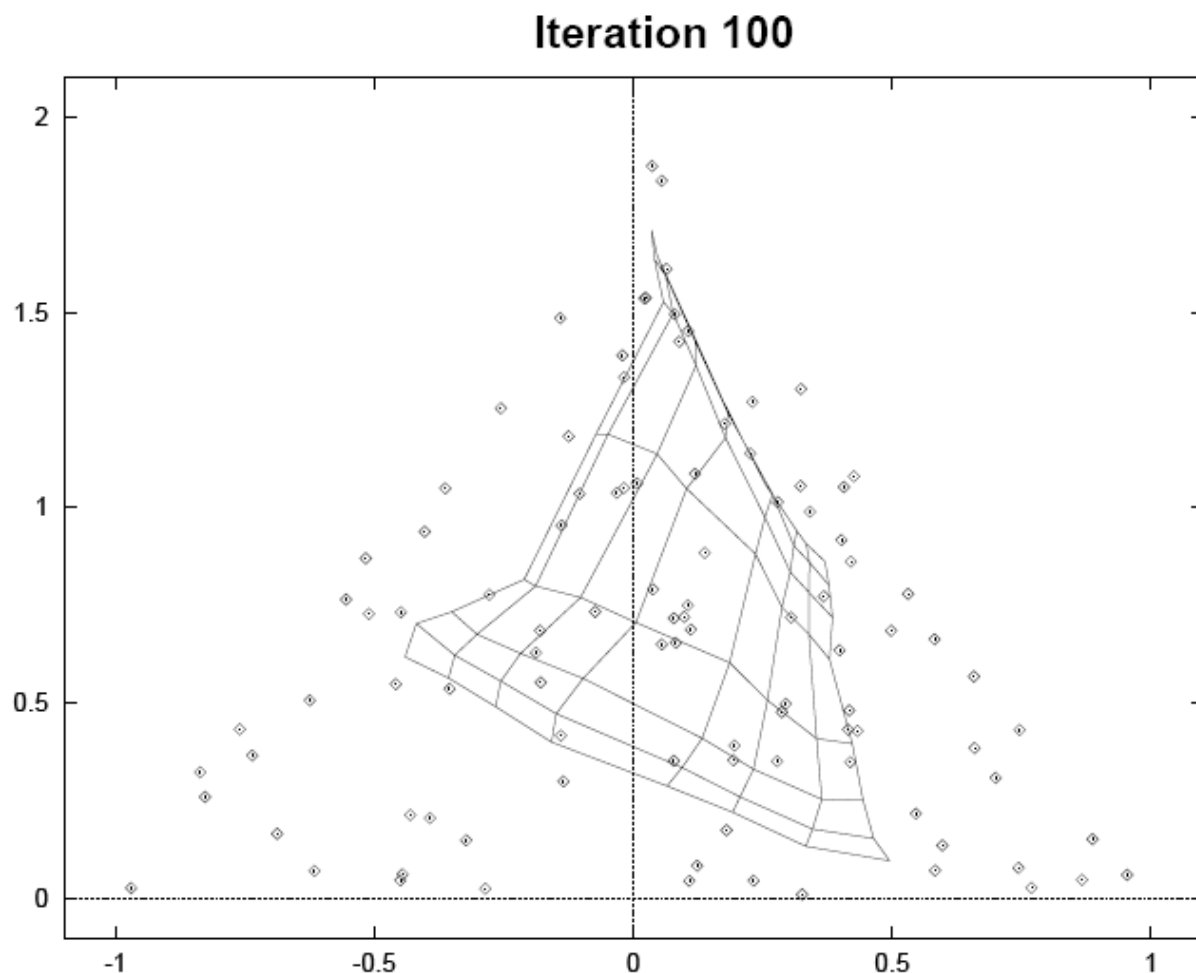
SOM - example



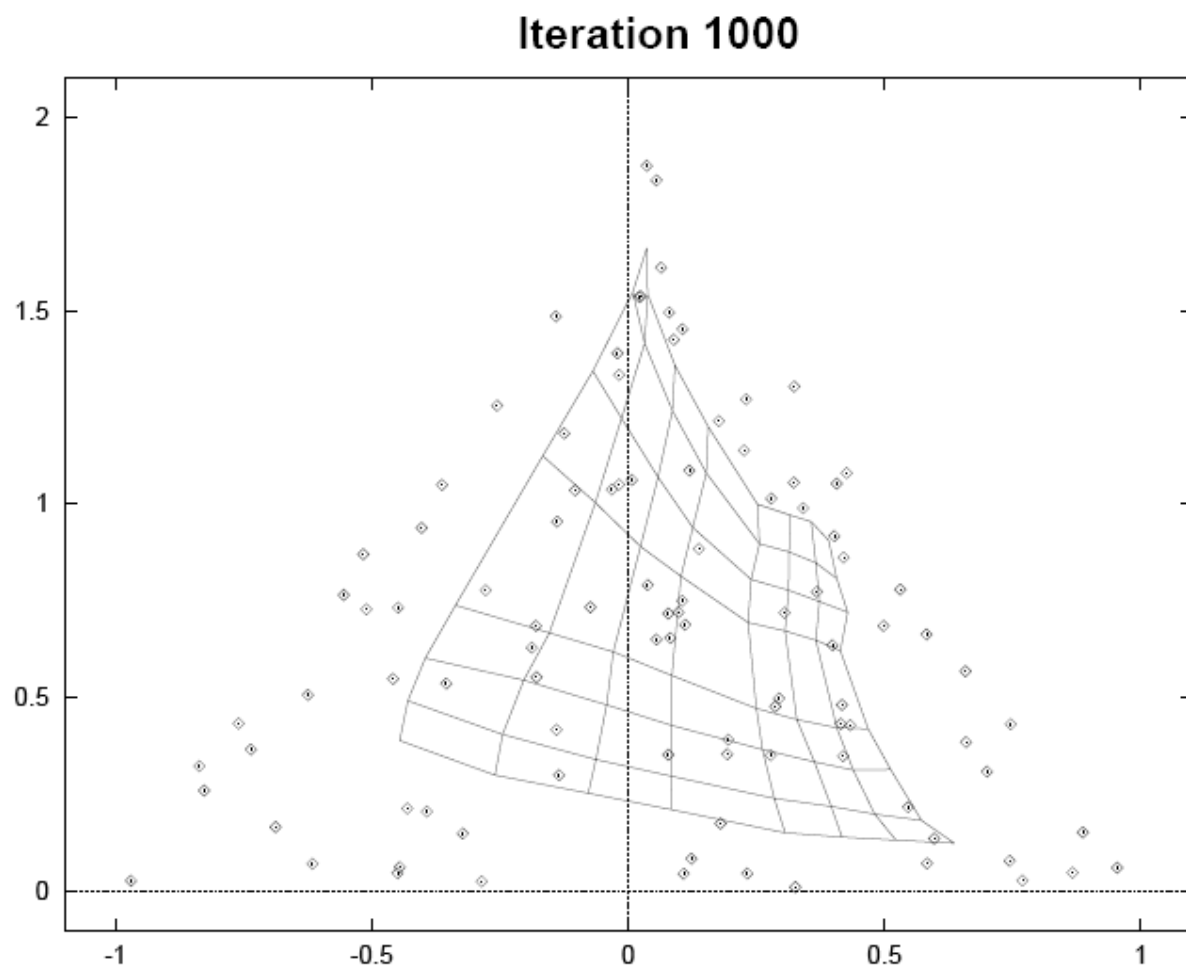
SOM - example



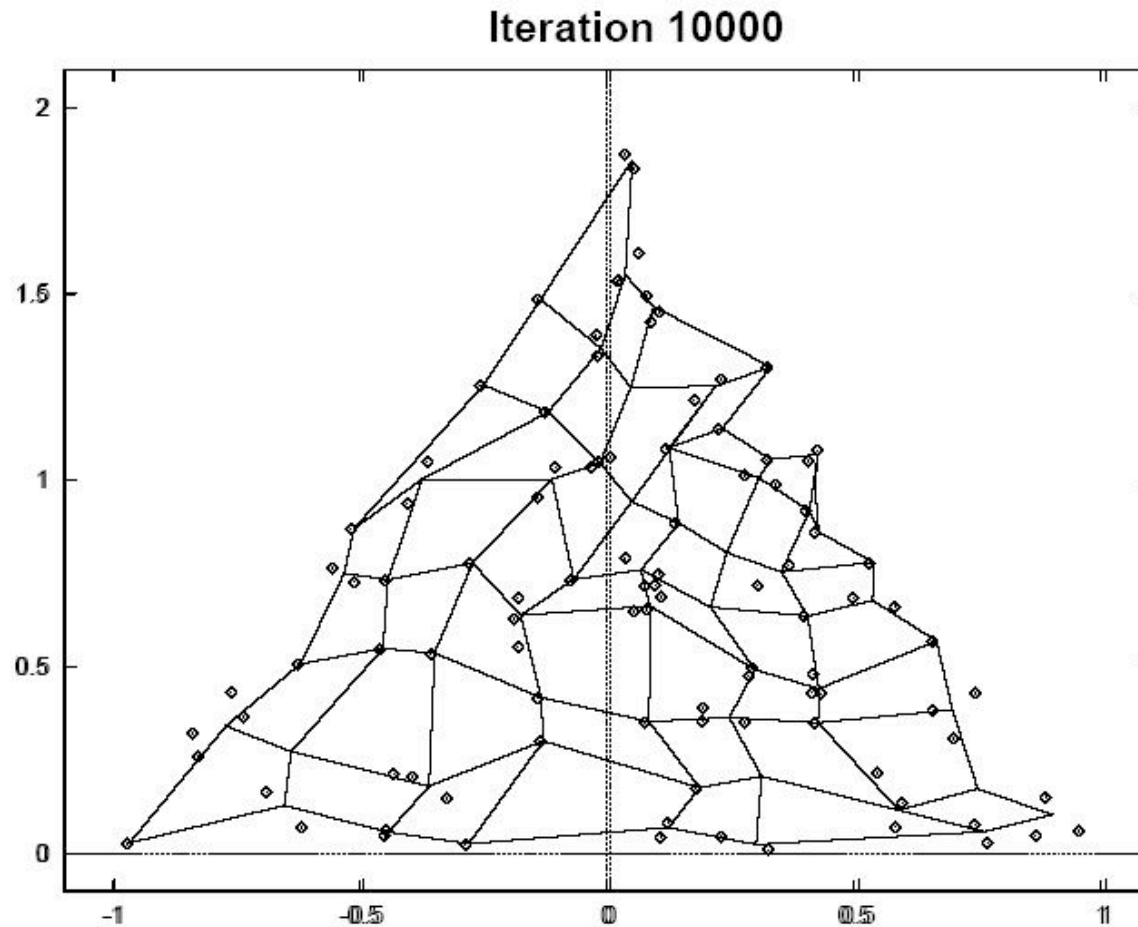
SOM - example



SOM - example



SOM - example



Comparison of clustering methods

- Hierarchical clustering
 - Distances between all variables
 - Time consuming with a large number of gene
 - Advantage to cluster on selected genes
- K-means clustering
 - Faster algorithm
 - Does only show relations between all variables
- SOM
 - Machine learning algorithm

Distance measures

- Euclidian distance

$$d(x_i, y_i) = \left(\sum_{i=1}^N (x_i - y_i)^2 \right)^{1/2}$$

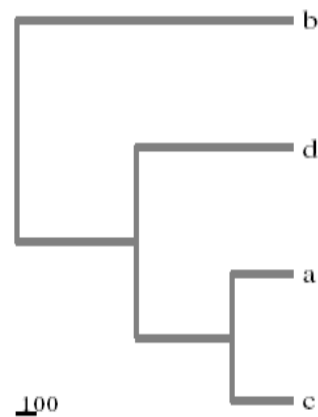
- Vector angle distance

$$d(x_i, y_i) = (1 - \cos \alpha) = 1 - \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

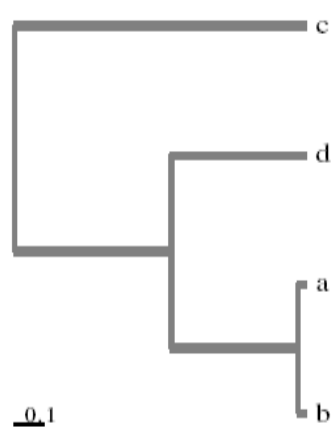
- Pearsons distance

$$d(x_i, y_i) = (1 - CC) = 1 - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

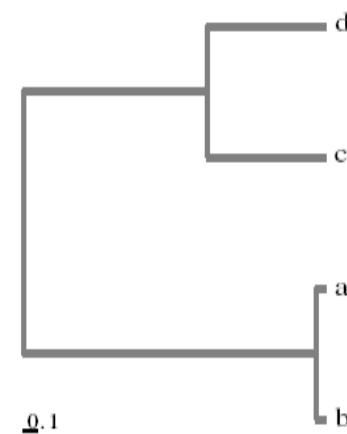
Comparison of distance measures



Euclidean



Vector angle



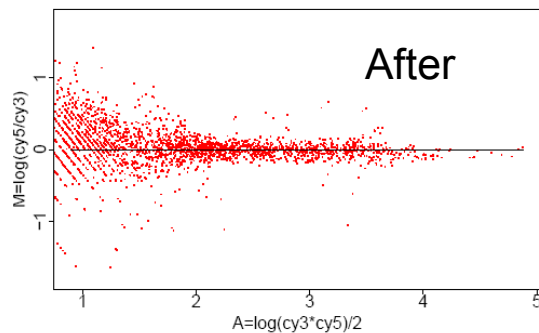
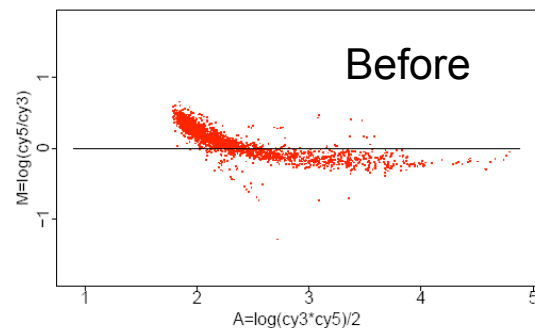
Pearson

Summary

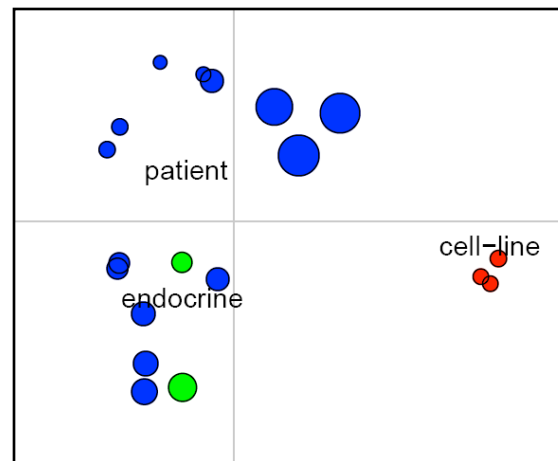
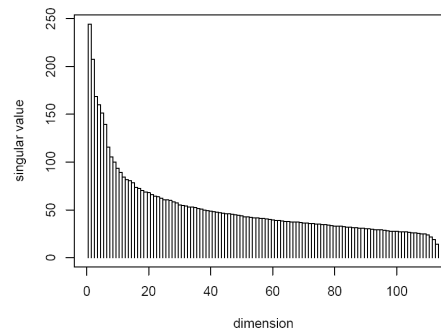
- Dimension reduction important to visualize data
 - Methods:
 - Principal Component Analysis
 - Clustering
 - Hierarchical
 - K-means
 - Self organizing maps
- (distance measure important)

DNA Microarray Analysis Overview/Review

Normalization



PCA (using SVD)



Cluster analysis

