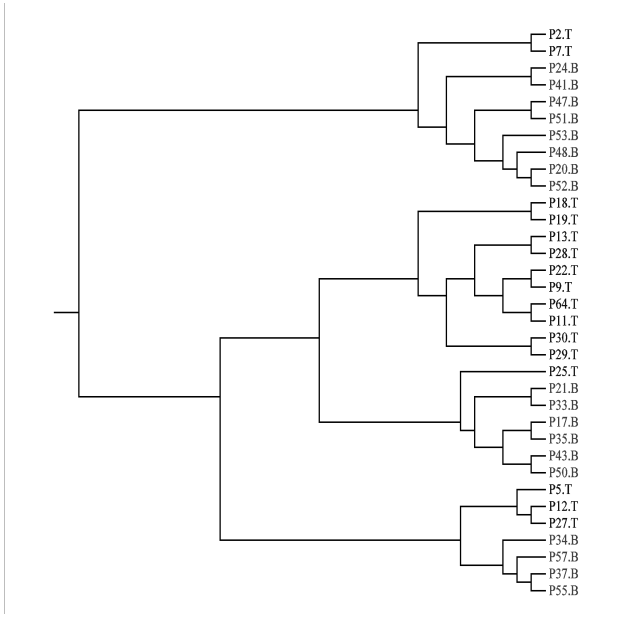
PCA, Clustering and Classification

By H. Bjørn Nielsen strongly inspired by Agnieszka S. Juncker

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



Motivation: Multidimensional data

	Pat1	Pat2	Pat3	Pat4	Pat5	Pat6	Pat7	Pat8
209619_at	7758	4705	5342	7443	8747	4933	7950	5031
32541_at	280	387	392	238	385	329	337	163
206398_s_at	1050	835	1268	1723	1377	804	1846	1180
219281_at	391	593	298	265	491	517	334	387
207857_at	1425	977	2027	1184	939	814	658	593
211338_at	37	27	28	38	33	16	36	23
213539_at	124	197	454	116	162	113	97	97
221497_x_at	120	86	175	99	115	80	83	119
213958_at	179	225	449	174	185	203	186	185
210835_s_at	203	144	197	314	250	353	173	285
209199_s_at	758	1234	833	1449	769	1110	987	638
217979_at	570	563	972	796	869	494	673	1013
201015_s_at	533	343	325	270	691	460	563	321
203332_s_at	649	354	494	554	710	455	748	392
204670_x_at	5577	3216	5323	4423	5771	3374	4328	3515
208788_at	648	327	1057	746	541	270	361	774
210784_x_at	142	151	144	173	148	145	131	146
204319_s_at	298	172	200	298	196	104	144	110
205049_s_at	3294	1351	2080	2066	3726	1396	2244	2142
202114_at	833	674	733	1298	862	371	886	501
213792_s_at	646	375	370	436	738	497	546	406
203932_at	1977	1016	2436	1856	1917	822	1189	1092
203963_at	97	63	77	136	85	74	91	61
203978_at	315	279	221	260	227	222	232	141
203753_at	1468	1105	381	1154	980	1419	1253	554
204891_s_at	78	71	152	74	127	57	66	153
209365_s_at	472	519	365	349	756	528	637	828
209604_s_at	772	74	130	216	108	311	80	235
211005_at	49	58	129	70	56	77	61	61
219686_at	694	342	345	502	960	403	535	513
38521_at	775	604	305	563	542	543	725	587
217853_at	367	168	107	160	287	264	273	113
217028_at	4926	2667	3542	5163	4683	3281	4822	3978
201137_s_at	4733	2846	1834	5471	5079	2330	3345	1460
202284_s_at	600	1823	1657	1177	972	2303	1574	1731
201999_s_at	897	959	800	808	297	1014	998	663

Outline

- Dimension reduction
  - PCA
  - Clustering
- Classification
- Example: study of childhood leukemia

Childhood Leukemia

- Cancer in the cells of the immune system
- Approx. 35 new cases in Denmark every year
- 50 years ago – all patients died
- Today – approx. 78% are cured
- Riskgroups
  - Standard
  - Intermediate
  - High
  - Very high
  - Extra high
- Treatment
  - Chemotherapy
  - Bone marrow transplantation
  - Radiation

Prognostic Factors

	Good prognosis	**Poor prognosis**
Immunophenotype	precursor B	T
Age	1-9	≥10
Leukocyte count	Low (<50*10^9/L)	High (>100*10^9/L)
Number of chromosomes	Hyperdiploidy (>50)	Hypodiploidy (<46)
Translocations	t(12;21)	t(9;22), t(1;19)
Treatment response	Good response	Poor response

Study of Childhood Leukemia

- Diagnostic bone marrow samples from leukemia patients
- Platform: Affymetrix Focus Array
  - 8763 human genes
- Immunophenotype
  - 18 patients with precursor B immunophenotype
  - 17 patients with T immunophenotype
- Outcome 5 years from diagnosis
  - 11 patients with relapse
  - 18 patients in complete remission

Principal Component Analysis (PCA)

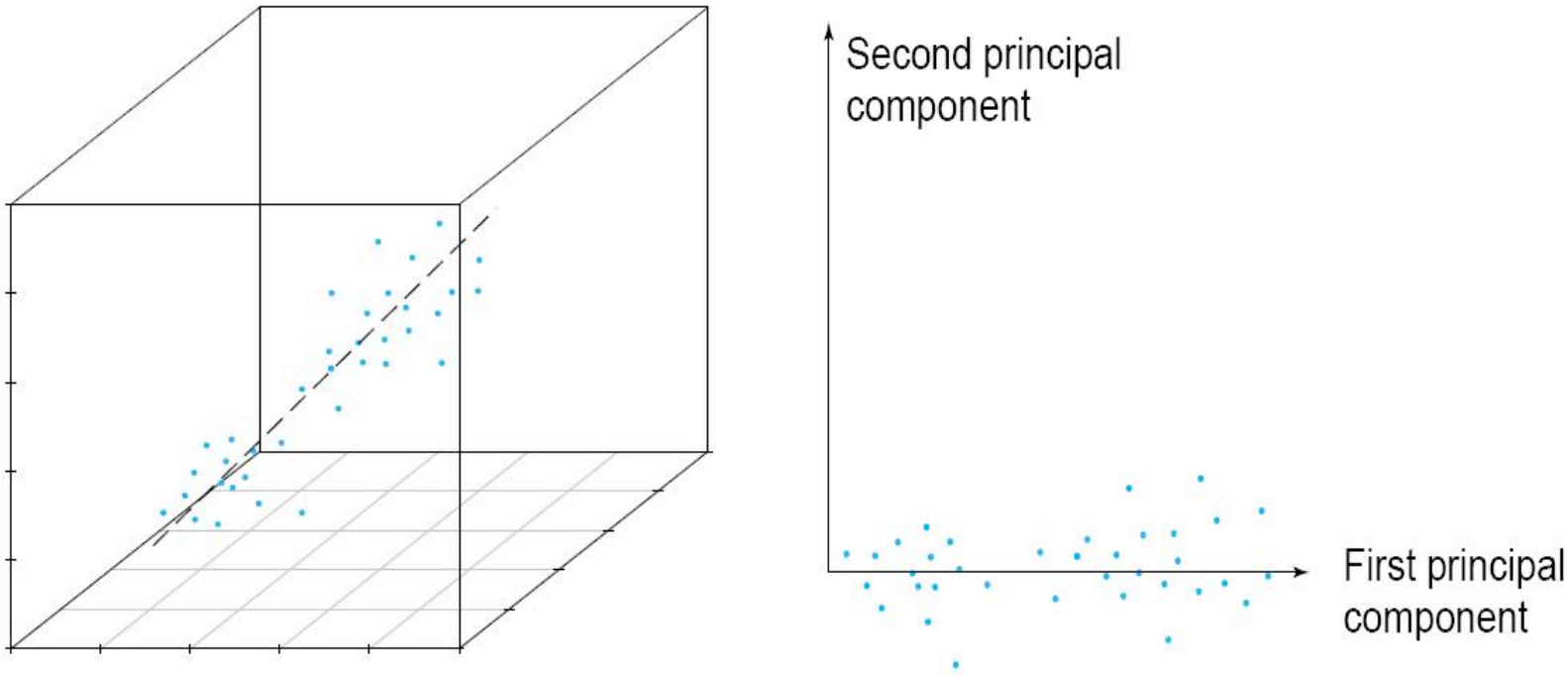
- used for visualization of complex data
- developed to capture as much of the variation in data as possible

Principal components

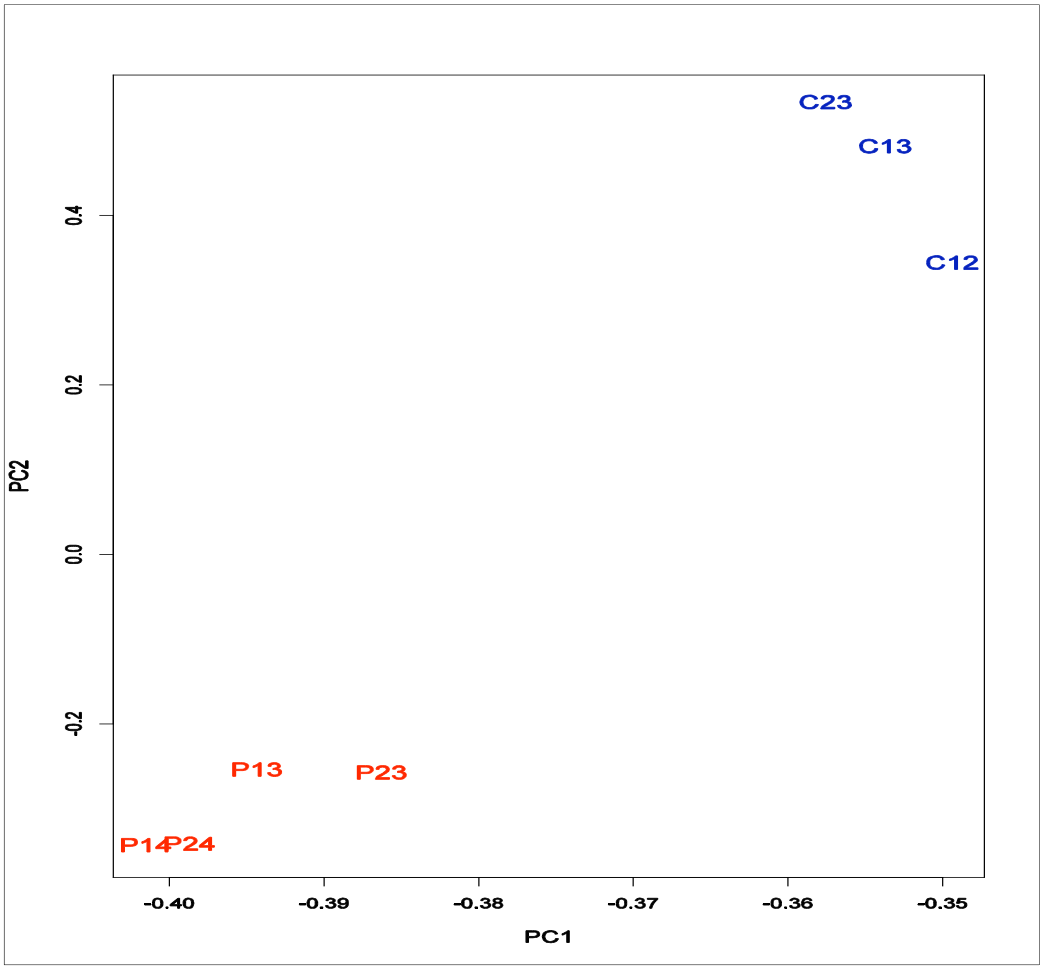
- 1. principal component (PC1)
  - the direction along which there is greatest variation
- 2. principal component (PC2)
  - the direction with maximum variation left in data, orthogonal to the 1. PC

- General about principal components
  - linear combinations of the original variables
  - uncorrelated with each other

Principal components



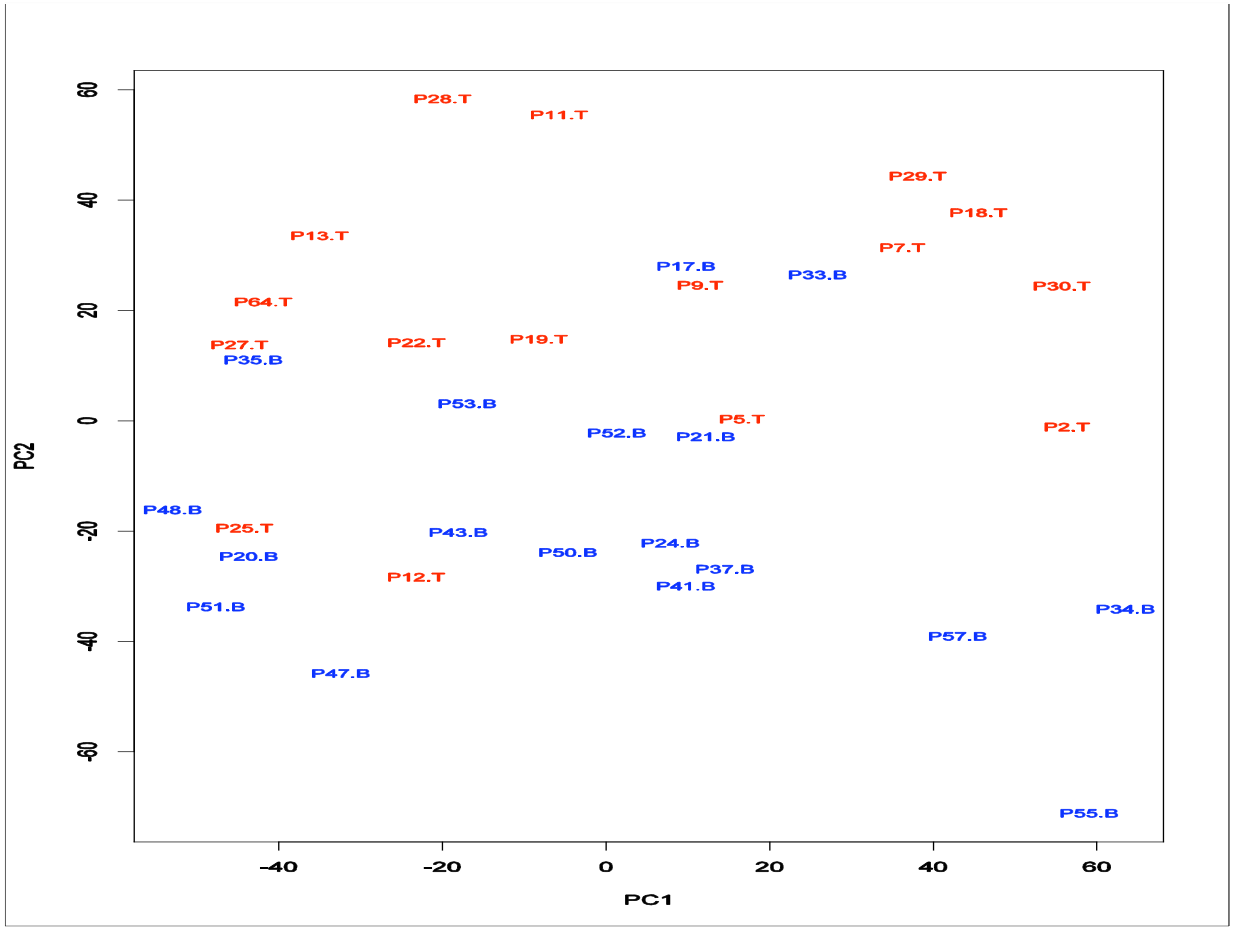
PCA - example


C23
C13
C12
P13
P23
P14
P24
0.4
0.2
0.0
-0.2
PC2
-0.40
-0.39
-0.38
-0.37
-0.36
-0.35
PC1

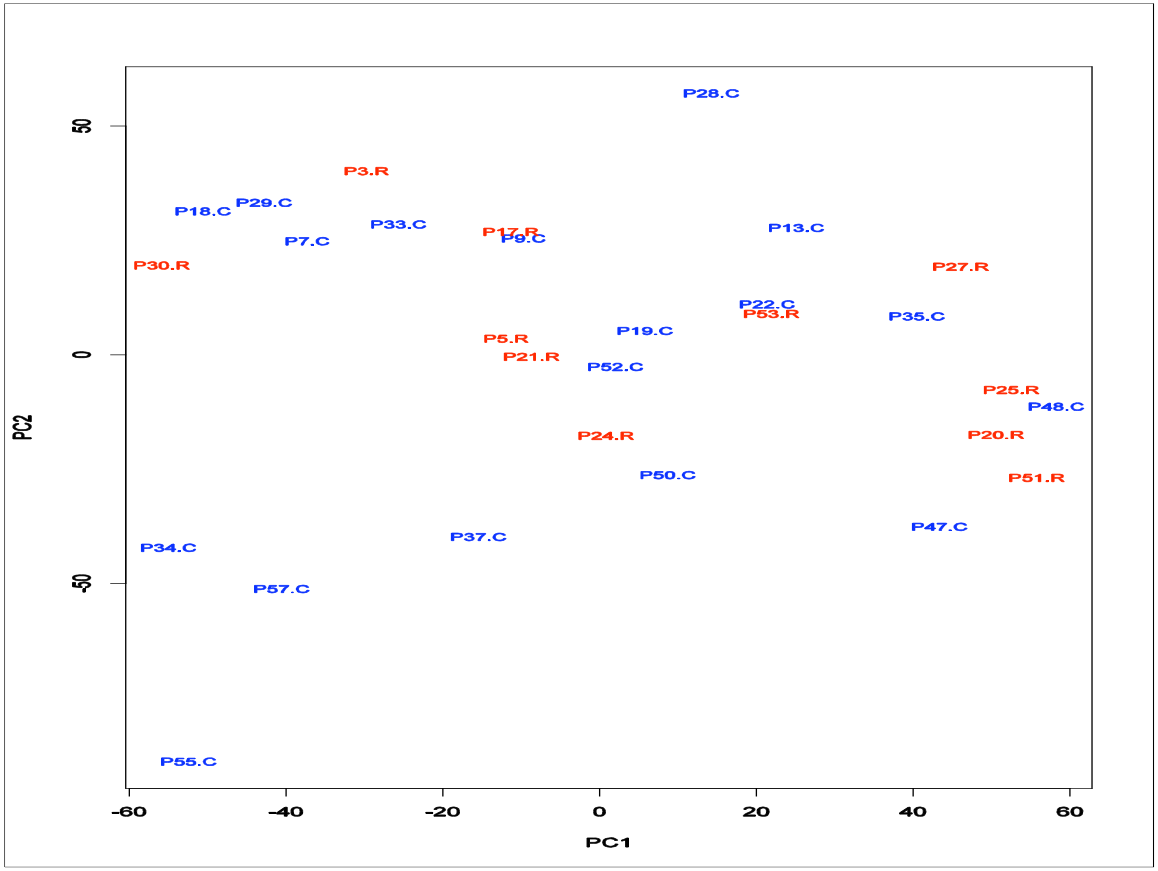
PCA on all Genes

Leukemia data, precursor B and T

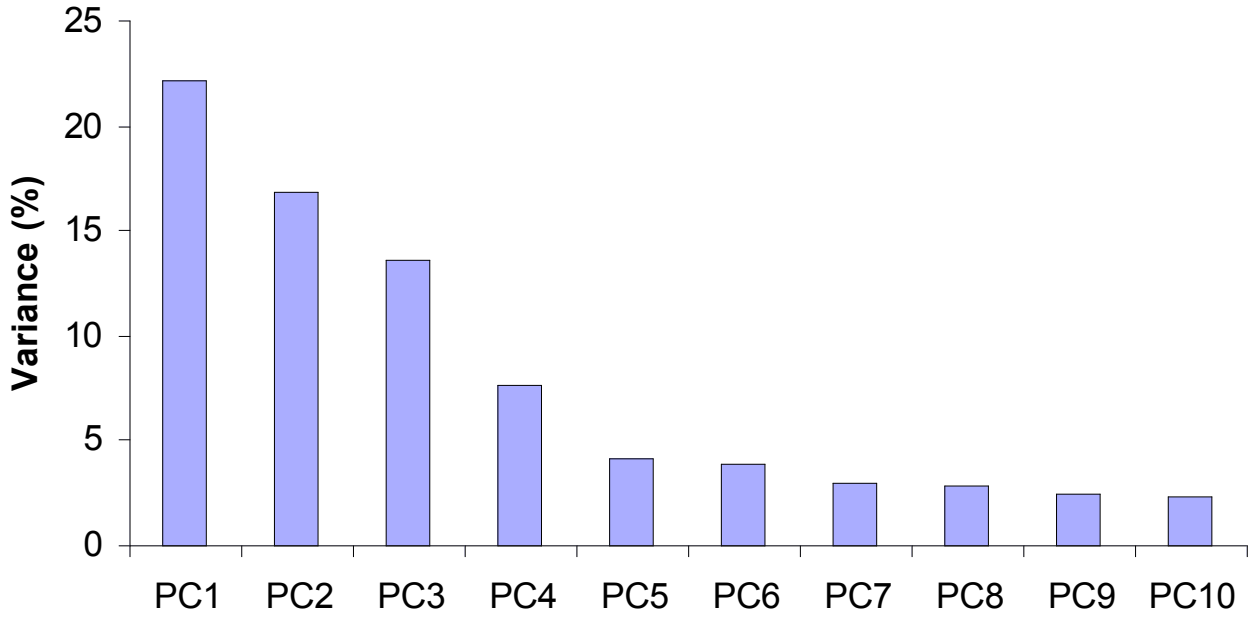
Plot of 34 patients, dimension of 8973 genes reduced to 2



Outcome: PCA on all Genes

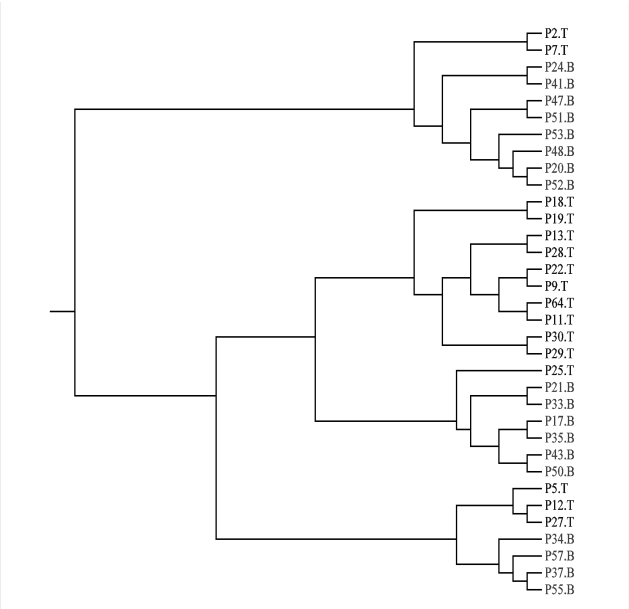


Principal components - Variance



Clustering methods

- Hierarchical
  - agglomerative (buttom-up) eg. UPGMA
  - divisive (top-down)



- Partitioning
  - eg. K-means clustering

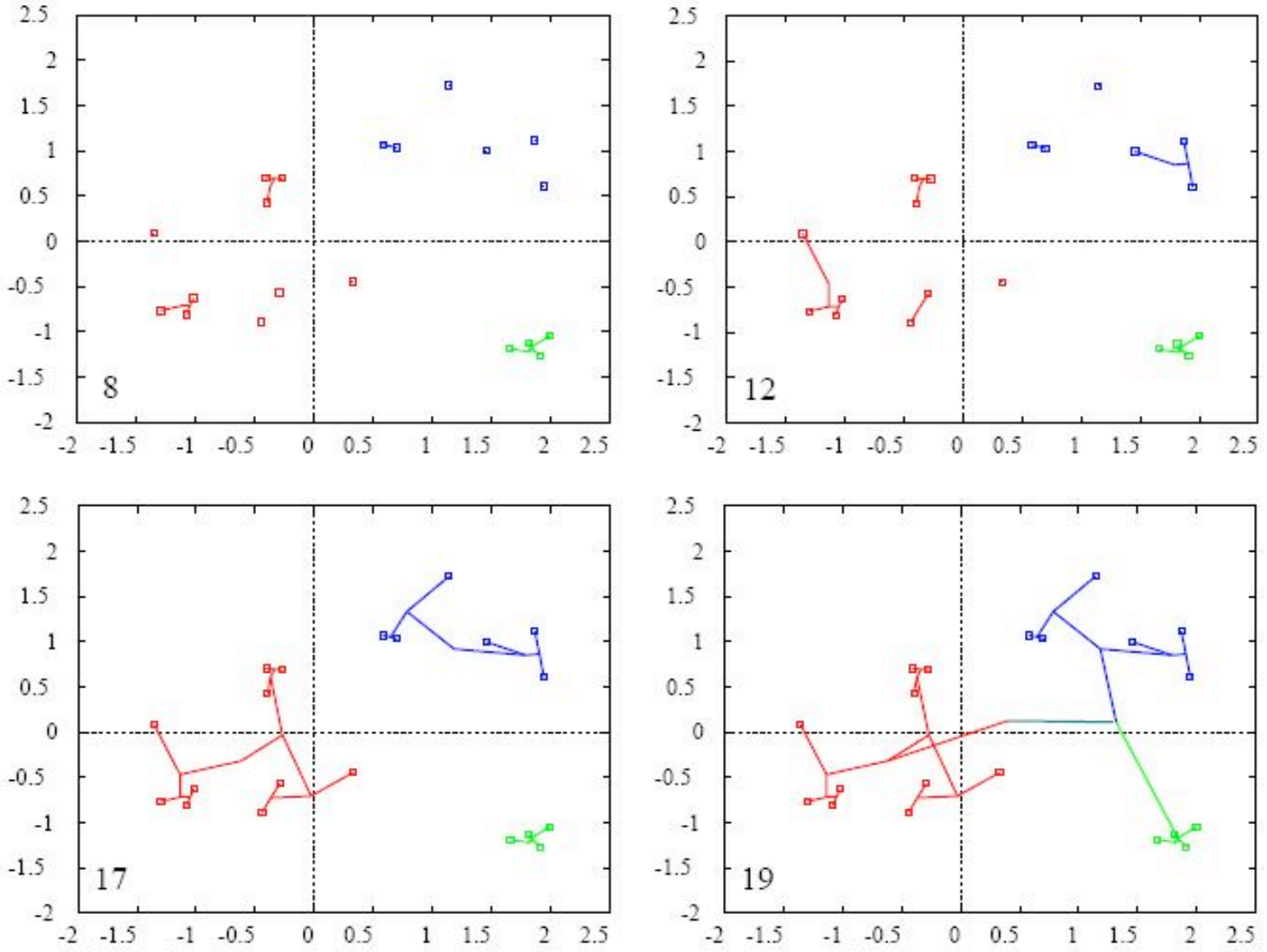
Hierarchical clustering

- Representation of all pairwise distances
- Parameters: none (distance measure)
- Results:
  - in one large cluster
  - hierarchical tree (dendrogram)
- Deterministic

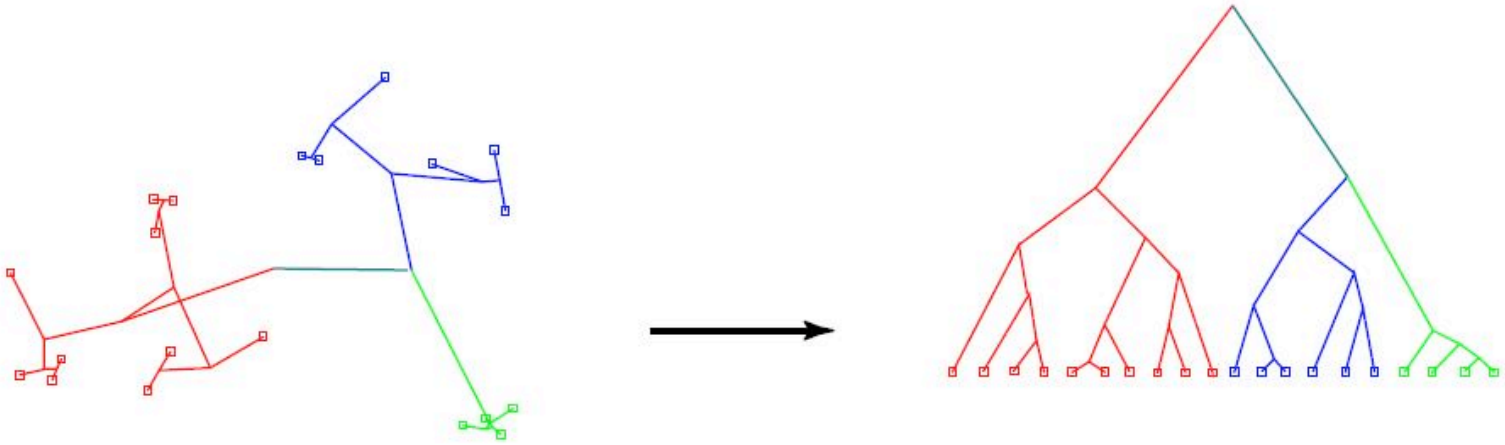
Hierarchical clustering – UPGMA Algorithm

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

Hierarchical clustering



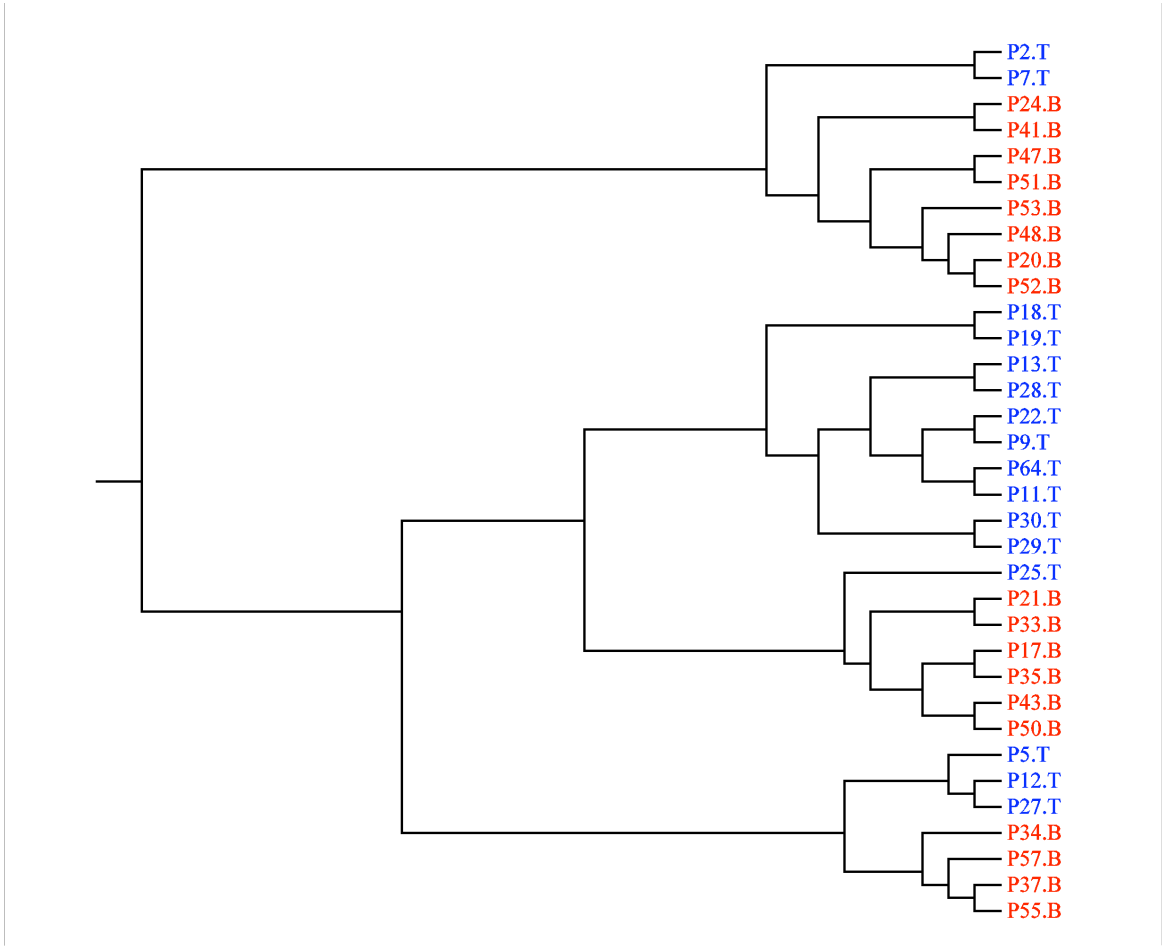
Hierarchical clustering



Data with clustering order and distances

Dendrogram representation

Leukemia data - clustering of patients



K-means clustering

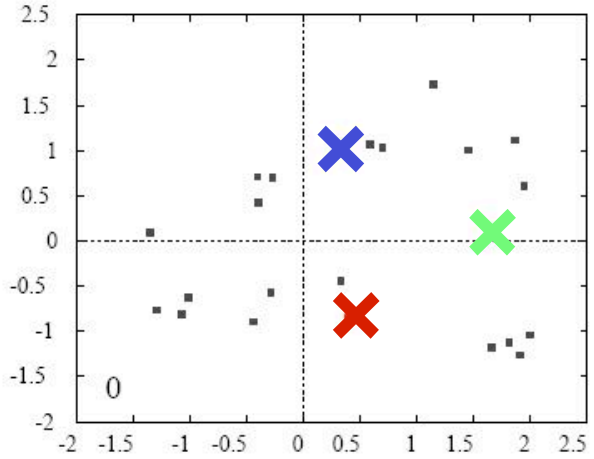
- Partition data into K clusters
- Parameter: Number of clusters (K) must be chosen
- Randomilized initialization:
  - different clusters each time

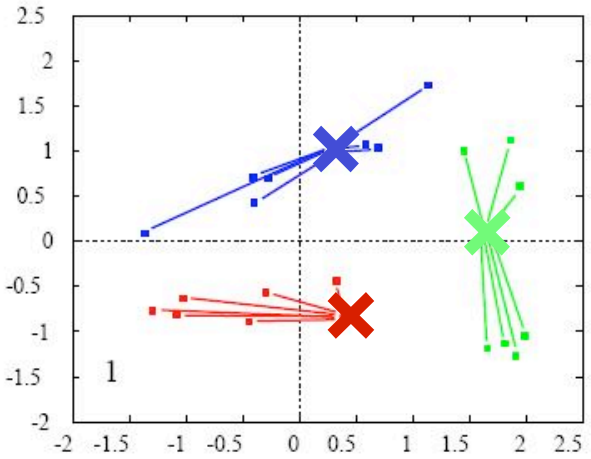
K-means - Algorithm

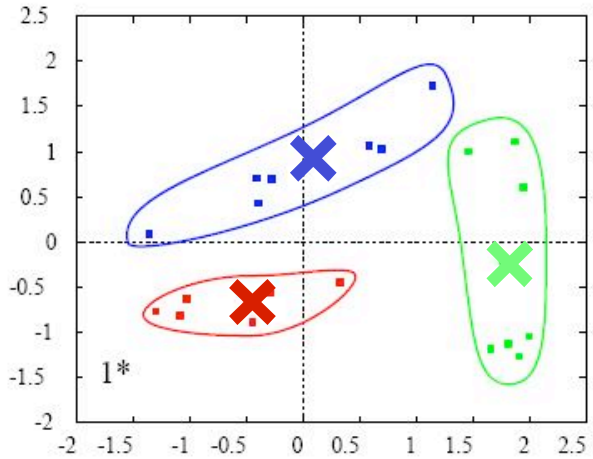
- Assign each item a class in 1 to K (randomly)
- For each class 1 to K
  - Calculate the centroid (one of the K-means)
  - Calculate distance from centroid to each item
- Assign each item to the nearest centroid
- Repeat until no items are re-assigned (convergence)

K-means clustering, K=3

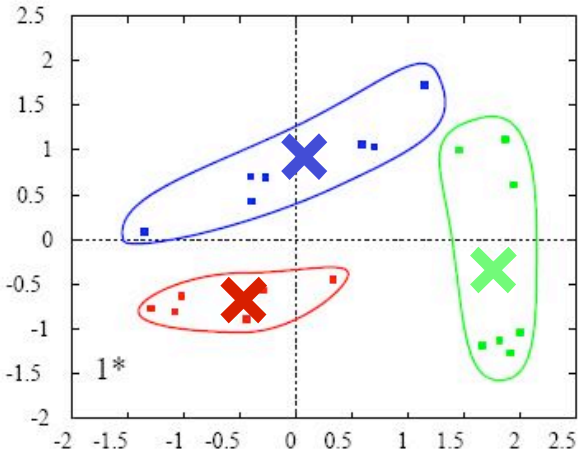
start

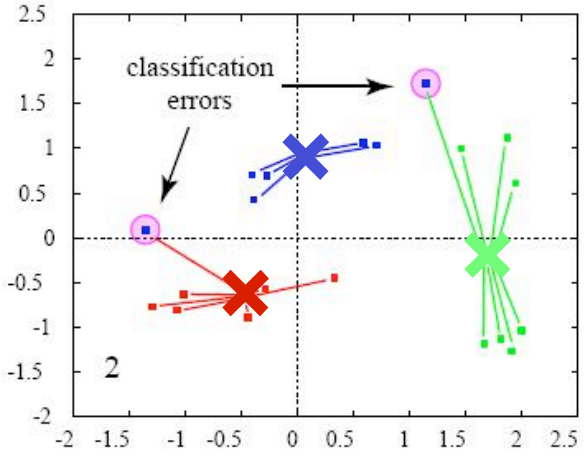


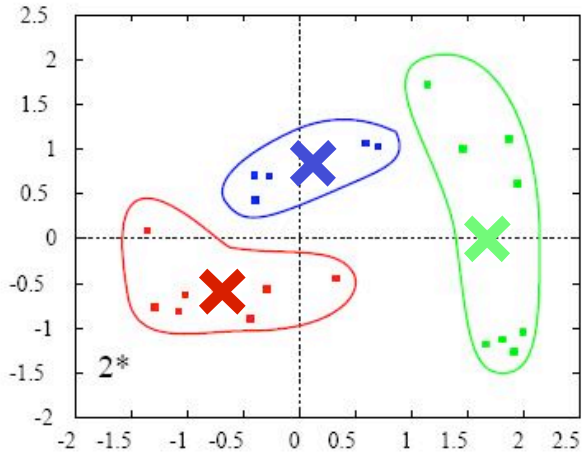




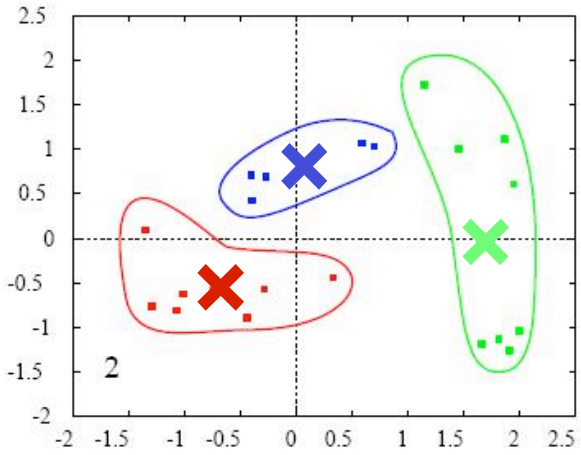
K-means clustering, K=3

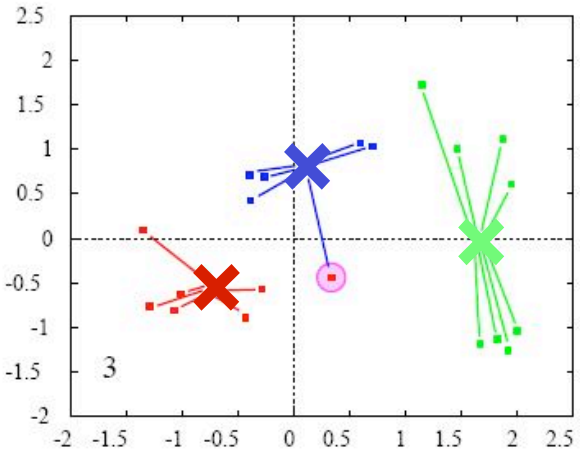


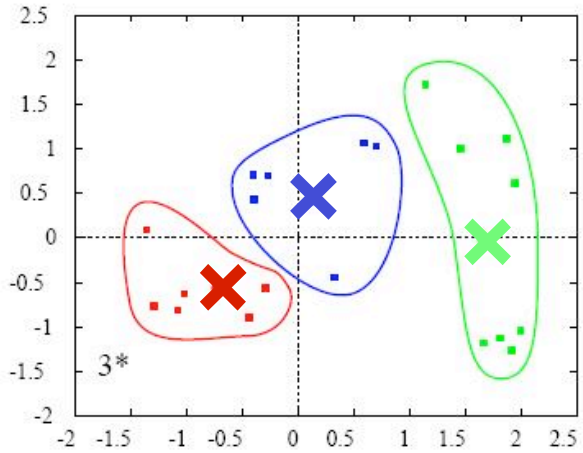




K-means clustering, K=3







end

Comparison of clustering methods

- Hierarchical clustering
  - Distances between all variables
  - Timeconsuming with a large number of gene
  - Advantage to cluster on selected genes

- K-mean clustering
  - Faster algorithm
  - Does not show relations between all variables

Distance measures

- Euclidian distance

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

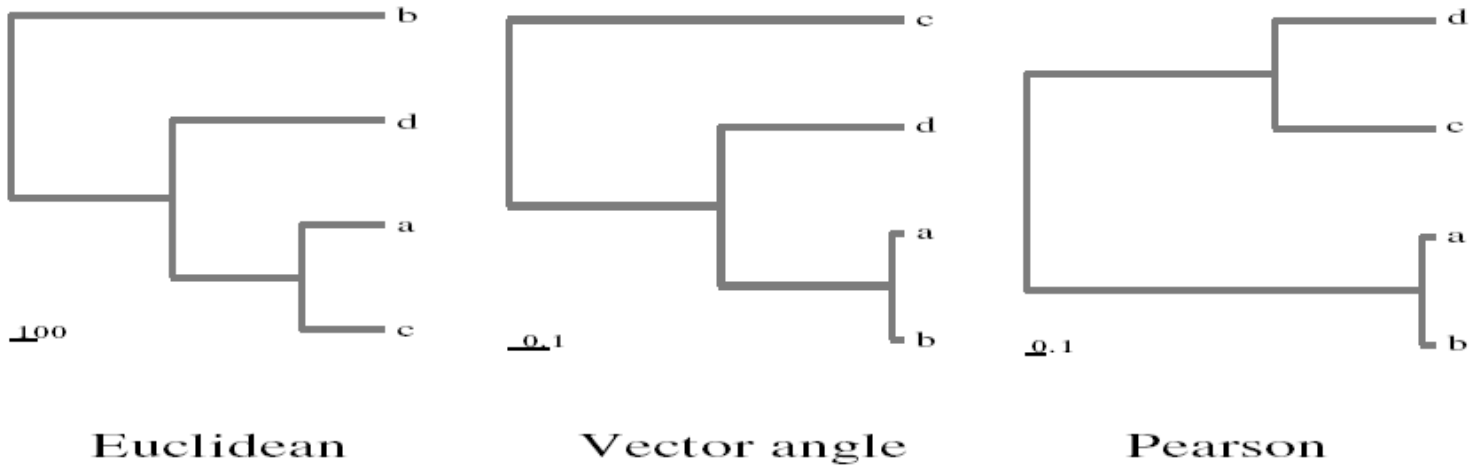
- Vector angle distance

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

- 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



Classification

- Feature selection
- Classification methods
- Cross-validation
- Training and testing

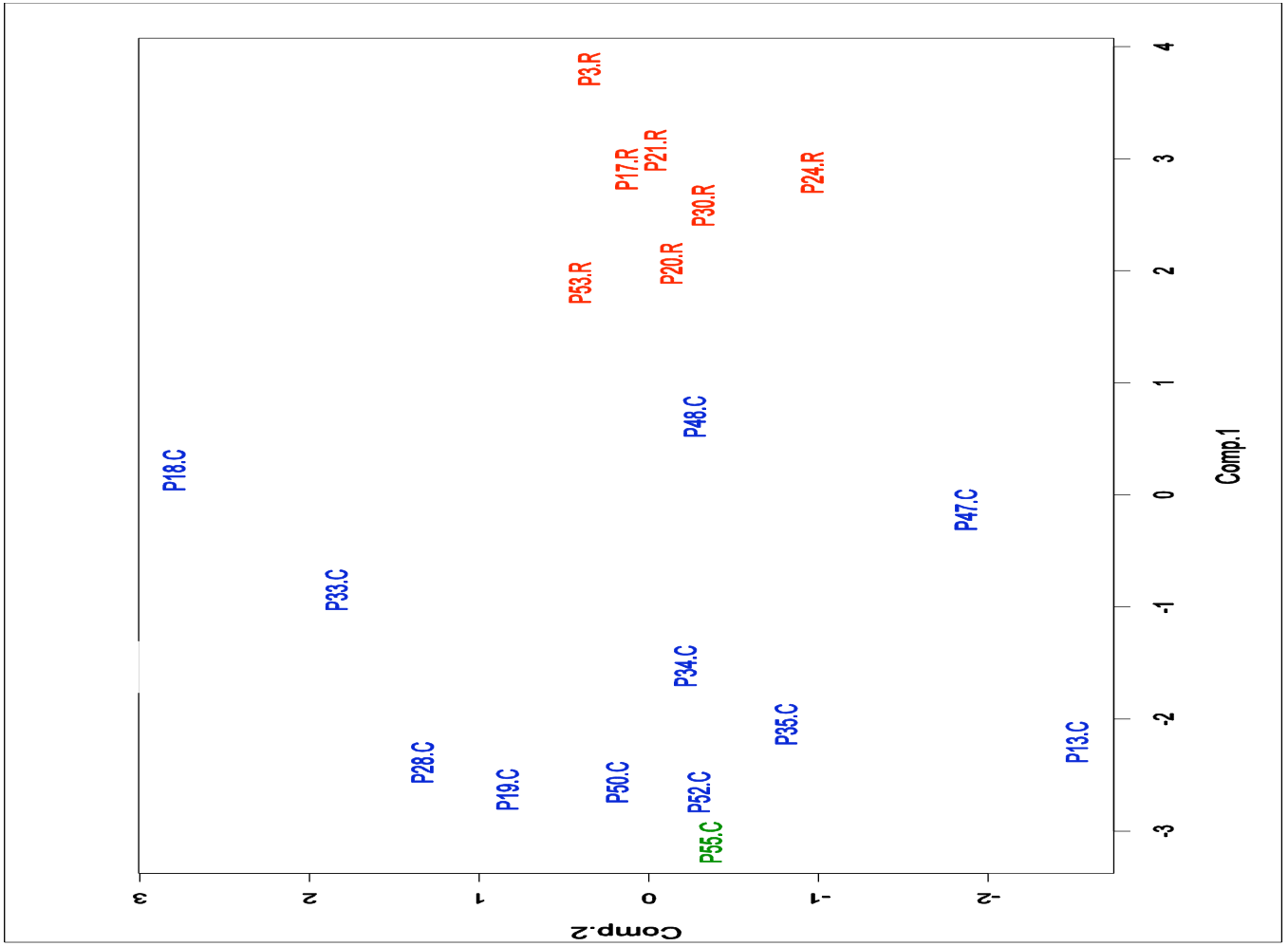
Reduction of input features

- Dimension reduction
  - PCA
- Feature selection (gene selection)
  - Significant genes: t-test
  - Selection of a limited number of genes

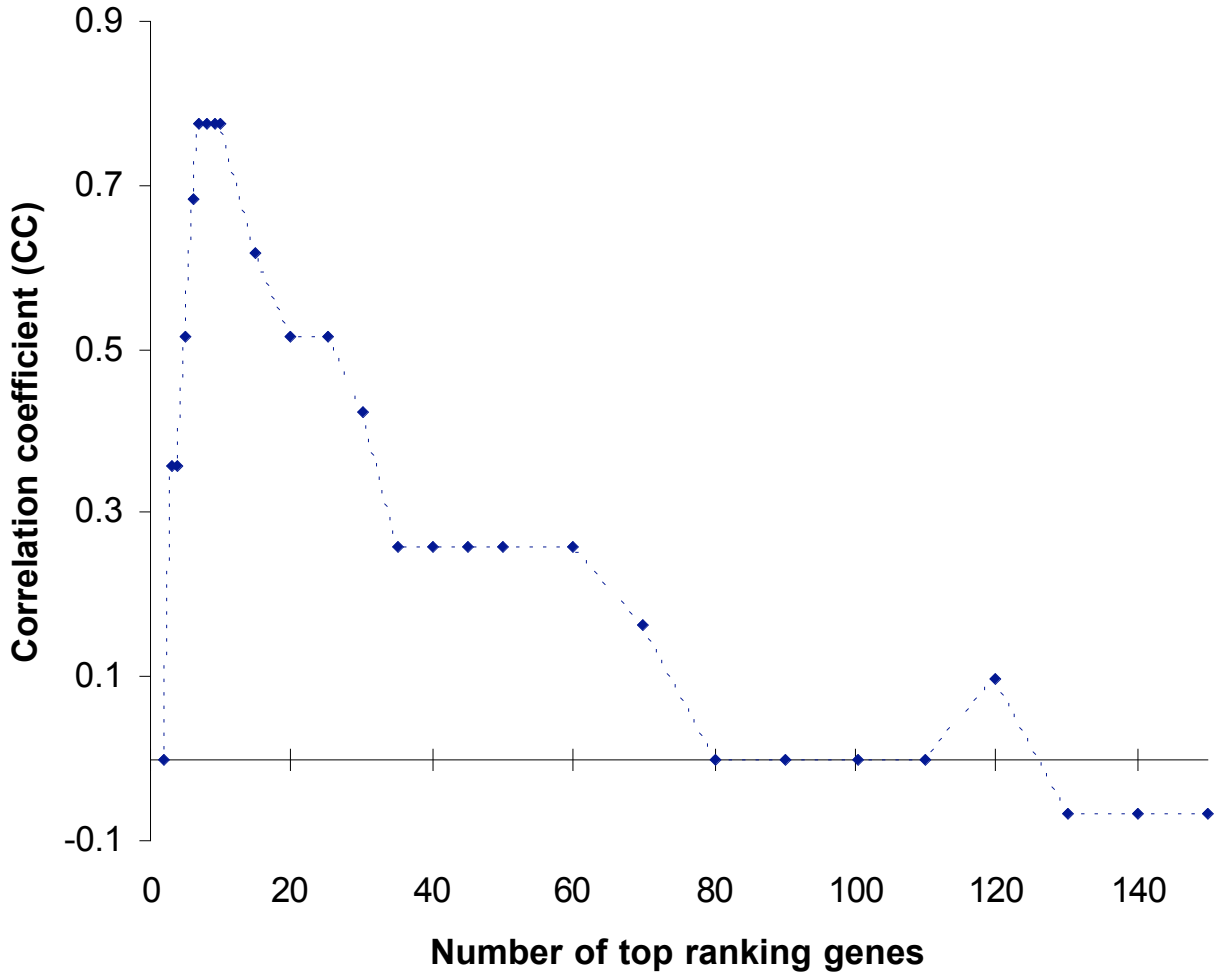
Microarray Data

Class	precursorB		T		...
	Patient1	Patient2	Patient3	Patient4	...
Gene1	1789.5	963.9	2079.5	3243.9	...
Gene2	46.4	52.0	22.3	27.1	...
Gene3	215.6	276.4	245.1	199.1	...
Gene4	176.9	504.6	420.5	380.4	
Gene5	4023.0	3768.6	4257.8	4451.8	
Gene6	12.6	12.1	37.7	38.7	
...	...	...	...	...	...
Gene8793	312.5	415.9	1045.4	1308.0	

Outcome: PCA on Selected Genes

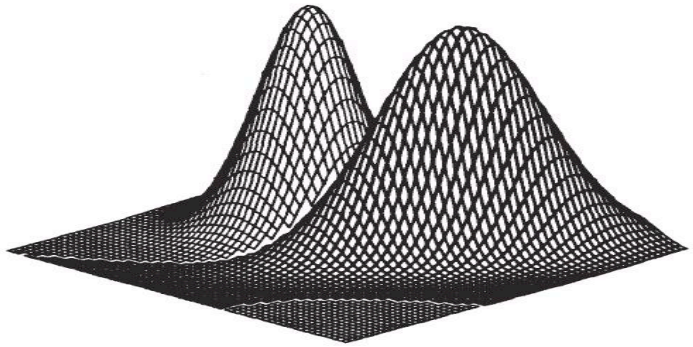


Outcome Prediction:
CC against the Number of Genes



Linear discriminant analysis

- Assumptions:
  - Data eg. Gaussian distributed
  - Variances and covariances the same for classes



Nearest Centroid

- Calculation of a centroid for each class

$$\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} / n_k$$

- Calculation of the distance between a test sample and each class cetroid

- Class prediction by the nearest centroid method

K-Nearest Neighbor (KNN)

- Based on distance measure
  - For example Euclidian distance

- Parameter k = number of nearest neighbors
  - k=1
  - k=3
  - k=...

- Prediction by majority vote for odd numbers

Support Vector Machines

- Machine learning
- Relatively new and highely theoretic
- Works on non-linearly seperable data

- Finding a hyperplane between the two classes by minimizing of the distance between the hyperplane and closest points

Comparison of Methods

Linear discriminant analysis **Nearest centroid** **KNN**	**Neural networks** **Support vector machines**
Simple method	Advanced methods
Based on distance calculation	Involve machine learning
Good for simple problems	Several adjustable parameters
Good for few training samples	Many training samples required (eg. 50-100)
Distribution of data assumed	Flexible methods

Cross-validation

Data: 10 samples

Cross-5-validation:

Training: 4/5 of data (8 samples)

Testing: 1/5 of data (2 samples)

-> 5 different models

Leave-one-out cross-validation (LOOCV)

Training: 9/10 of data (9 samples)

Testing: 1/10 of data (1 sample)

-> 10 different models

Validation

- Definition of
  - true and false positives
  - true and false negatives

Actual class	B	B	T	T
Predicted class	B	T	T	B
	TP	FN	TN	FP

Accuracy

- Definition: $$\frac{TP + TN}{TP + TN + FP + FN}$$

- Range: 0 – 100%

Matthews correlation coefficient

- Definition: $$\frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}}$$

- Range: (-1) – 1

Sensitivity and Specificty

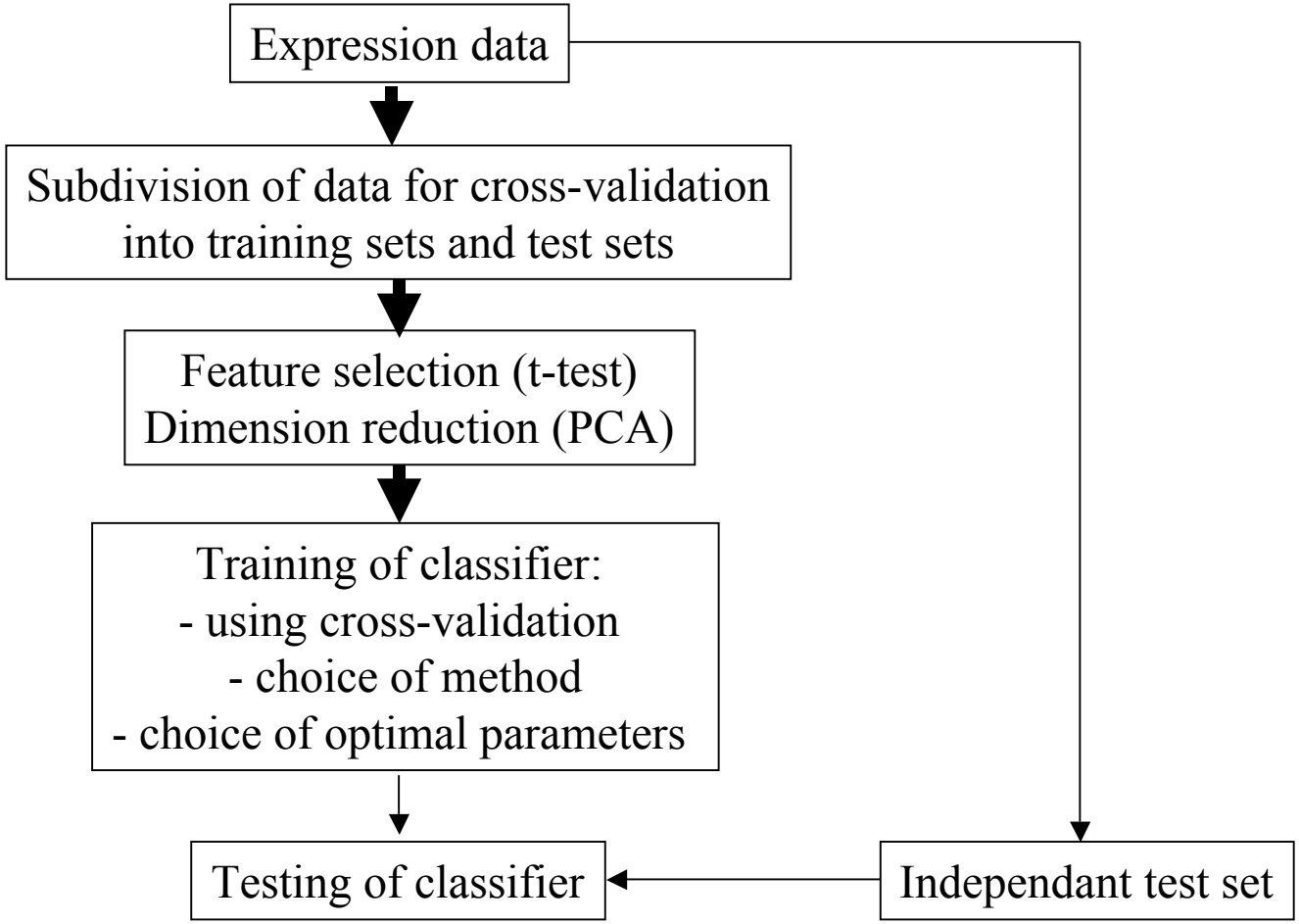
Sensitivity: the ability to **detect "true positives"**

$$\frac{TP}{TP + FN}$$

Specificity: the ability to **avoid "false positives"**

$$\frac{TP}{TP + FP}$$

Overview of Classification



Important Points

- Avoid overfitting
- Validate performance
  - Test on an independant test set
  - by using cross-validation
- Include feature selection in cross-validation

Why?

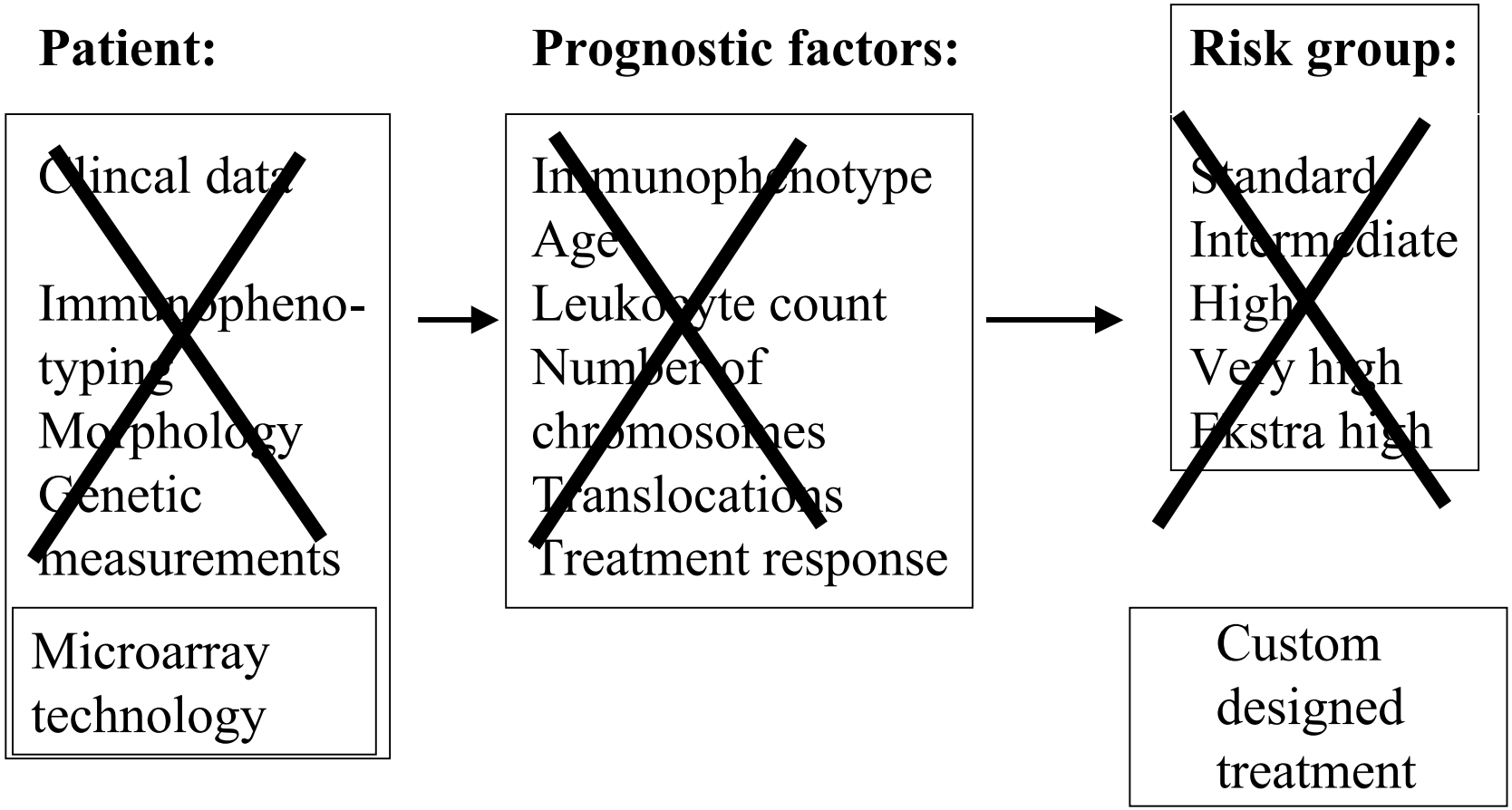
- **To avoid overestimation of performance!**
- **To make a general classifier**

Study of Childhood Leukemia: Results

- **Classification of immunophenotype (precursorB og T)**
  - 100% accuracy
    - During the training
    - When testing on an independant test set
  - Simple classification methods applied
    - K-nearest neighbor
    - Nearest centroid

- **Classification of outcome (relapse or remission)**
  - 78% accuracy (CC = 0.59)
  - Simple and advanced classification methods applied

Risk classification in the future ?



Summary

- Dimension reduction important to visualize data
  - Principal Component Analysis
  - Clustering
    - Hierarchical
    - Partitioning (K-means)

    (distance measure important)

- Classification
  - Reduction of dimension often nessesary (t-test, PCA)
  - Several classification methods avaliable
  - Validation