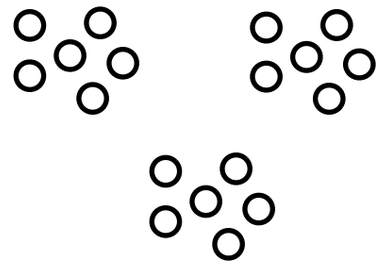
10. Clustering

Introduction to Bioinformatics
30.9.2008
Jarkko Salojärvi

Based on lecture slides by Samuel Kaski

Definition of a cluster



Typically either

1. A group of mutually similar samples, or
2. A mode of the distribution of the samples (more dense than the surroundings)

The definitions depend on the similarity measure or the metric of the data space.

Q: Why clustering? A: Exploratory (descriptive) data analysis

Goal: To make sense of unknown, large data sets by "looking at the data" through

- statistical descriptions
- visualizations

Often additionally: Hunt for discoveries to generate hypotheses for further confirmatory analyses.

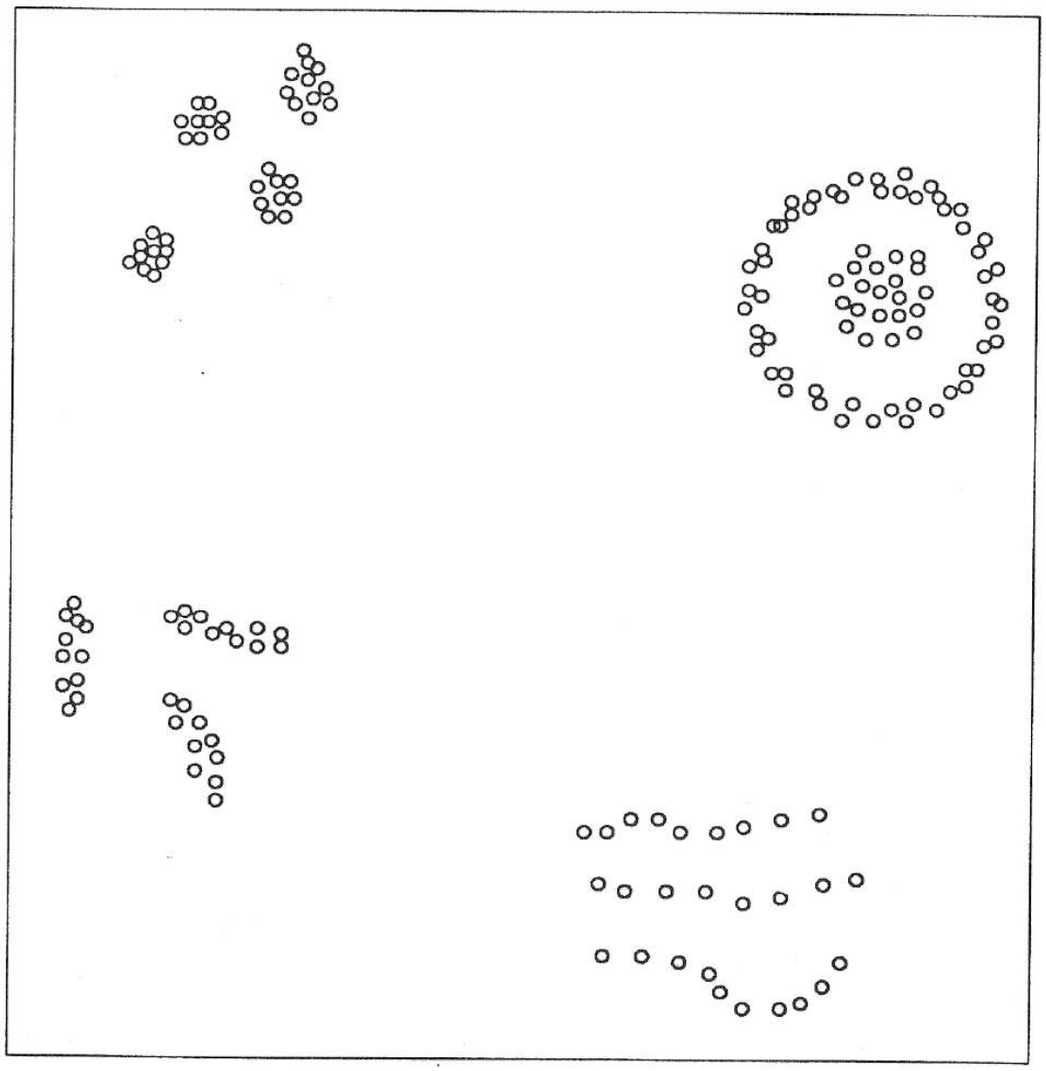
This means flexible model families with additional constraints set by the discovery task, computational and modeling resources, and interpretability.

Goals of clustering

1. **Compression.** Because it is easy to define the cost function for compression, there is a natural goal and criterion for clustering as well: As effective compression as possible.

2. **Discovery of "natural clusters" and description of the data.** There does not exist any single well-posed and generally accepted criterion.

Which are clusters?



Note:

Distinguish between the goal of clustering and the clustering algorithm.

The goal can be defined by

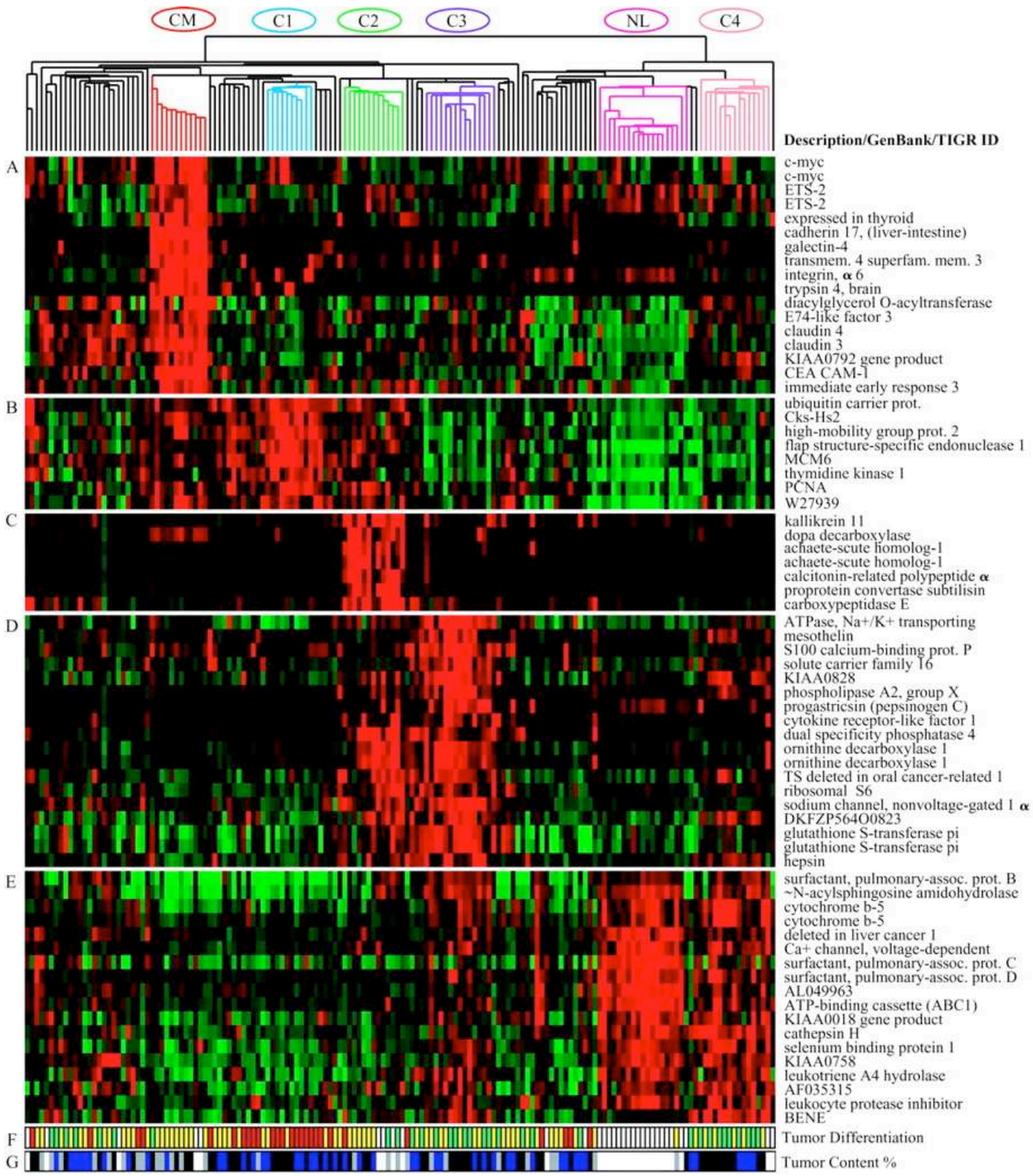
- a cost function to be optimized
- a (statistical) model
- characterizing somehow what a “good” cluster is like
- indirectly by introducing an algorithm

All are only partial solutions; so far nobody has proposed a globally satisfactory definition of a cluster!

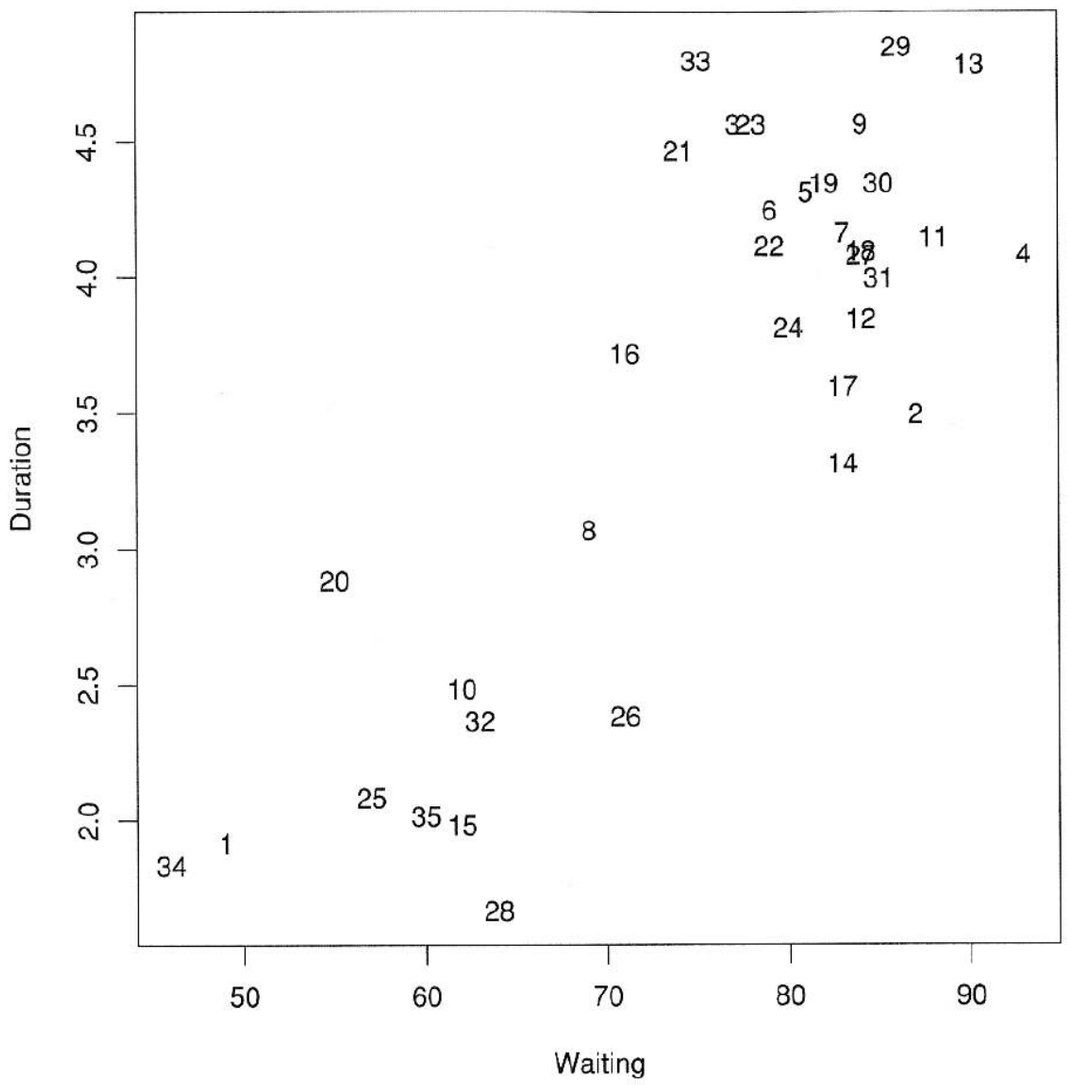
A clustering algorithm describes how the clusters are found, given the goal.

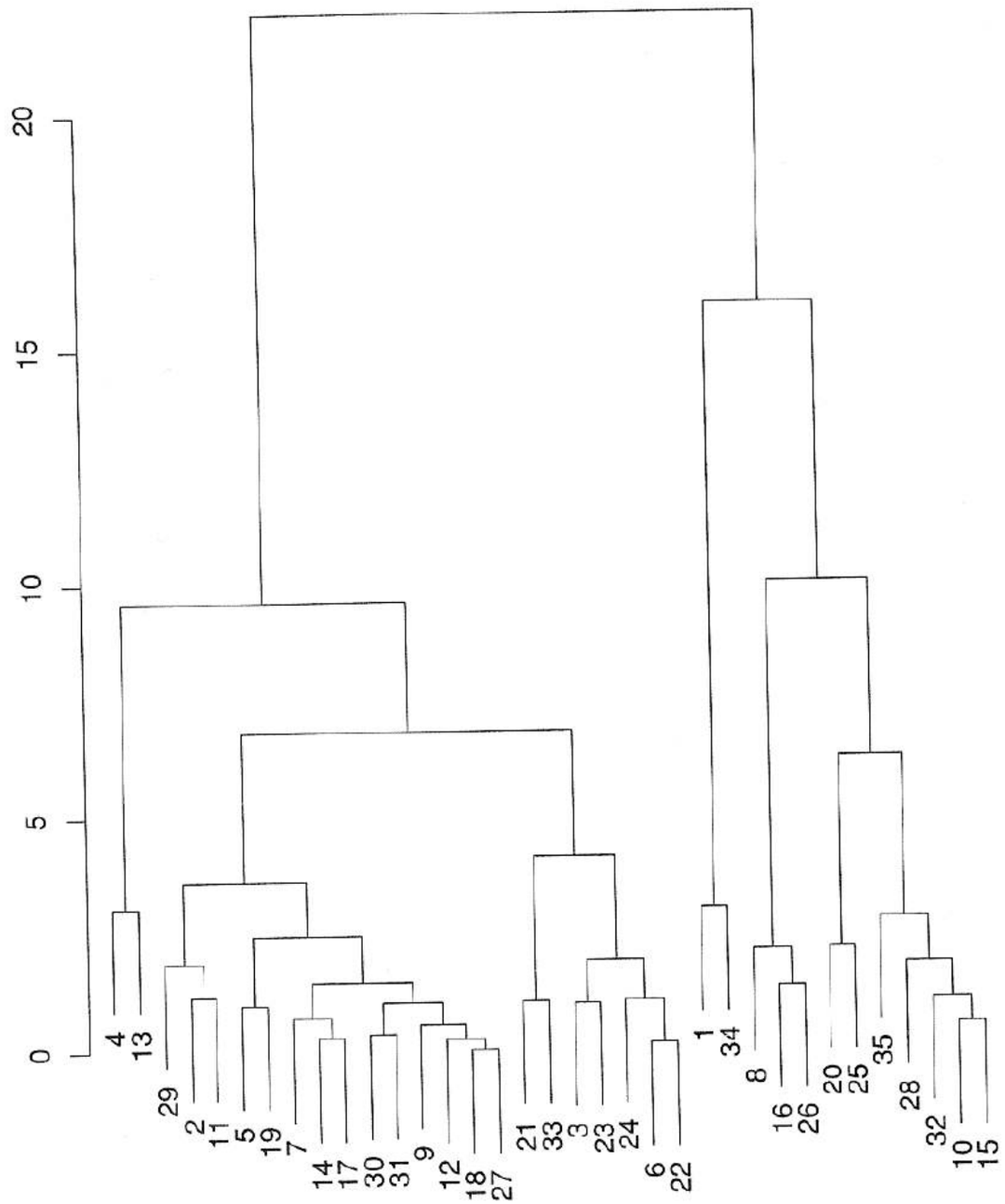
Example: Hierachical clustering of gene expression data

- Data: Expression (activity) of a set of genes measured by DNA chips in tissue samples
- The samples are adenocarcinomas from humans
- The goal is to find sets of mutually similar tissue samples. Maybe subcategories will be found that respond differentially to treatments.


CM
C1
C2
C3
NL
C4
Description/GenBank/TIGR ID
A
c-myc
c-myc
ETS-2
ETS-2
expressed in thyroid
cadherin 17, (liver-intestine)
galectin-4
transmem. 4 superfam. mem. 3
integrin, α 6
trypsin 4, brain
diacylglycerol O-acyltransferase
E74-like factor 3
claudin 4
claudin 3
KIAA0792 gene product
CEA CAM-1
immediate early response 3
B
ubiquitin carrier prot.
Cks-Hs2
high-mobility group prot. 2
flap structure-specific endonuclease 1
MCM6
thymidine kinase 1
PCNA
W27939
C
kallikrein 11
dopa decarboxylase
achaete-scute homolog-1
achaete-scute homolog-1
calcitonin-related polypeptide α
proprotein convertase subtilisin
carboxypeptidase E
D
ATPase, Na+/K+ transporting
mesothelin
S100 calcium-binding prot. P
solute carrier family 16
KIAA0828
phospholipase A2, group X
progastricsin (pepsinogen C)
cytokine receptor-like factor 1
dual specificity phosphatase 4
ornithine decarboxylase 1
ornithine decarboxylase 1
TS deleted in oral cancer-related 1
ribosomal S6
sodium channel, nonvoltage-gated 1 α
DKFZP564O0823
glutathione S-transferase pi
glutathione S-transferase pi
hepsin
E
surfactant, pulmonary-assoc. prot. B
~N-acylsphingosine amidohydrolase
cytochrome b-5
cytochrome b-5
deleted in liver cancer 1
Ca+ channel, voltage-dependent
surfactant, pulmonary-assoc. prot. C
surfactant, pulmonary-assoc. prot. D
AL049963
ATP-binding cassette (ABC1)
KIAA0018 gene product
cathepsin H
selenium binding protein 1
KIAA0758
leukotriene A4 hydrolase
AF035315
leukocyte protease inhibitor
BENE
F
Tumor Differentiation
G
Tumor Content %

How was the clustering carried out?





Variants

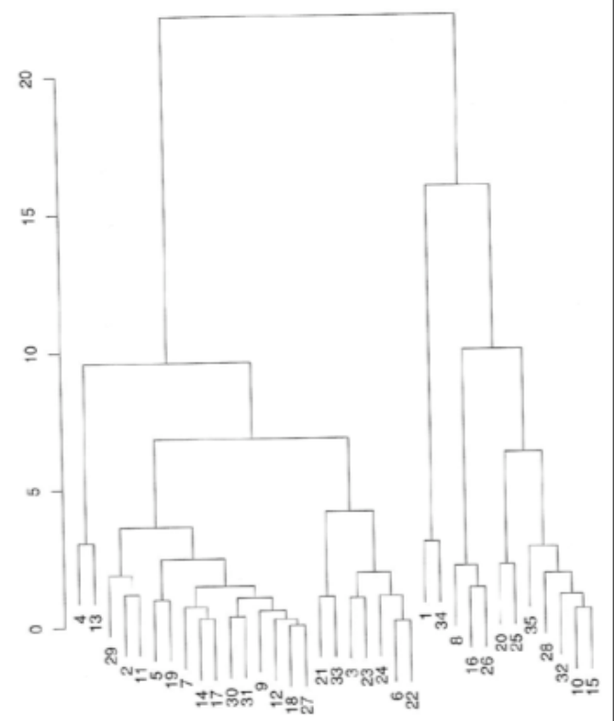
Agglomerative vs. divisive clustering

Different criteria for agglomeration and division:

- single linkage
- complete linkage
- average linkage (UPGMA)
- Ward etc.

Pros and cons of hierarchical clustering

+ The result is intuitive and easily interpretable.
+ The dendrogram can be used for both (i) displaying similarity relationships between clusters and (ii) partitioning by cutting at different heights.
+ Possibly tedious to interpret for large data sets
- Sensitivity to noise
- Clustering has been defined by an algorithm. Can the result be described as such? Is there a goodness criterion?



Partitional clustering

Definition of a cluster:

Assume a distance measure $d(\mathbf{x},\mathbf{y})$ and define a cluster based on it:

A cluster consists of a set of samples having small mutual distances, that is,

$$E_k = \sum_{w(\mathbf{x})=w(\mathbf{y})=k} d^2(\mathbf{x},\mathbf{y})$$

is small. Here the cluster of sample $\mathbf{x}$ has been indexed by $w(\mathbf{x})$.

Partitional clustering algorithm

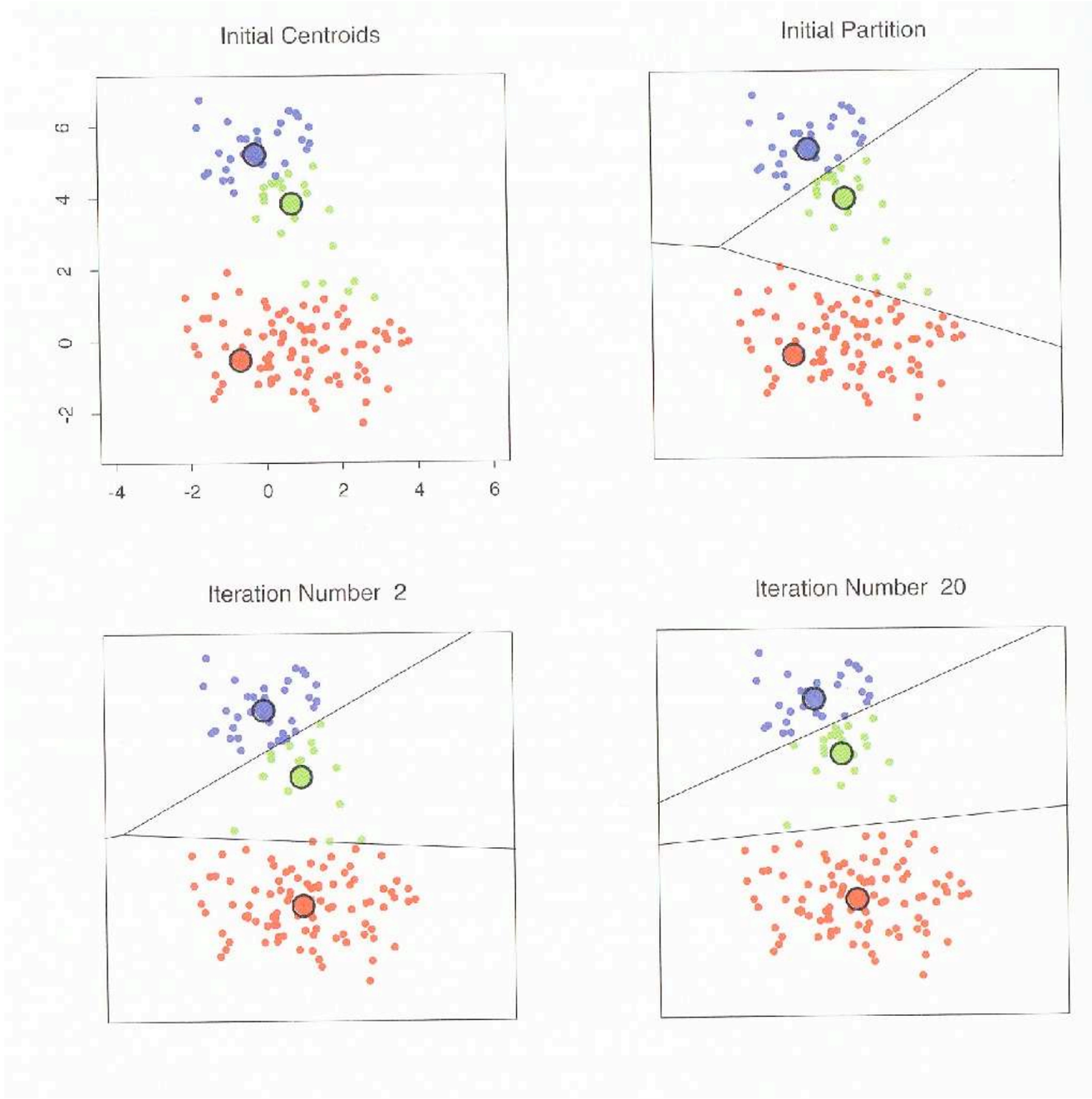
A partitional clustering algorithm tries to assign the samples to clusters such that mutual distances are small in *all clusters*.

In other words, the cost function

$$E = \sum_k E_k$$

is minimized.

In the K-means algorithm the distance measure is Euclidean, and the clusters are defined by a set of K *cluster prototypes*: Samples are assigned to the cluster with the closest prototype.


Initial Centroids
Initial Partition
Iteration Number 2
Iteration Number 20
-4
-2
0
2
4
6

Pros and cons of partitional clustering

+ Fast (although not faster than hierarchical clustering)
+ The result is intuitive, although possibly tedious to interpret for large data sets

- The number of clusters K must be chosen, which may be difficult
- Tries to find “spherical” clusters in the sense of the given distance measure. (This may be the desired result, though.)

Model-based clustering: Mixture density model

Assume that each sample $\mathbf{x}$ has been generated by one generator $k(\mathbf{x})$, but it is not known which one.

Assume that the generator k produces the probability distribution $p_k(\mathbf{x};\theta_k)$, where θ_k contains the parameters of the density.

Assume further that the probability that generator k produces a sample is p_k.

The probability density generated by the mixture is

$$p(\mathbf{x}) = \sum_k p_k(\mathbf{x};\theta_k)p_k$$

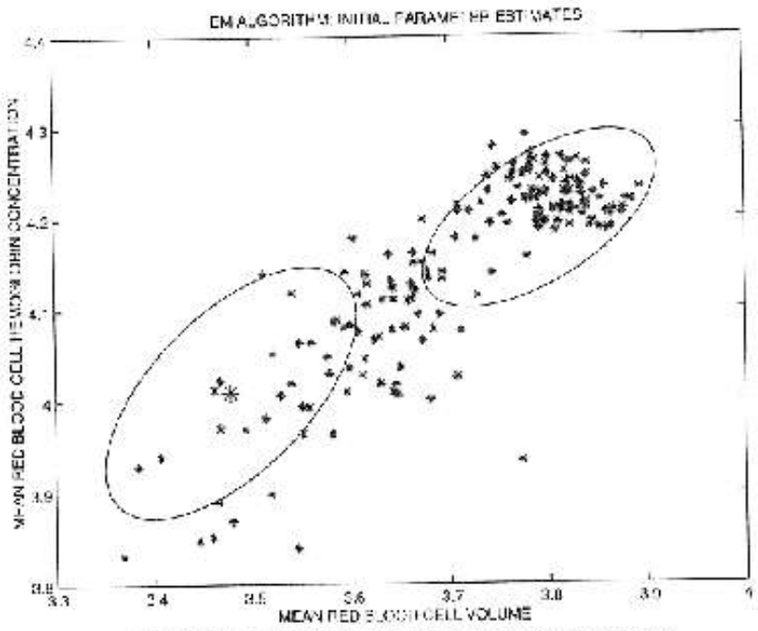
The model can be fitted to the data set with basic methods of statistical estimation:

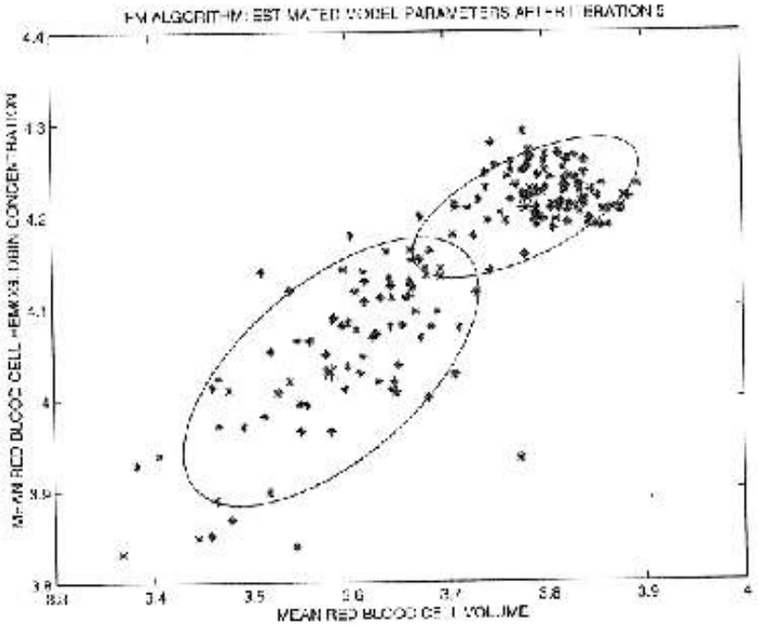
- maximum likelihood
- maximum a posterior

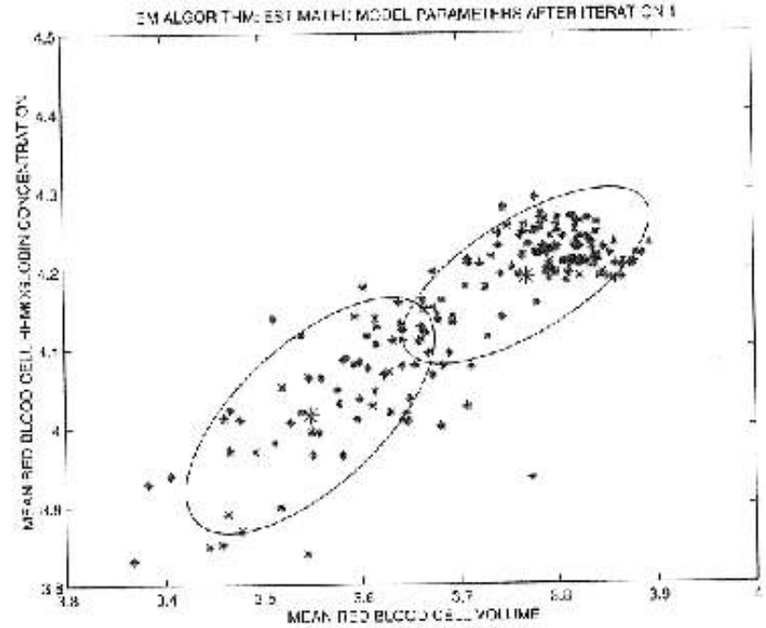
Conveniently optimizable by EM-based algorithms.

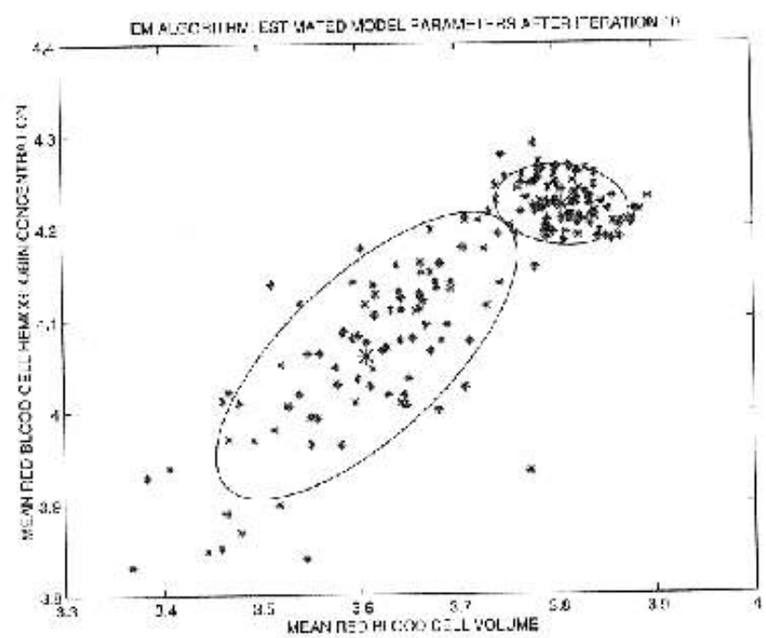
Suitable model complexity (number of clusters) can be learned by Bayesian methods, approximated by BIC (or AIC, MDL, ...)

Note that K-means is obtained as the limit when generators of normal distributions sharpen.


EM ALGORITHM: INITIAL PARAMETER ESTIMATES
MEAN RED BLOOD CELL HEMOGLOBIN CONCENTRATION
MEAN RED BLOOD CELL VOLUME


EM ALGORITHM: ESTIMATED MODEL PARAMETERS AFTER ITERATION 5
MEAN RED BLOOD CELL HEMOGLOBIN CONCENTRATION
MEAN RED BLOOD CELL VOLUME


EM ALGORITHM: ESTIMATED MODEL PARAMETERS AFTER ITERATION 1
MEAN RED BLOOD CELL HEMOGLOBIN CONCENTRATION
MEAN RED BLOOD CELL VOLUME


EM ALGORITHM: ESTIMATED MODEL PARAMETERS AFTER ITERATION 10
MEAN RED BLOOD CELL HEMOGLOBIN CONCENTRATION
MEAN RED BLOOD CELL VOLUME

EM ALGORITHM: ESTIMATED MODEL PARAMETERS AFTER ITERATION 2
MEAN RED BLOOD CELL HEMOGLOBIN CONCENTRATION
MEAN RED BLOOD CELL VOLUME

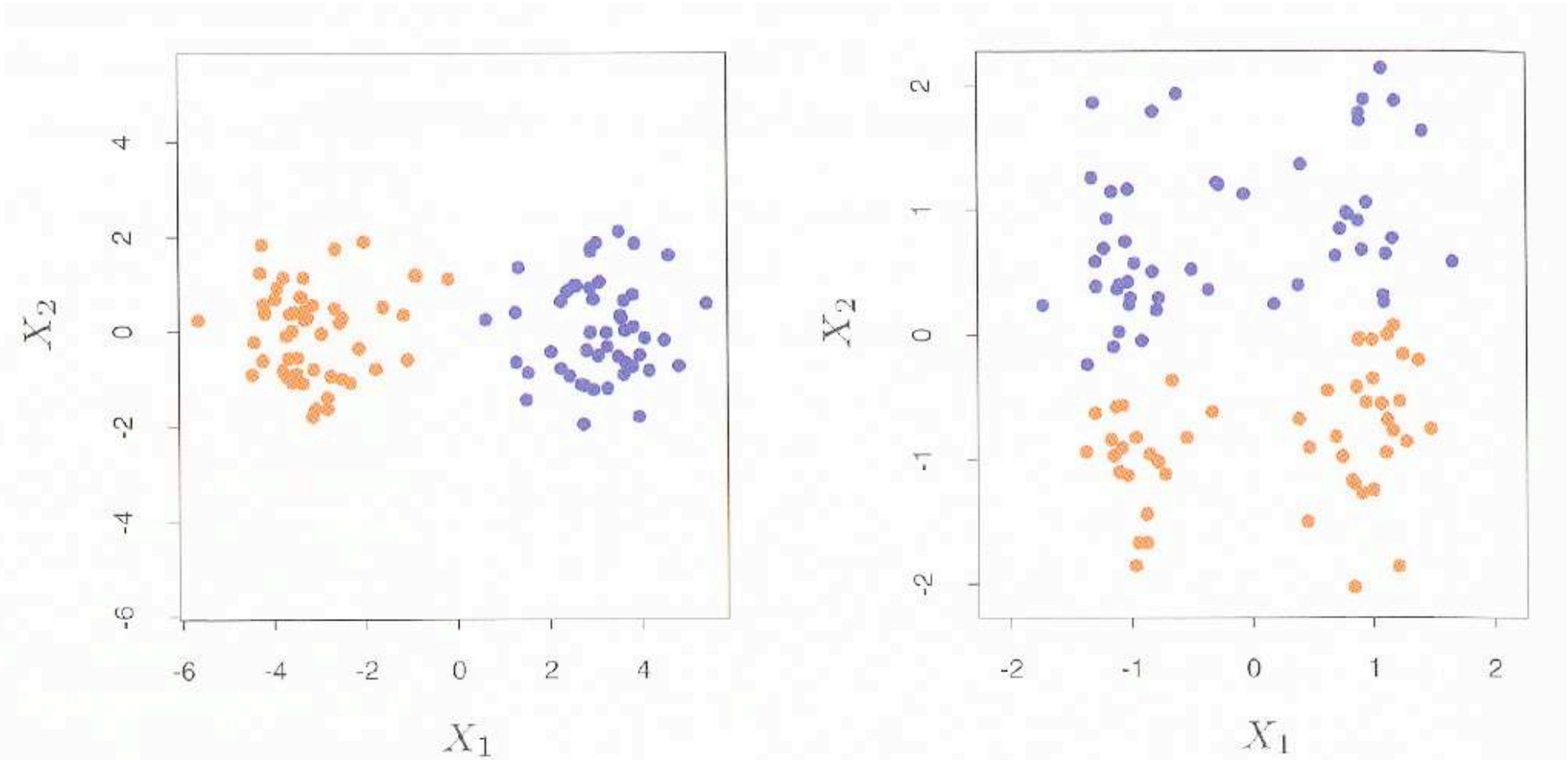
EM ALGORITHM: ESTIMATED MODEL PARAMETERS AFTER ITERATION 25
MEAN RED BLOOD CELL HEMOGLOBIN CONCENTRATION
MEAN RED BLOOD CELL VOLUME

Pros and cons of clustering by mixture density models

+ The model is well-defined. It is based on explicit and clear assumptions on the uncertainty within the data
+ As a result, all tools of probabilistic inference are applicable:
  + evaluation of the generalizability and quality of the result
  + choosing the number of clusters
- Is the goal of clustering the same as the goal of density estimation? The probabilistic tools work properly only if the assumptions are correct!

Pitfalls

Clusteredness depends on scaling



GIGO Principle

Supervised learning:

Garbage in $\Rightarrow$ weaker results out

Unsupervised learning:

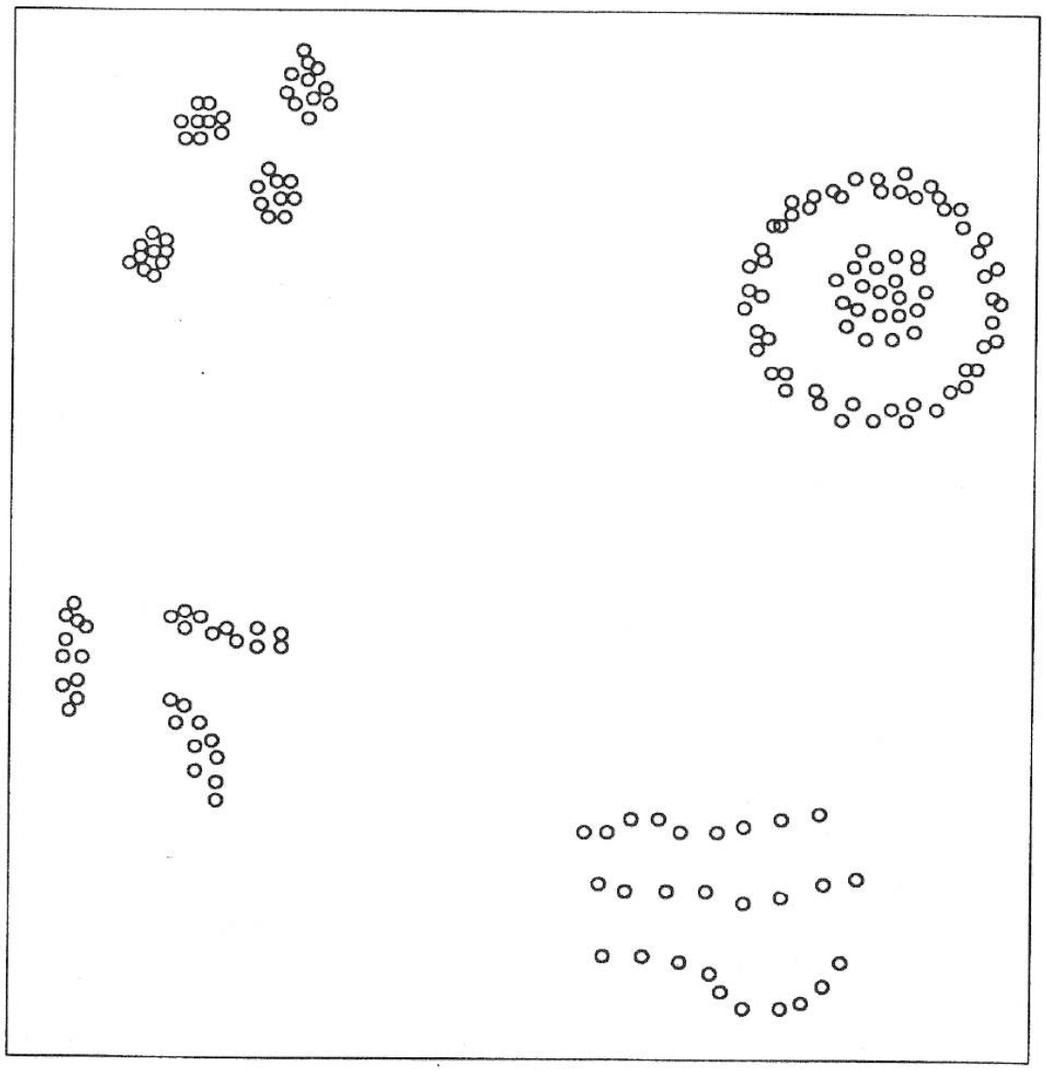
Garbage in $\Rightarrow$ garbage out

(Successful) unsupervised learning is always implicitly supervised

by

- feature extraction
- variable selection
- model selection

Which are clusters?



Distance measures

Absolute magnitudes / Zero level	Reliable	Unreliable
Interesting	Euclidean metric	(Euclidean with mean subtracted)
Not interesting	Inner product	Correlation

Accoding to some studies (including ours) the correlation may be best.

About metrics

Euclidean metric:

$$d_E^2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2 = (\mathbf{x}-\mathbf{y})^T\mathbf{I}(\mathbf{x}-\mathbf{y})$$

Becomes (essentially) inner products for normalized vectors, $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$:

$$d_E^2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}^T\mathbf{y} = 2(1-\mathbf{x}^T\mathbf{y})$$

Correlation (with vector components interpreted as samples of the same random variable, and σ_x being standard deviation of $\mathbf{x}$)

$$\rho(\mathbf{x},\mathbf{y}) = \frac{(\mathbf{x}-\bar{\mathbf{x}})^T(\mathbf{y}-\bar{\mathbf{y}})}{\sigma_x\sigma_y}$$

becomes inner products by Z-score normalization, $z = (\mathbf{x}-\bar{\mathbf{x}})/\sigma_x$.

Global metric for $\mathbf{A} = \mathbf{S}^T\mathbf{S}$ is

$$d_A^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{A}(\mathbf{x}-\mathbf{y}) = \|\mathbf{S}\mathbf{x}-\mathbf{S}\mathbf{y}\|^2$$

Local (Riemannian) metric for $\mathbf{y} = \mathbf{x} + d\mathbf{x}$ is

$$d_{\mathbf{A}(\mathbf{x})}^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{A}(\mathbf{x})(\mathbf{x}-\mathbf{y})$$

Number of clusters?

In principle: Use the normal model complexity selection methods.

Lots of more or less heuristic solutions exist.

One possible solution: Visualization

Cluster validation

(Selecting the number of clusters is a sub-problem of this.)

Since the data exploration process necessarily is partly subjective, the results must be validated: Are the clusters/other findings real?

Fundamentally boils down to generalizability to new data (which can be assessed by measuring more data!)

Bayesian averaging over models is hard because of

- label switching
- the end result will be discovery or "understanding of data." Since we do not know how humans do that, it is hard to assign proper priors (=choose model families) for the analysis.

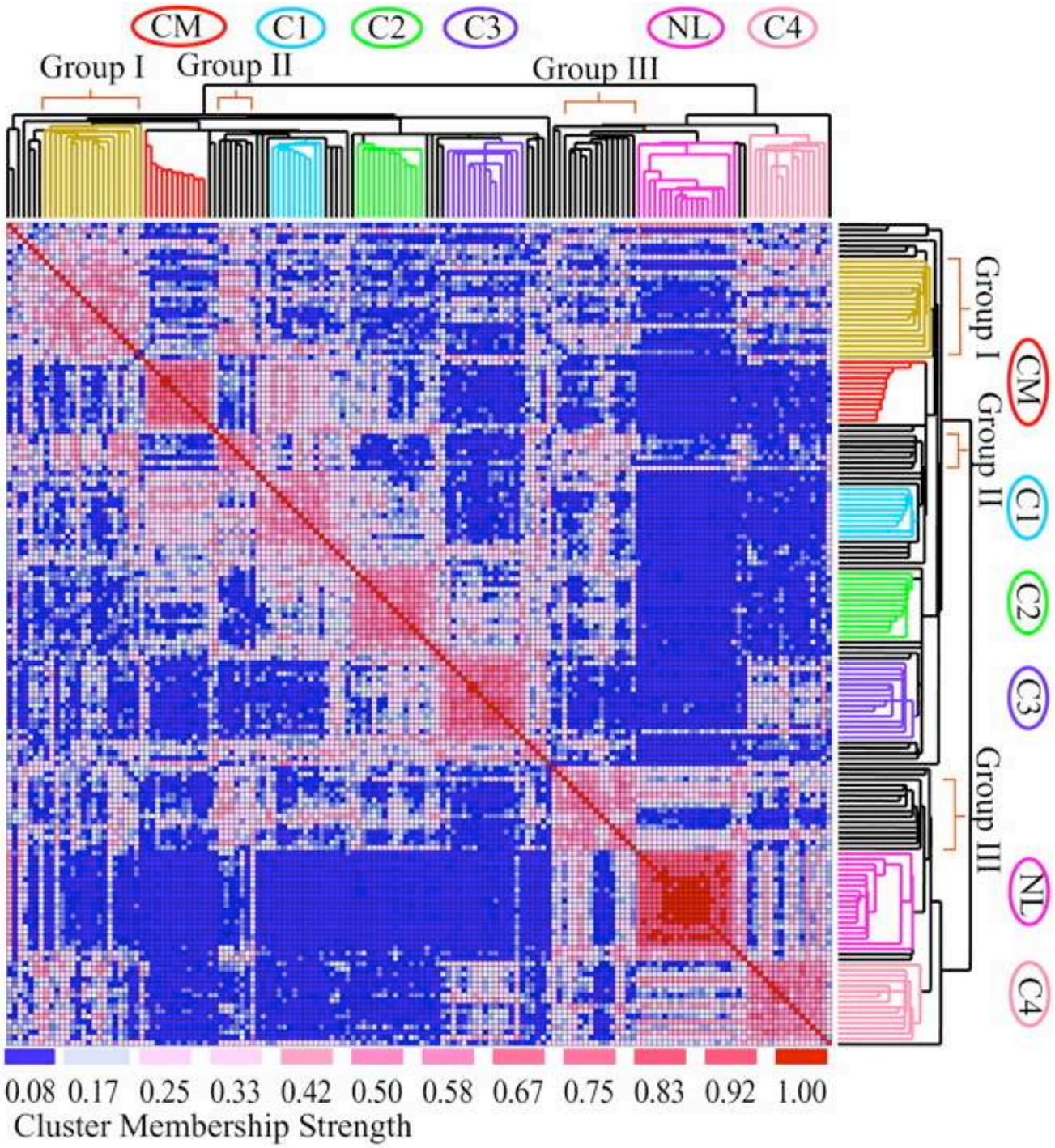
A temporary solution is to use cross-validation or bootstrap.

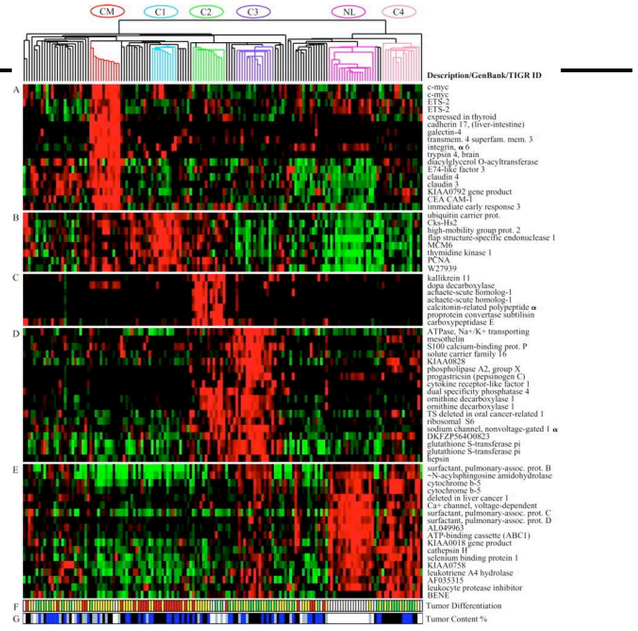
Bhattacharjee et al: Similarity of samples from a mixture model

Quantize the robustness of the clustering results to random variations in the observed data:

- Construct lots of (200) bootstrapped data sets by sampling with replacement from the original data
- Cluster each new set
- For each pair of samples $(\mathbf{x}, \mathbf{y})$, compute the strength of association as the percentage of times they become clustered into the same cluster

Recap: hierarchical clustering





Discussion

- Strengthens the faith to the hierarchical clustering
- Not a very illustrative visualization without the hierarchical clustering
- Would there exist a better clustering in the new similarity measure induced by the bootstrapping procedure?
- Is robustness to variation a good indication of clusteredness? The robust features may not be biologically interesting? ($\Rightarrow$ external criteria might be better)

Conclusions

Ill-defined problem with lost of proposed solutions.

Words of advice:

- The reason is that there actually are lots of different clustering tasks with different goals and not enough prior knowledge to define the problem exactly.
- This does not imply that sloppy application of clustering methods would be acceptable!
- In contrast, you have to understand the principles and key ideas, in order to use your prior knowledge to choose suitable methods to your specific task.
- Check the validity of the results somehow.

Material

A.K. Jain, M.N. Murty and P.J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999. (A good review.)

V. Estivill-Castro. Why so many clustering algorithms—A position paper. SIGKDD Explorations, 4(1):65-75. (I do not agree with everything but describes many of the problems in defining clusters.)

These papers contain some of the case studies discussed in the lectures:

A. Bhattacharjee, W. G. Richards, and J. S. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98:13790–13795, 2001.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

+ the same old book