Visualizing data structures using R :

An introduction to unsupervised learning: Clustering

Ziv Shkedy

Hasselt University, Belgium

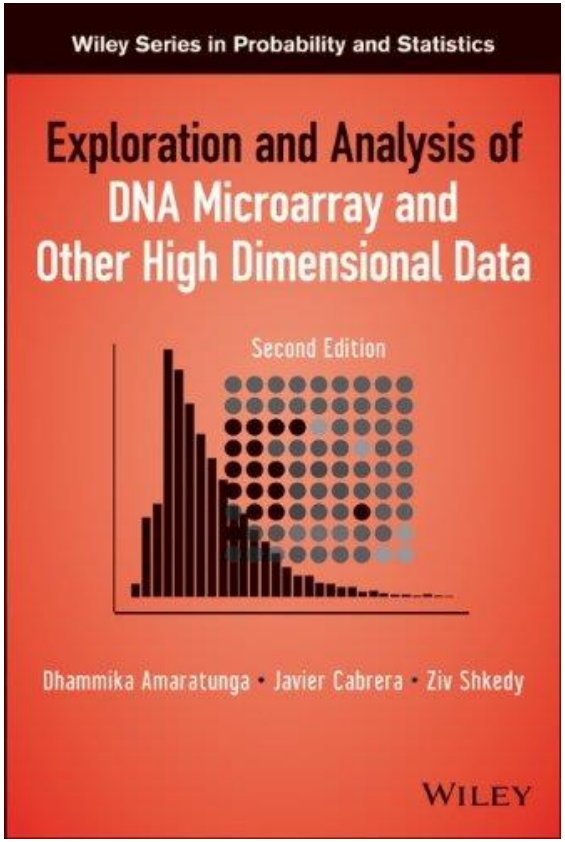
Master in statistics: Bioinformatics

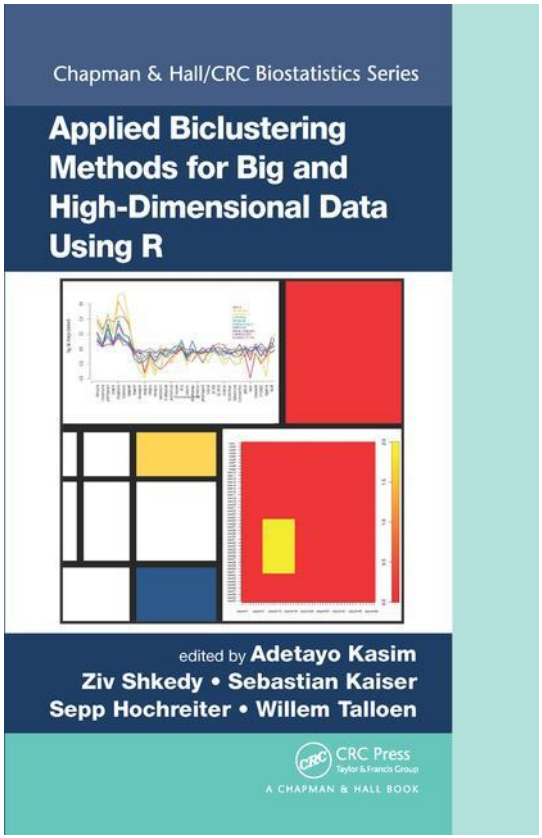
September-December 2021

Contributors

- Slides and materials were developed jointly with:
  - Dan Lin
  - Adetayo Kasim
  - Ziv Shkedy
  - Dhammika Amaratunga
  - Javier Cabrera
  - Martin Otava
  - Nolen J Perualila

References





- Chapter 10:
  - pattern discovery.

- Chapter 2:
  - clustering & biclsutering

Introduction

Gene Expression Data Analysis

Three main types of statistical problems associated with the microarray data:

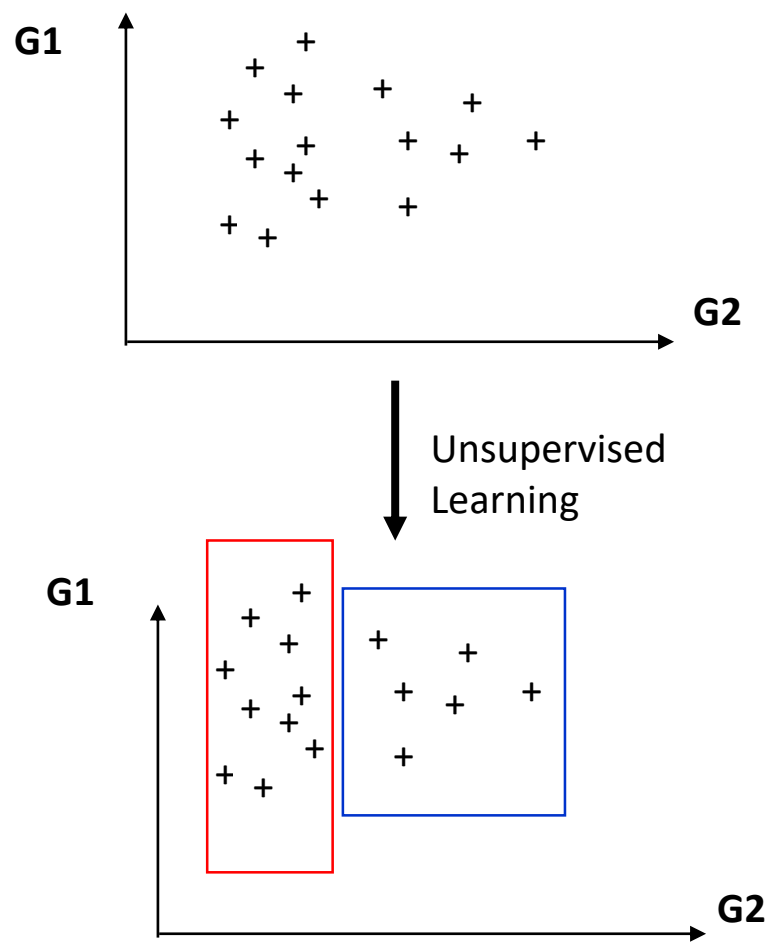
- Identification of "marker" genes that characterize the different tumor classes (feature or variable selection).
- Identification of new/unknown tumor classes using gene expression profiles (unsupervised learning – clustering)
- Classification of sample into known classes (supervised learning – classification)

Classification vs. Clustering

Classification	Clustering
• known number of classes	• unknown number of classes
• based on a training set	• no prior knowledge
• used to classify future observations	• used to understand (explore) data
• Classification is a form of supervised learning	• Clustering a form of unsupervised learning

Classification vs. Clustering

- Data elements are simply defined in terms of G1 and G2 values
- find groups whose elements are similar to another based on the feature values



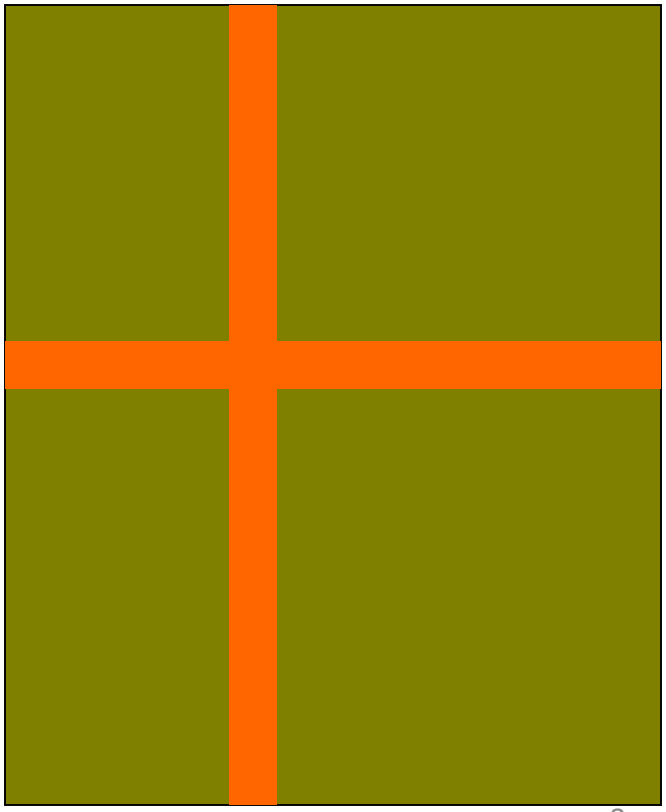
Clustering microarray data

- Genes and experiments/samples are given as the row and column vectors of a gene expression data matrix.
- Clustering may be applied either to genes or experiments (regarded as vectors in $\mathbf{R}^p$ or $\mathbf{R}^n$).

gene expression data matrix

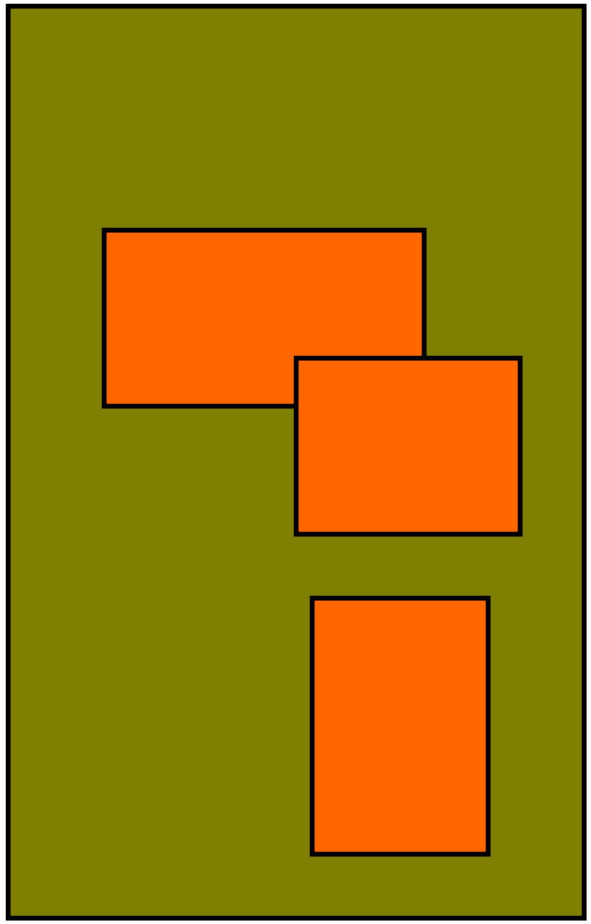
n experiments

p genes



Clustering microarray data

- We can cluster genes (rows), mRNA samples (cols), or both at once.
- Identify groups of possibly co-regulated genes (e.g. in conjunction with sequence data).
- Identify typical temporal or spatial gene expression patterns (e.g. cell cycle data).



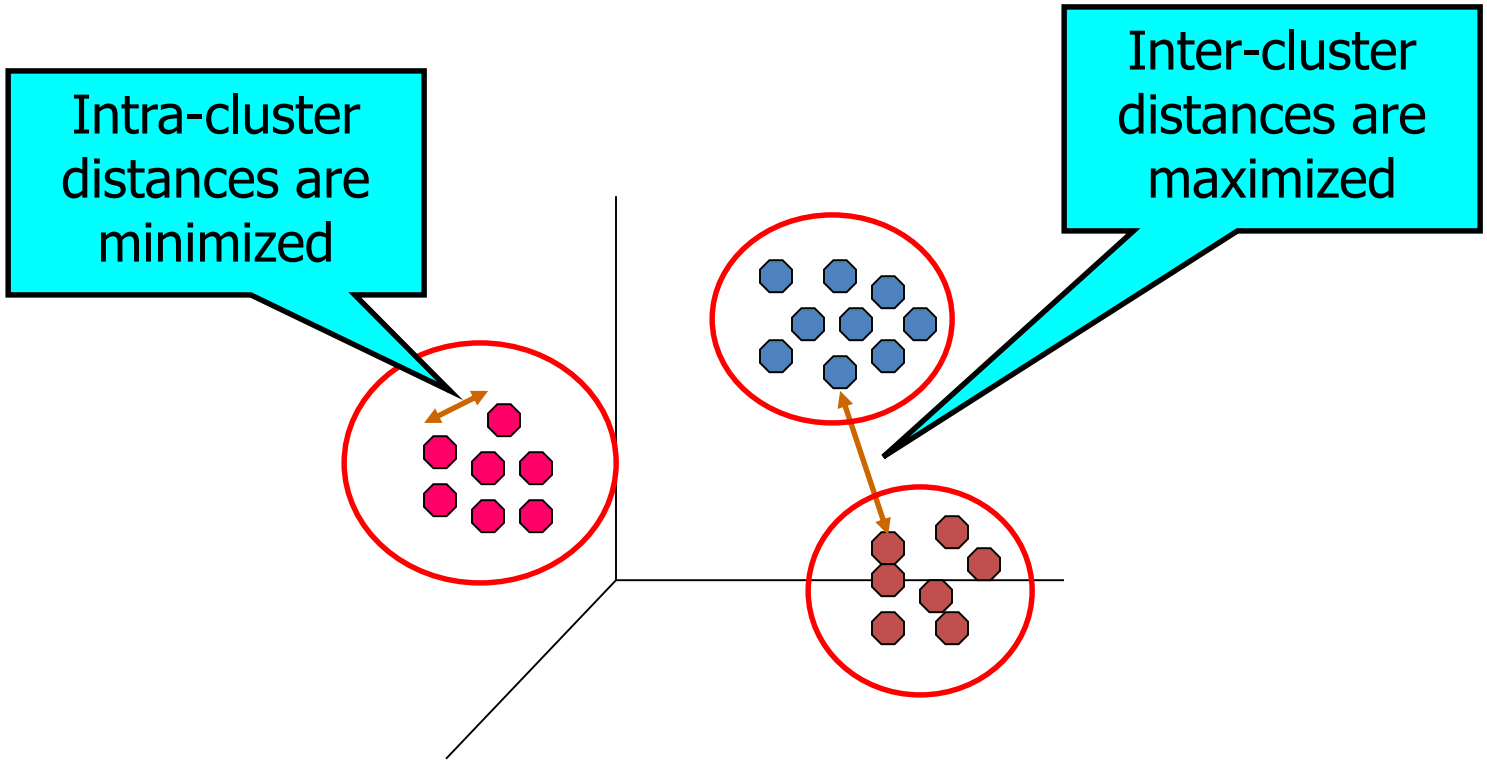
Why cluster genes?

- Clustering leads to readily interpretable figures.
- Clustering can be helpful for identifying patterns in time or space.
- Clustering is useful, perhaps essential, when seeking new subclasses of cell samples (tumors, etc).

⇨ Pattern discovery.

Cluster Analysis

- Find groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.



Three generic clustering problems

Three important tasks (which are generic) are:

1. Estimating the number of clusters;
2. Assigning each observation to a cluster;
3. Assessing the strength/confidence of cluster assignments for individual observations.

.

Basic principles of clustering

Aim: to group observations that are “similar” based on predefined criteria.

Issues: Which feature / samples to use?

Which similarity or dissimilarity measure?

Which clustering algorithm?

- It is advisable to **reduce** the number of featuers from the full set to some more manageable number, before clustering.

Data structure

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix}.$$

Variables, features…

Observations, samples, conditions ….

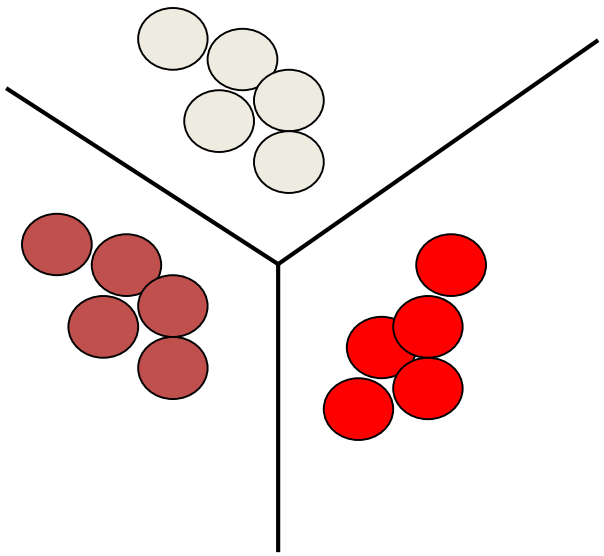
Global patterns

- Find variables (observations) that can be grouped together due to a pattern in the data matrix.

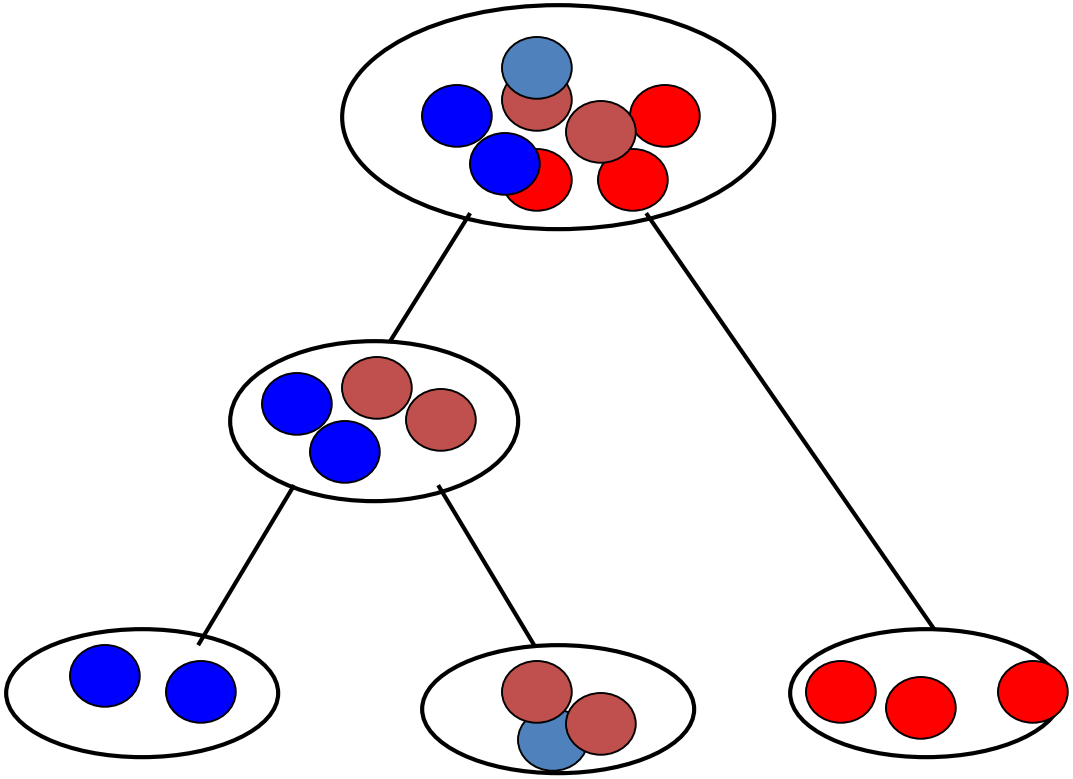
- Examples:
  - All costumers in a supermarkets that have a tendency to buy the same products.
  - Genes with the same expression profiles in an expression matrix.

Two basic types of methods

Partitioning



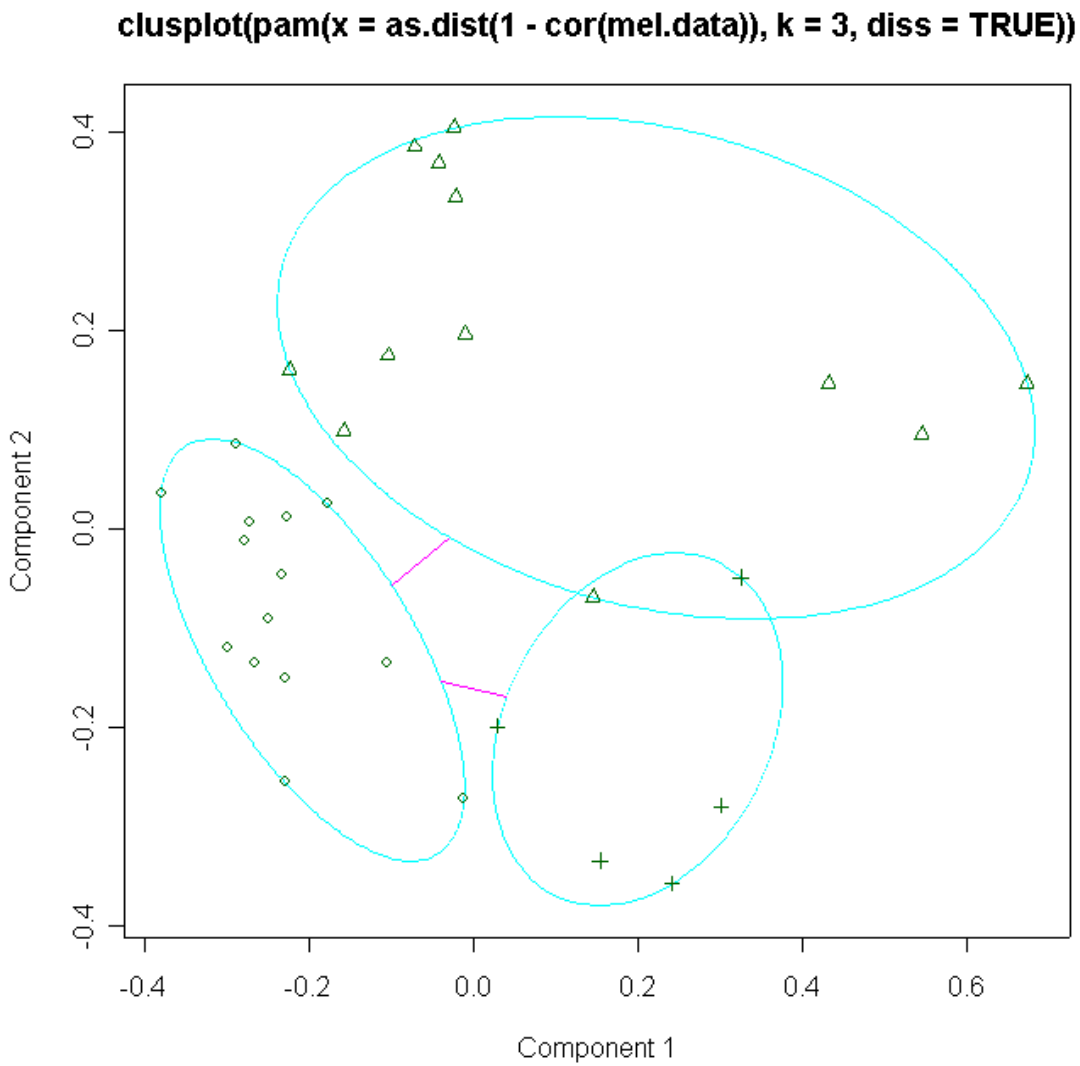
Hierarchical



Two commonly seen clustering approaches in gene expression data analysis

- Hierarchical clustering:
  - Clustering tree (=Dendrogram).
  - Do not need to select the number of clusters in advance.
- K-means/K-medoids:
  - Partitioning method.
  - Requires user to define K = # of clusters a priori.

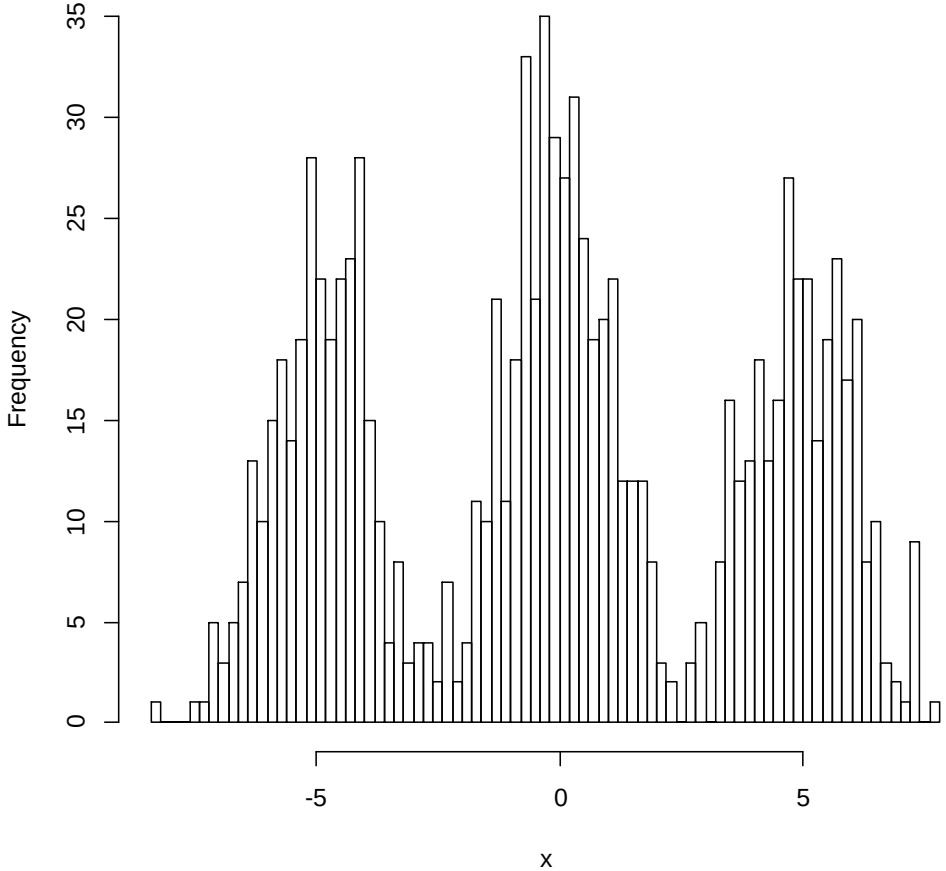
Example


clusplot(pam(x = as.dist(1 - cor(mel.data)), k = 3, diss = TRUE))
Component 2
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
0.4
0.6
Component 1
These two components explain 37.03 % of the point variability.

Issues in Clustering

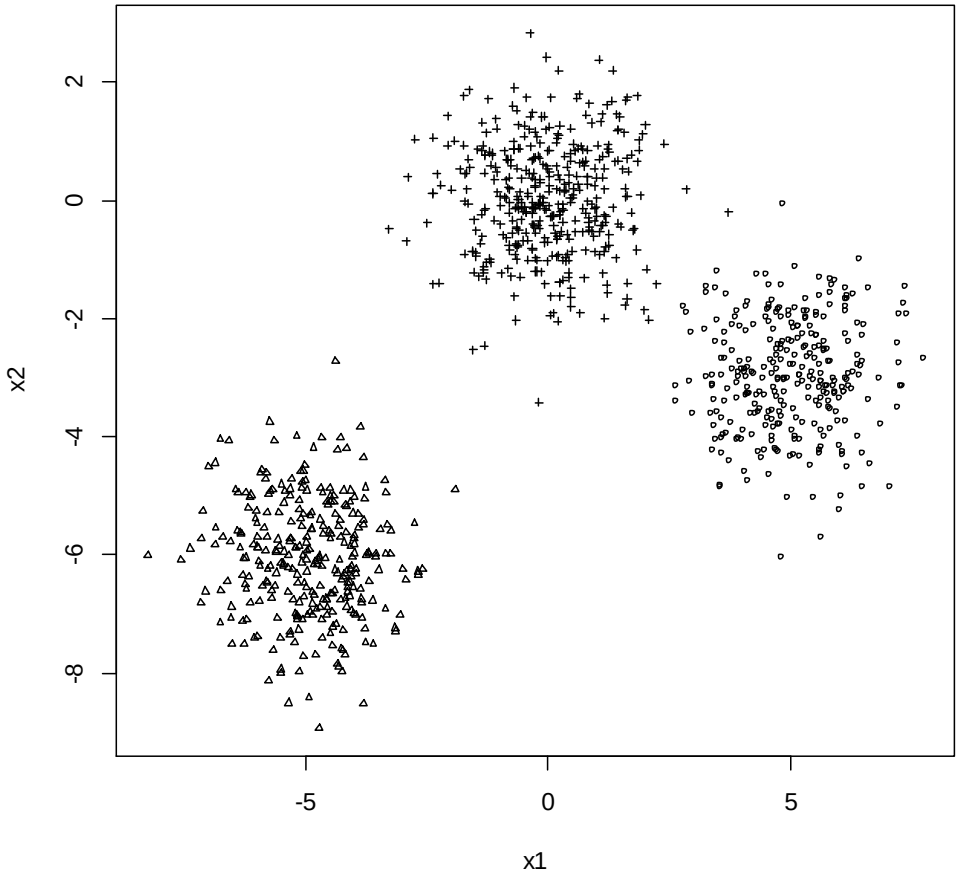
- Which genes (variables) are used.
- Which samples are used.
- Which distance measure is used.
- Which algorithm is applied.
- How to decide the number of clusters *K.*

Example 1



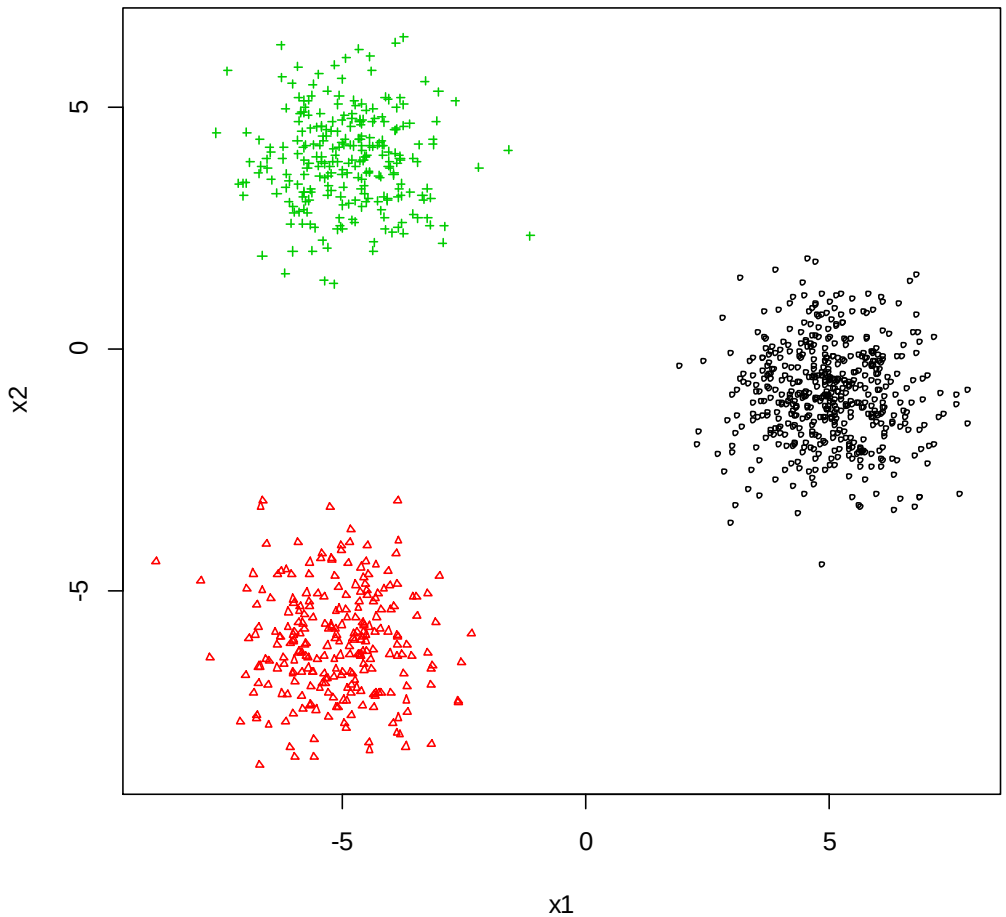
3 clusters of observations in a data matrix with one variable.

Example 2



3 clusters of observations in a data matrix with two variables.

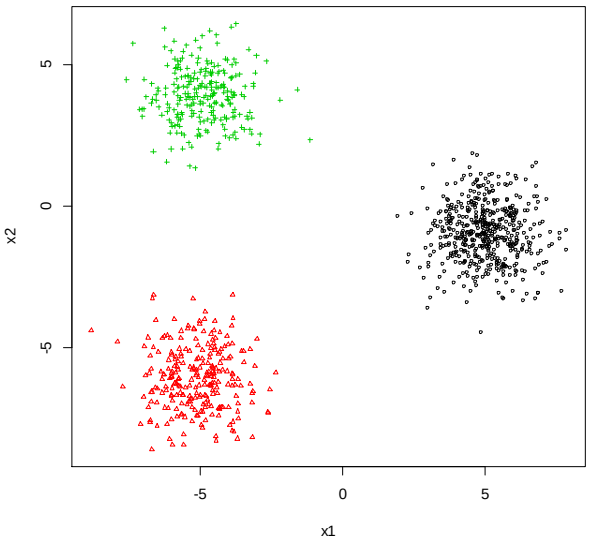
Example 3



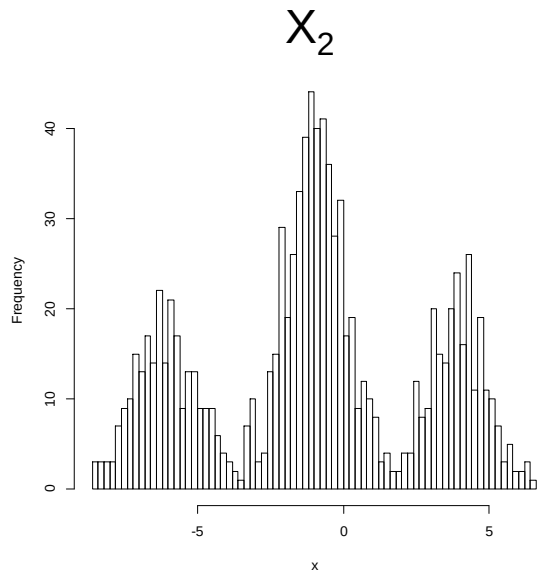
How many clusters of observations in a data matrix with two variables ?

X_1: two clusters.
X_2: three clusters.

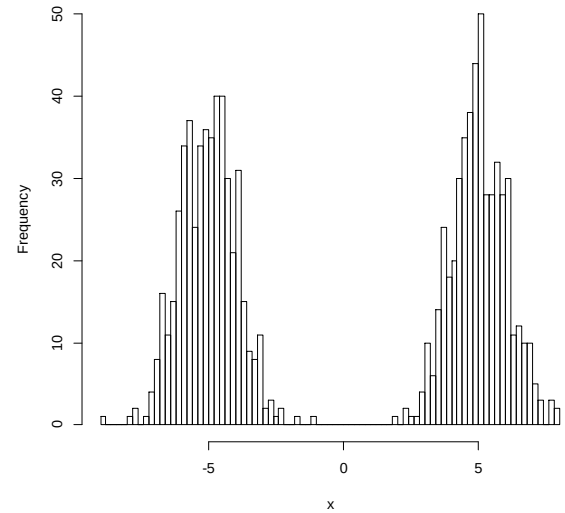
Example 3: how many clusters ?



X_1



X_2



X_1

Two clusters can be identified on the dimension of X_1 but three on the dimension of X_2.

Hierarchical clustering

Software

- In R:
  - `hclust()`
  - `dist()`
  - …..

Hierarchical methods

Hierarchical clustering methods produce a clustering tree or dendrogram.

They avoid specifying how many clusters are appropriate by providing a partition for each *k* obtained from cutting the tree at some level.

The tree can be built in two distinct ways

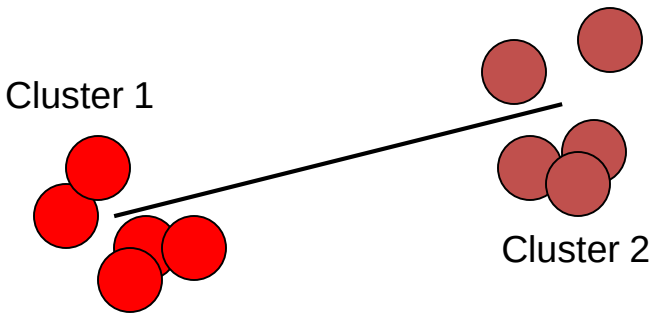
- bottom-up: agglomerative clustering;
- top-down: divisive clustering.

Agglomerative Methods

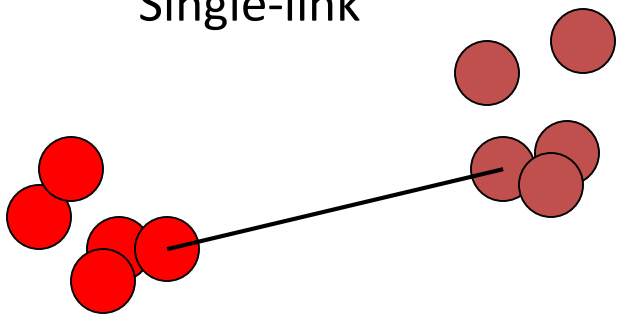
- Bottom-up: start with n clusters (each observation is a cluster).
- At each step, merge the two closest clusters using a measure of between-cluster dissimilarity which reflects the shape of the clusters.
- Examples of between-cluster dissimilarities:
  - Average: average of pairwise dissimilarities
  - Single-link: minimum of pairwise dissimilarities
  - Complete-link: maximum of pairwise dissimilarities

Simple examples of distance measures

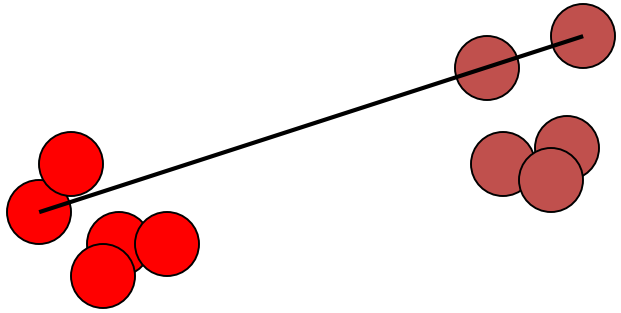
Distance between centroids



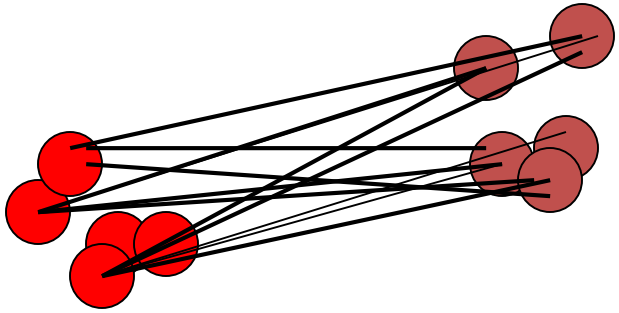
Single-link



Complete-link

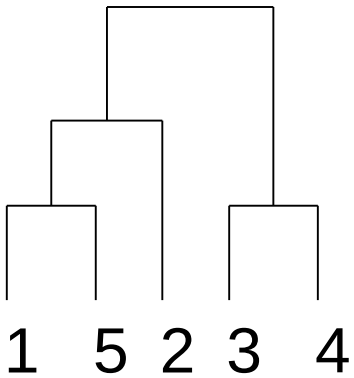


Mean-link

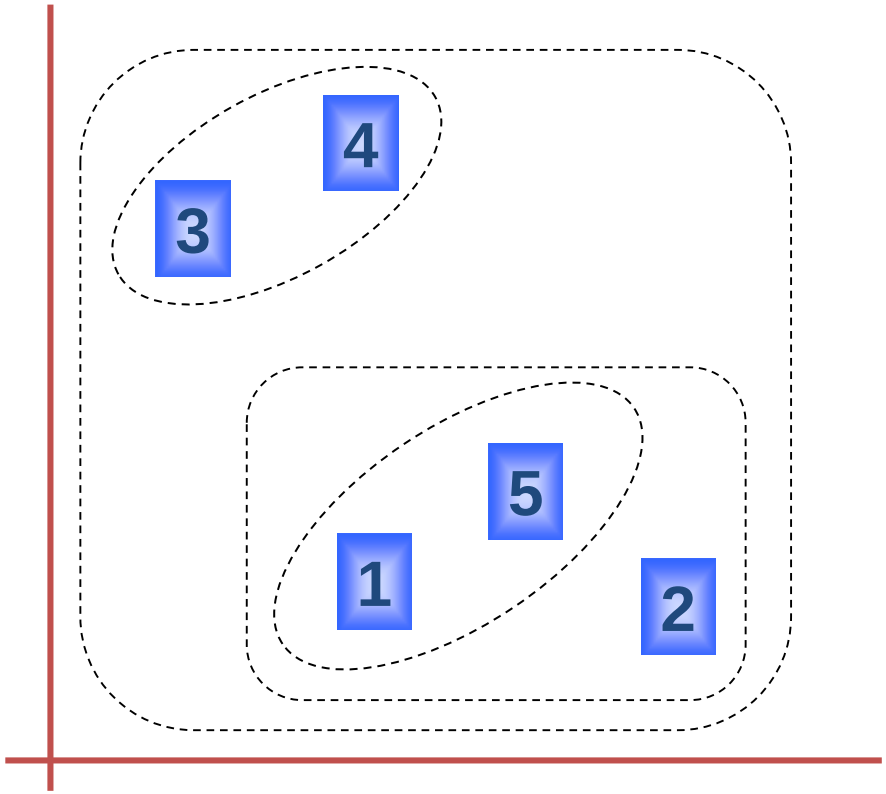


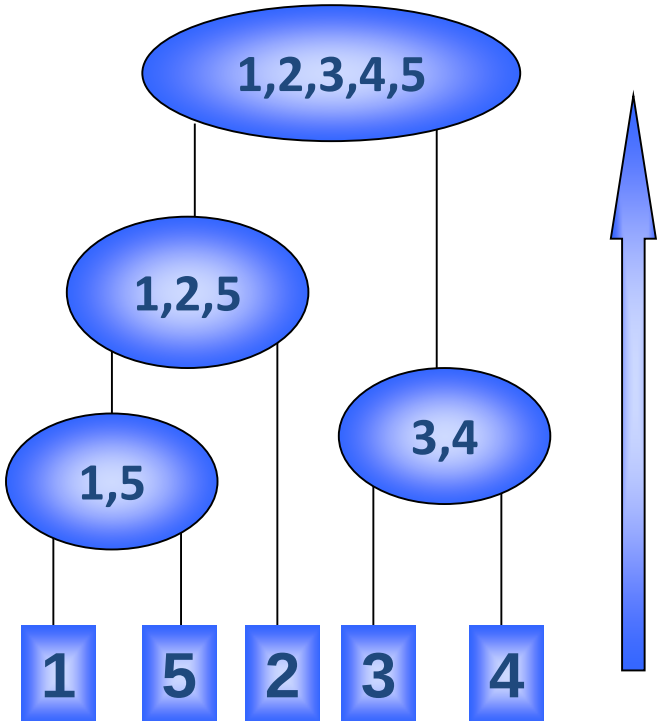
See also in slide 44-46

Bottom up clustering



Agglomerative

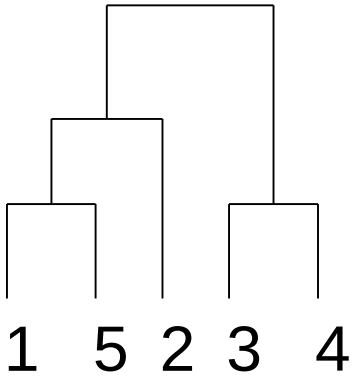


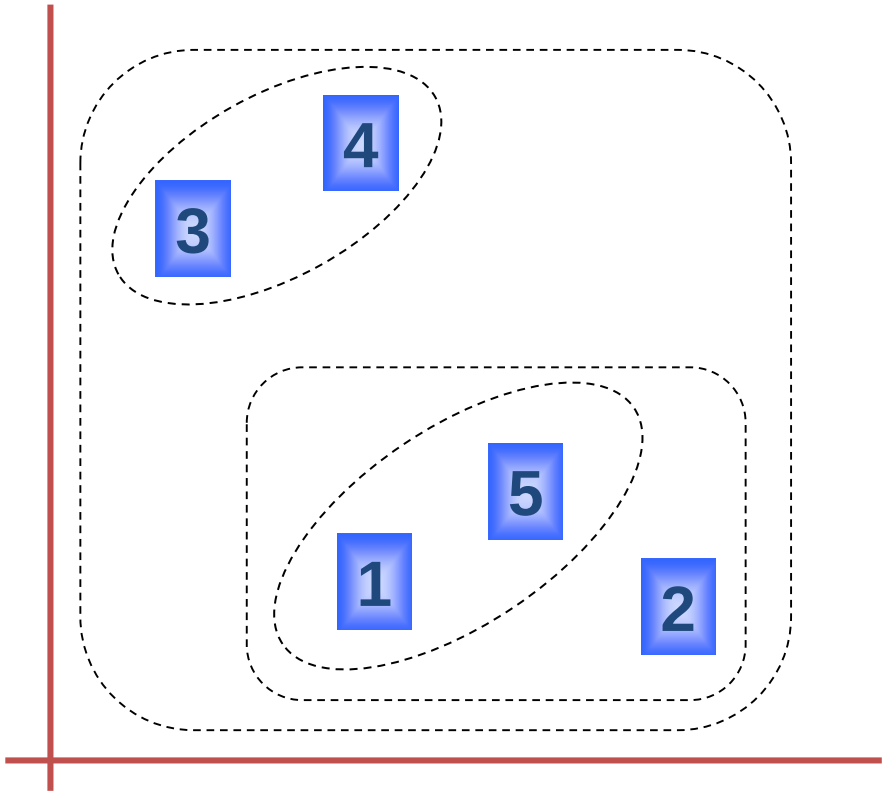


Divisive methods

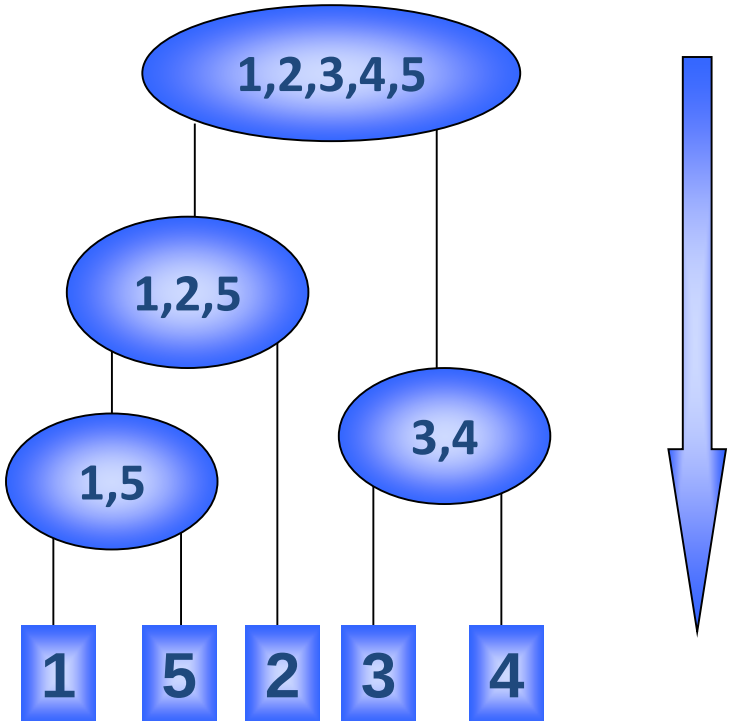
- Top-down : start with only one cluster (all observations in one cluster).
- At each step, split clusters into two parts.
- Split to give greatest distance between two new clusters.

Top down cluttering





Divisive



Two main classes of measures of dissimilarity

- Correlation (similarity)
- Distance (dissimilarity)
  - Manhattan
  - Euclidean
  - Many more ….

Dissimilarity

- Define an inter-sample distance and calculate the distance between each pair of samples.
- Similarity:
  - s_{ij} indicates the strength of relationship between two objects i and j
  - Usually $0 \leq s_{ij} \leq 1$
  - Correlation-based similarity ranges from –1 to 1
  - Associated with similarity measures s_{ij} bounded by 0 and 1 is a dissimilarity

    $d_{ij} = 1 - s_{ij}$

Distance

- Distance measures are statistics that statesquantitatively how dissimilar x_g and x_h are to each other.
- Euclidean Distance:

$$D_E\left(x_i, x_j\right) = \sqrt{\sum_{k=1}^{P}\left(x_{ik} - x_{jk}\right)^2}$$

Dissimilarity axioms:

$$D_E\left(x_i, x_j\right) > 0$$

$$D_E\left(x_i, x_j\right) = 0, \qquad if\ x_i = x_j$$

$$D_E\left(x_i, x_j\right) = D_E\left(x_j, x_i\right)$$

$D_E\left(x_i, x_j\right)$ increasing as the distance between x_i and x_j increase

Distance

- Manhattan distance

$$D_M\left(x_i, x_j\right) = \sum_{k=1}^{P} \left| x_{ik} - x_{jk} \right|$$

Pearson's Correlation

$$R_M\left(x_i, x_j\right) = \frac{\sum_{k=i}^{P}\left(x_{ik} - \bar{x}_{i.}\right)\left(x_{jk} - \bar{x}_{j.}\right)}{\sqrt{\sum_{k=i}^{P}\left(x_{ik} - \bar{x}_{i.}\right)^2 \sum_{k=1}^{P}\left(x_{jk} - \bar{x}_{j.}\right)^2}}$$

Similarity measures

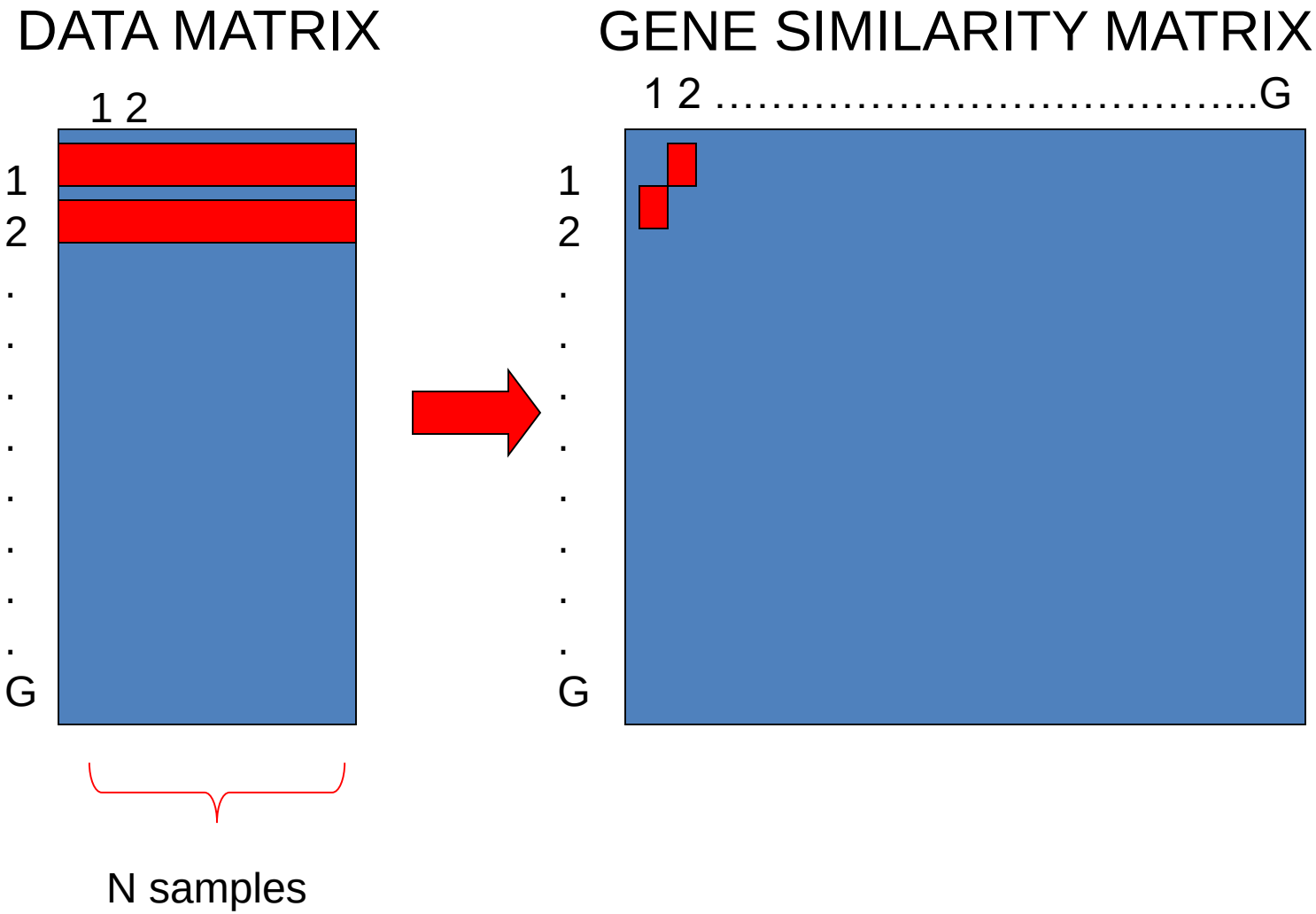
$$-1 \le R\left(x_i, x_j\right) \le 1$$

$$R\left(x_i, x_j\right) = 1, \quad if\ x_i = x_j$$

$$R\left(x_i, x_j\right) = -1, \quad if\ x_i = -x_j$$

$R\left(x_i, x_j\right) = 0$ if x_i and x_j are not associated

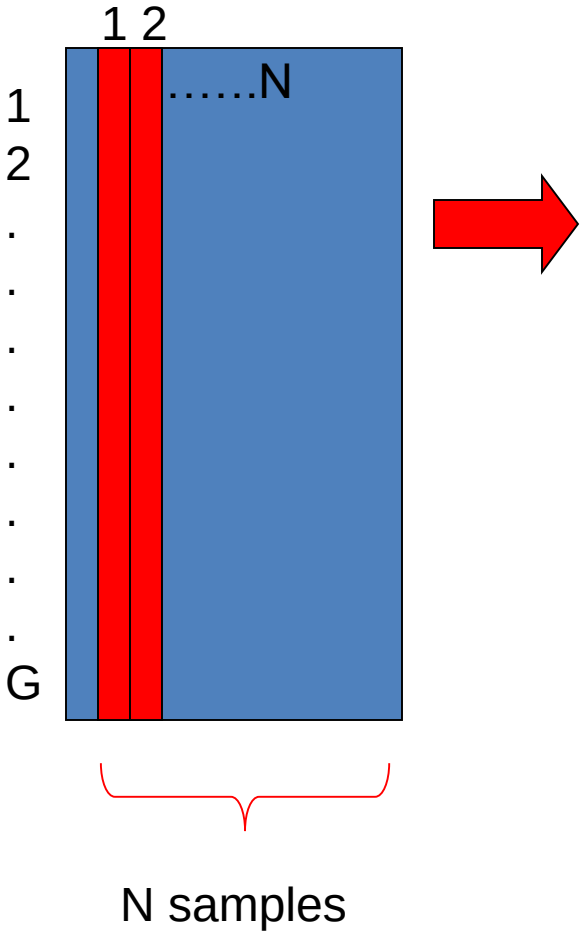
The similarity/distance matrices



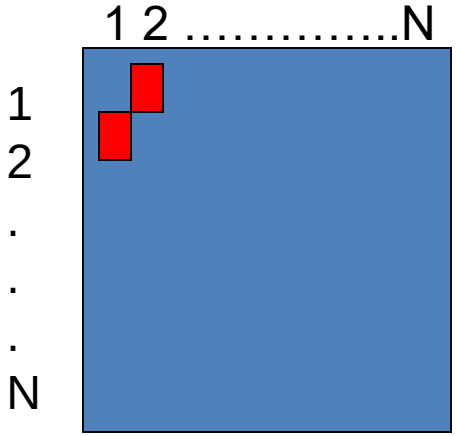
Row's similarity

The similarity/distance matrices

DATA MATRIX

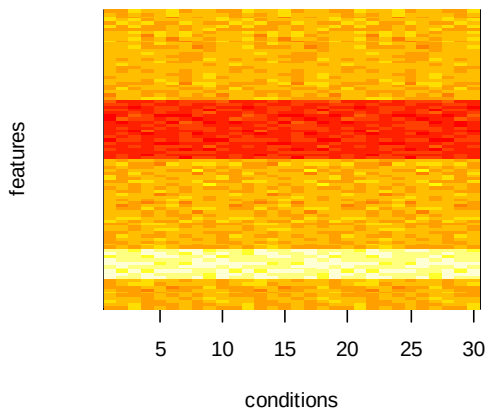


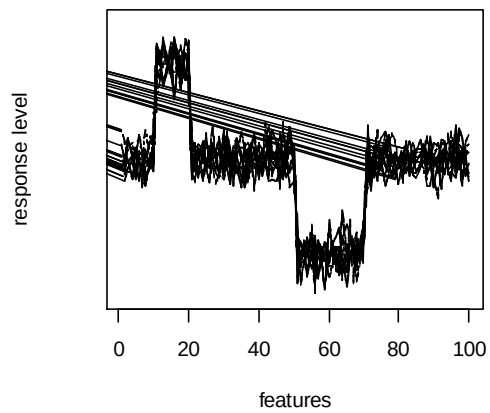
SAMPLE SIMILARITY MATRIX

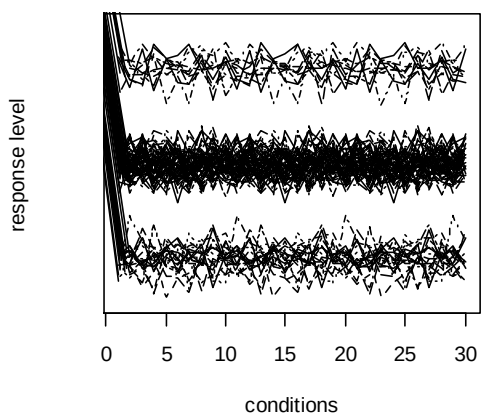


Column's similarity

Example

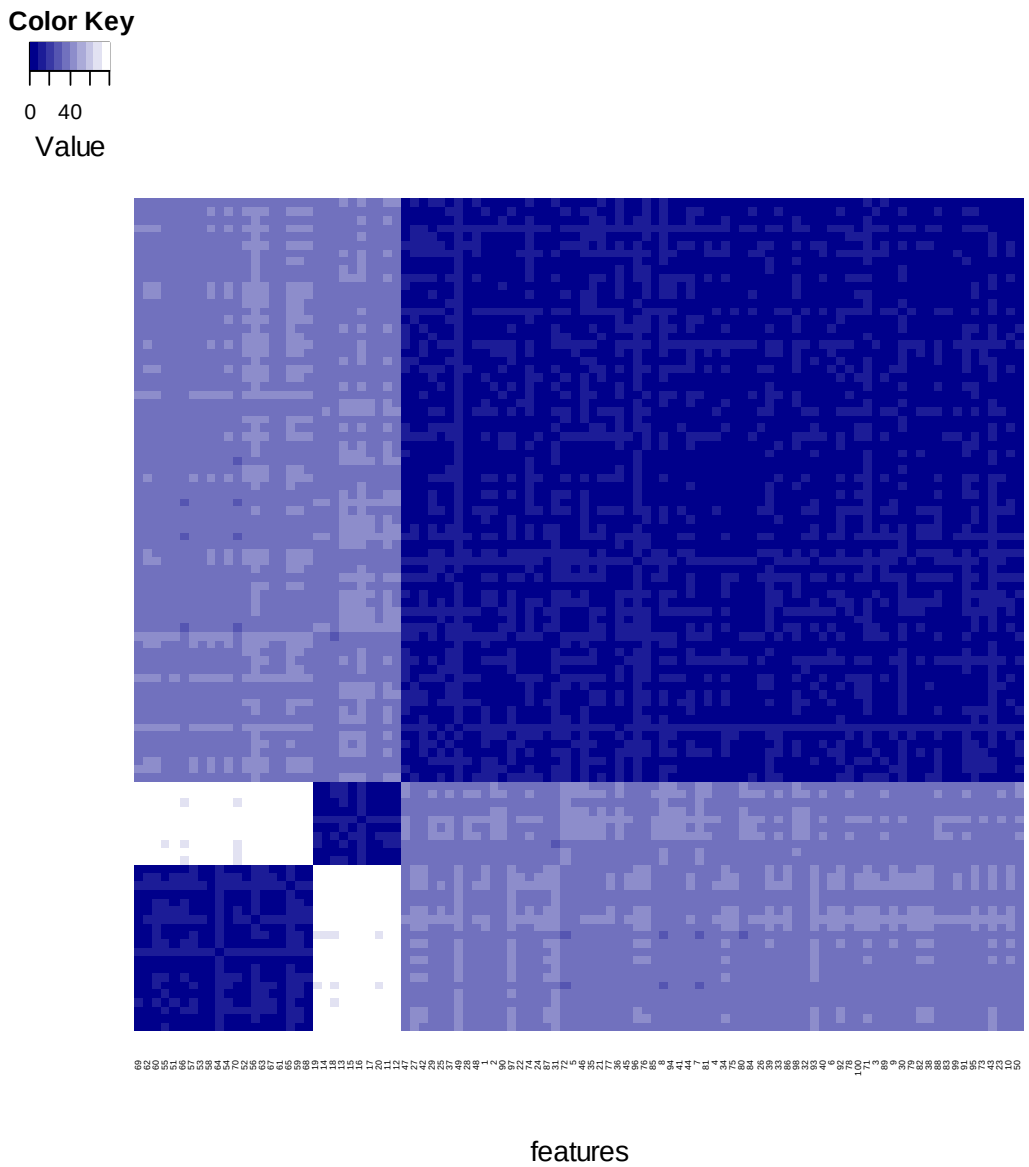






- 100 rows (observations).
- 30 columns (samples).
- Three clusters (of observations).

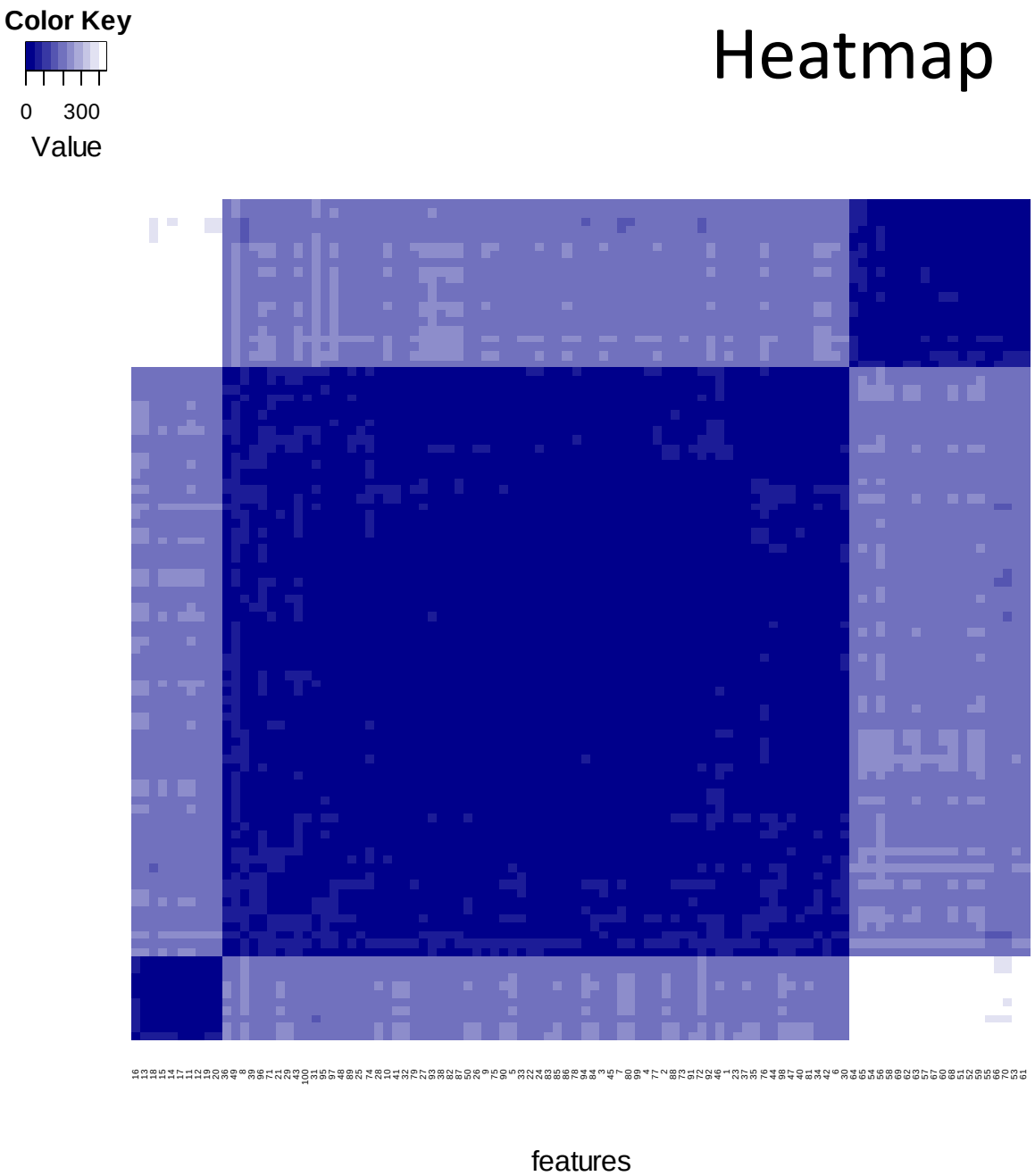
Heatmap



Euclidean distance. Three clusters (of observations).

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

Heatmap



Manhattan distance. Three clusters (of observations).

$$D_M(x_i, x_j) = \sum_{k=1}^{P} \left| x_{ik} - x_{jk} \right|$$

Hierarchical Clustering

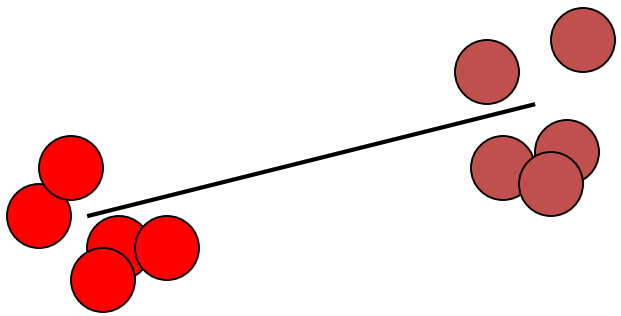
- Bottom-up clustering (also known as agglomerative hierarchical clustering)
- Algorithms are initiated with each gene situated in its own cluster.
- At the next and subsequent steps, the closest pair of clusters is agglomerated (i.e., combined).
- In principal, the process can be continued until all the data falls into one cluster.

Hierarchical Clustering

- Whenever two clusters are agglomerated, the distances between the new cluster and all the other clusters are recalculated.
- Different hierarchical clustering schemes calculate the distance between two clusters differently.

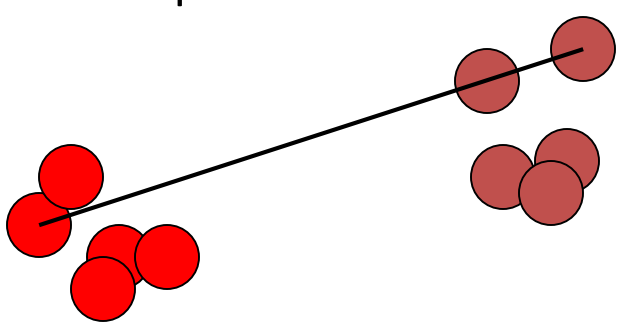
Simple examples of distance measures

Distance between centroids



In centroid clustering, the distance between two clusters is taken to be the dissimilarity measure between the cluster centers.

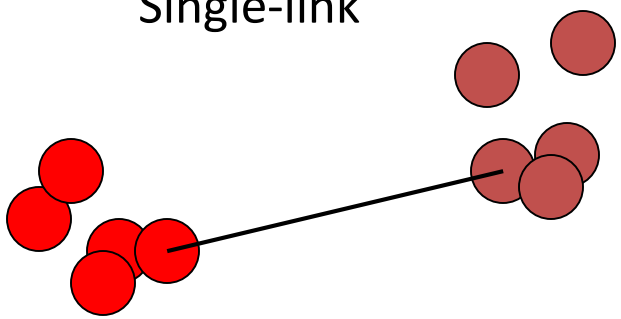
Complete-link



In *complete linkage hierarchical clustering* (or *farthest neighbor clustering*),the distance between two clusters is taken to be the largest dissimilarity measure between any two members in different clusters.

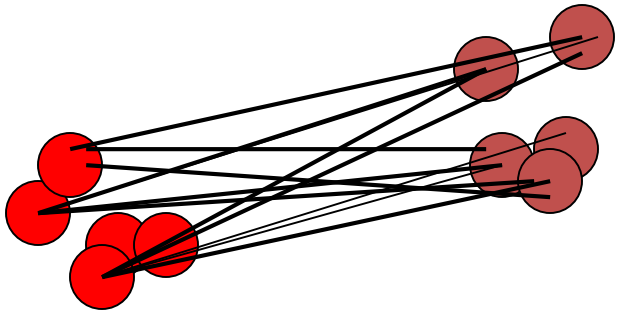
Simple examples of distance measures

Single-link



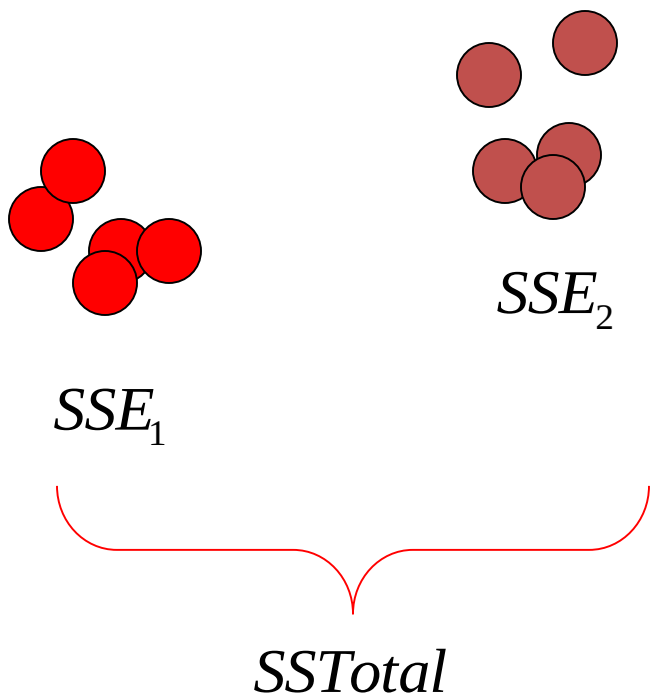
In single linkage hierarchical clustering (or nearest neighbor clustering), the distance between two clusters is taken to be the smallest dissimilarity measure between any two members in different clusters.

Mean-link



In average linkage hierarchical clustering, the distance between two clusters is taken to be the arithmetic mean of the dissimilarity measures between all pairs of members in different clusters.

Ward's Minimum-Variance Method



$$R^2 = \frac{SSTotal - (SSE_1 + SSE_2)}{SSTotal}$$

In Ward's clustering, the distance between two clusters is taken to be the sum of squares between clusters divided by the total sum of squares.

Equivalently, the change in R^2 when a cluster is split into two clusters, where the coefficient of determination, R^2, is the percent of the variation that can be explained by the clustering.

Ward's Minimum-Variance Method

- For the kth cluster, define the Error Sum of Squares as:
  - ESS_k = sum of squared deviations from the cluster centroid
- If there are C clusters, define the Total (within cluster) Error Sum of Squares as

$$ESS(C) = \sum_{k=1}^{C} ESS_k$$

Ward's Minimum-Variance Method

- Consider the union of every possible pair of clusters.
- Combine the 2 clusters whose combination results in the smallest increase in ESS.

Ward’s Minimum-Variance Method

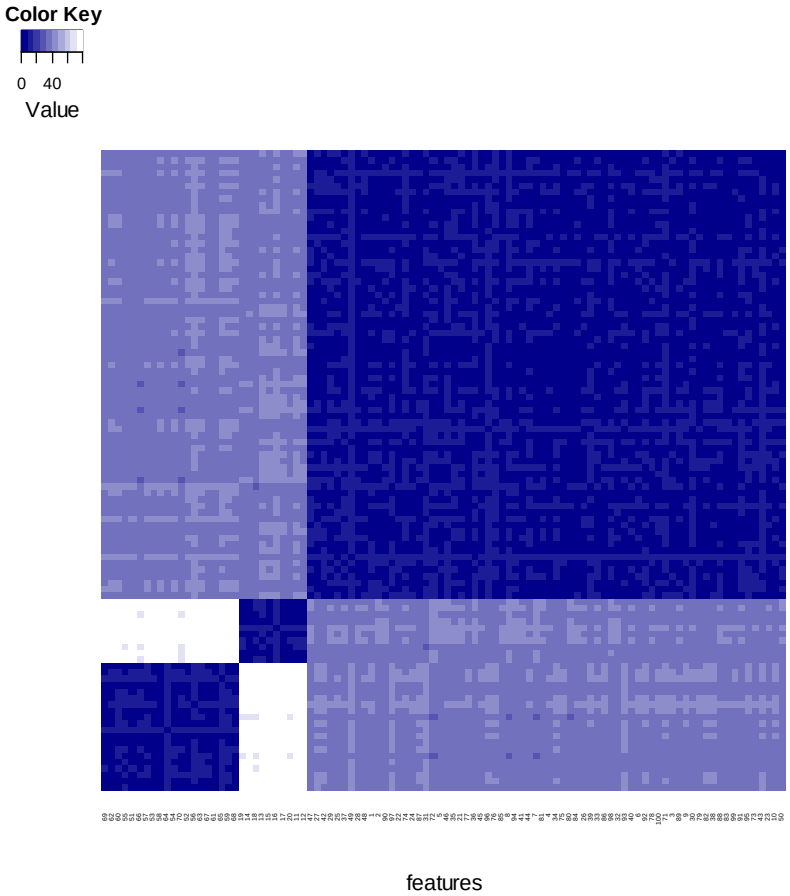
- In Ward's minimum-variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables.
- At each step, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous step.
- The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variance (squared correlations).

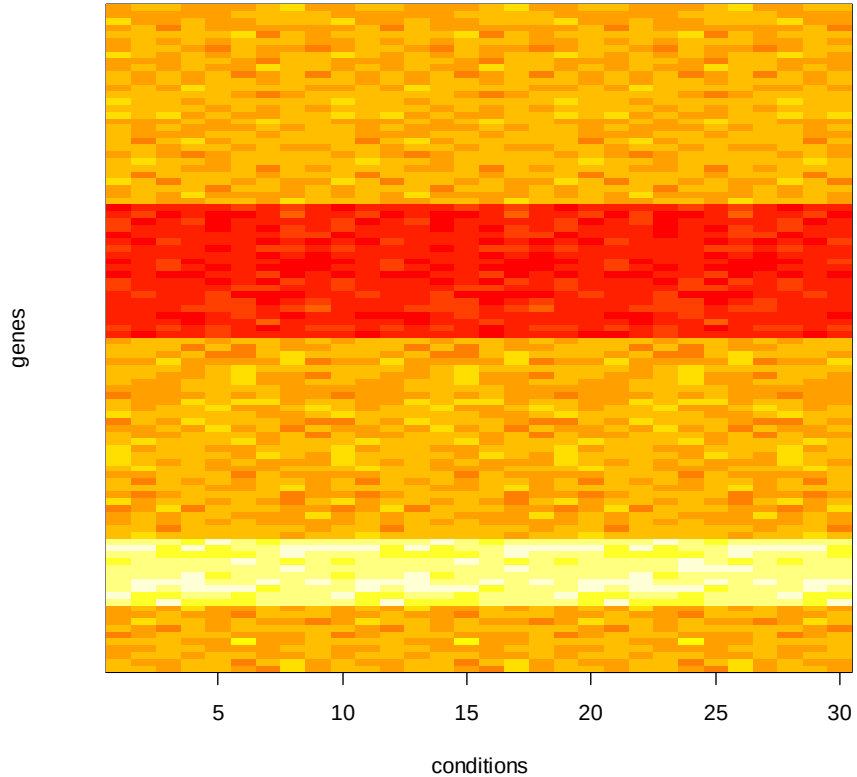
Ward’s Minimum-Variance Method

- Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions:
  - multivariate normal mixture
  - equal covariance matrices
  -

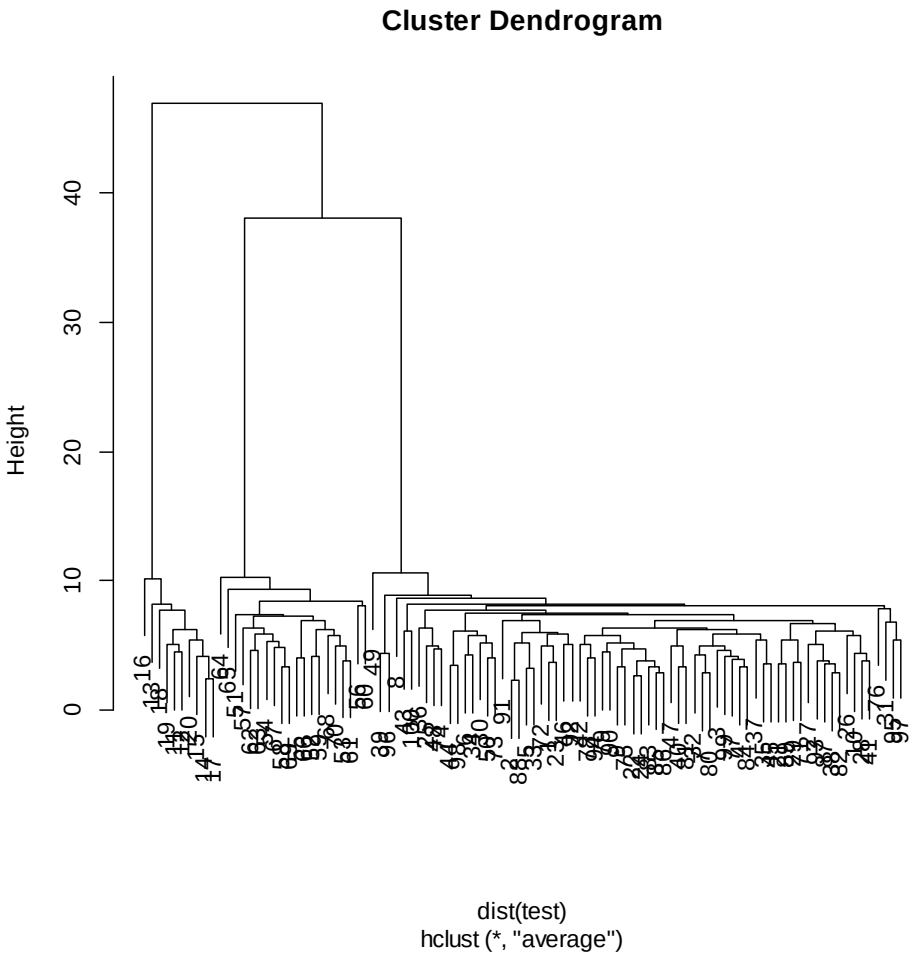
Example

100 variables, 30 samples, three clusters.



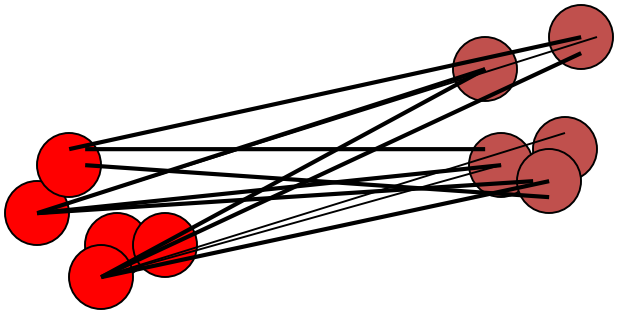


The average method




```
hc <- hclust(dist(test),"average")
par(mfrow=c(1,1))
plot(hc,cex=0.35)
```

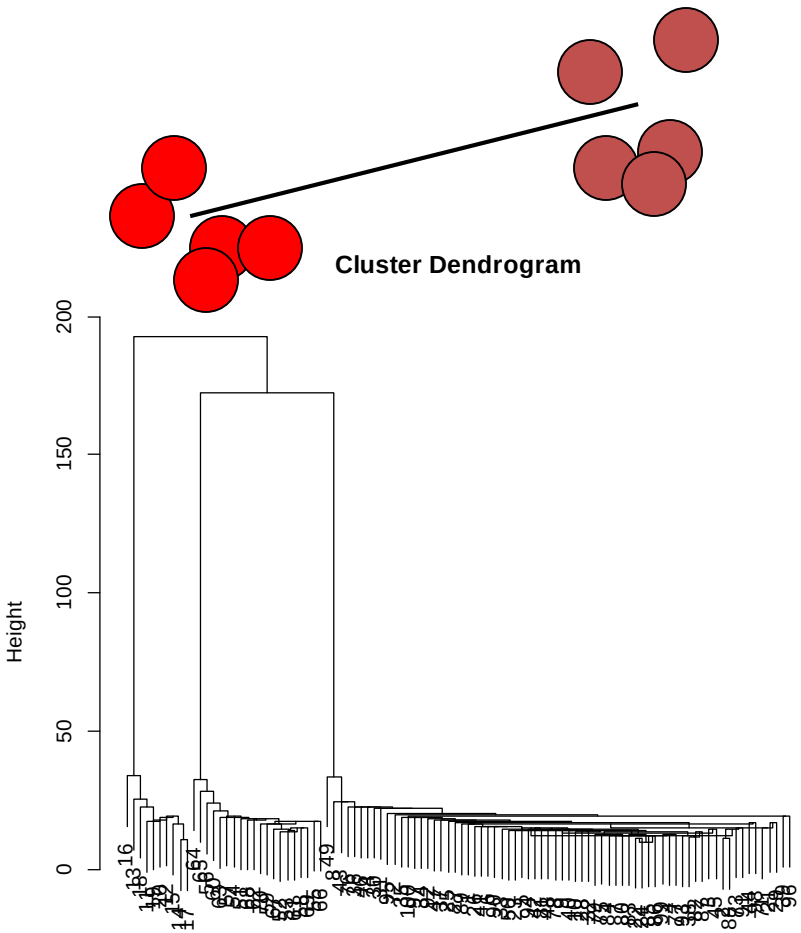
Mean-link



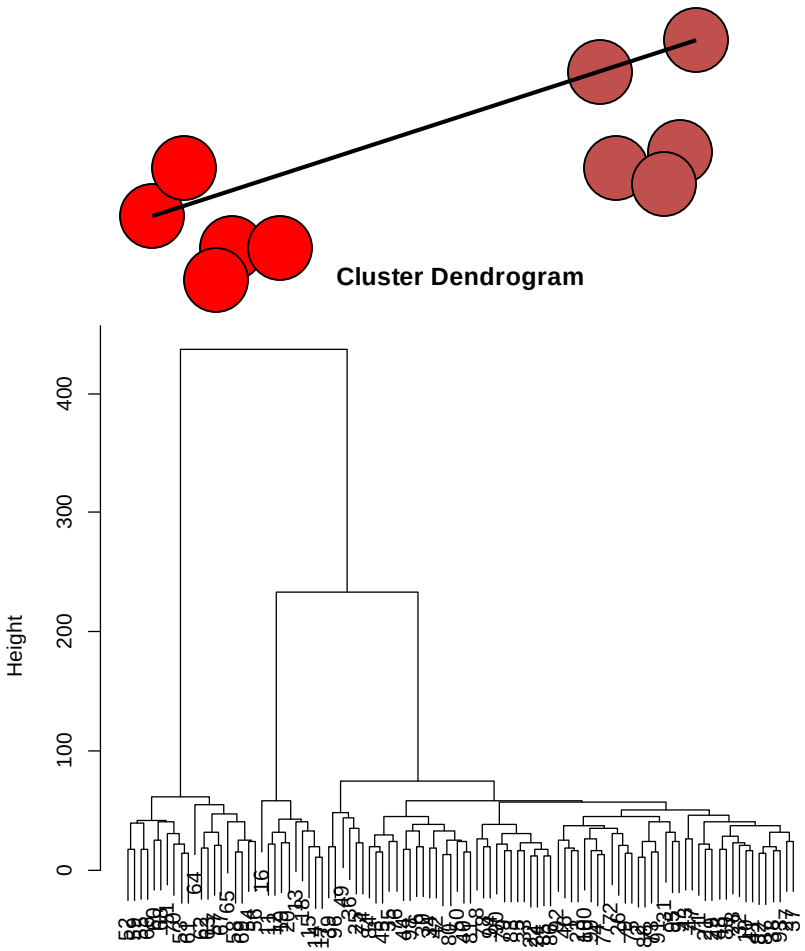
dist(test): The default is Euclidean

Centroid and complete linkage

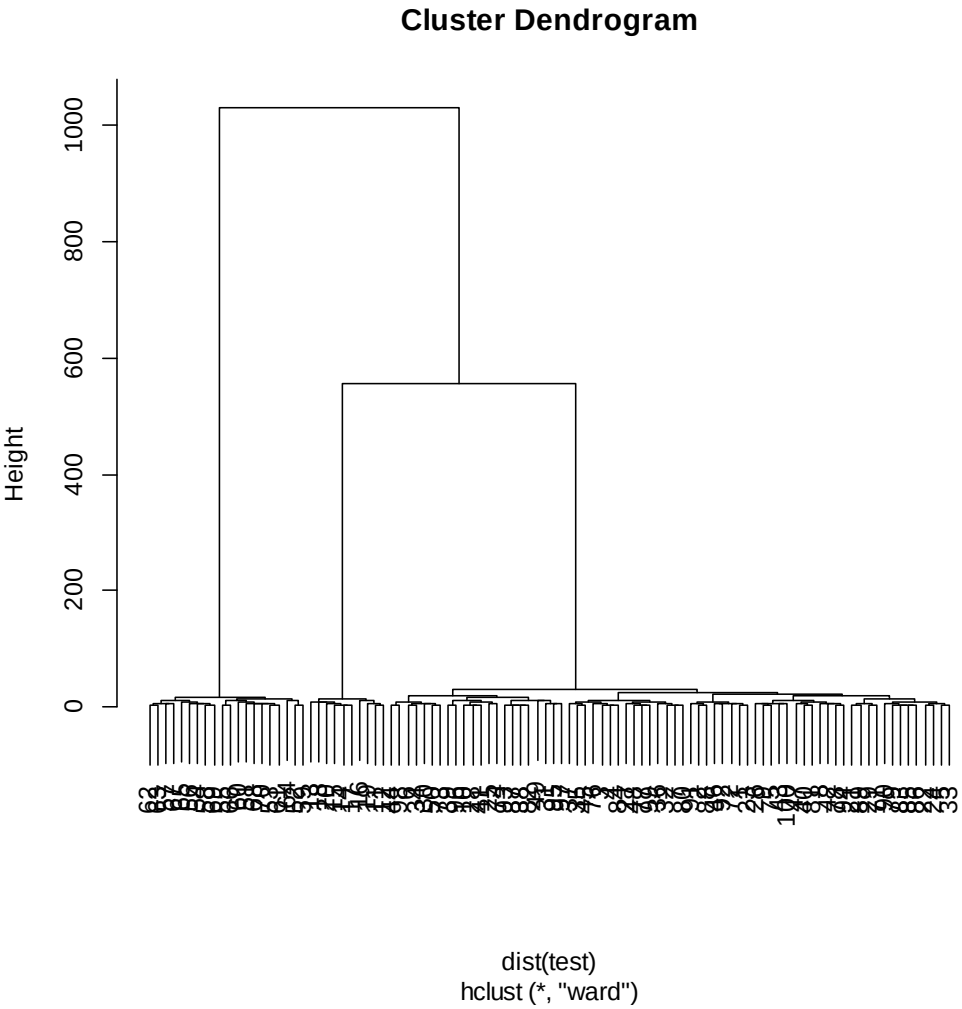
Centroid link



Complete-link



The ward method




```
hc <- hclust(dist(test),"ward")
par(mfrow=c(1,1))
plot(hc,cex=0.35)
```

Example: cluster analysis for the Golub data

Distinguishing two types of acute leukemia (AML vs. ALL)

- Golub, T.R. et al 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.
- http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi (near bottom of page)

Distinguishing AML vs. ALL

- 38 BM samples (27 childhood ALL, 11 adult AML) were hybridized to Affymetrix GeneChips
  - GeneChip included 6,817 human genes.
  - Affymetrix MAS 4.0 software was used to perform image analysis.
  - MAS 4.0 Average Difference expression summary method was applied to the probe level data to obtain probe set expression summaries.
  - Scaling factor was used to normalize the GeneChips.
  - Samples were required to meet quality control criteria.

Data structure

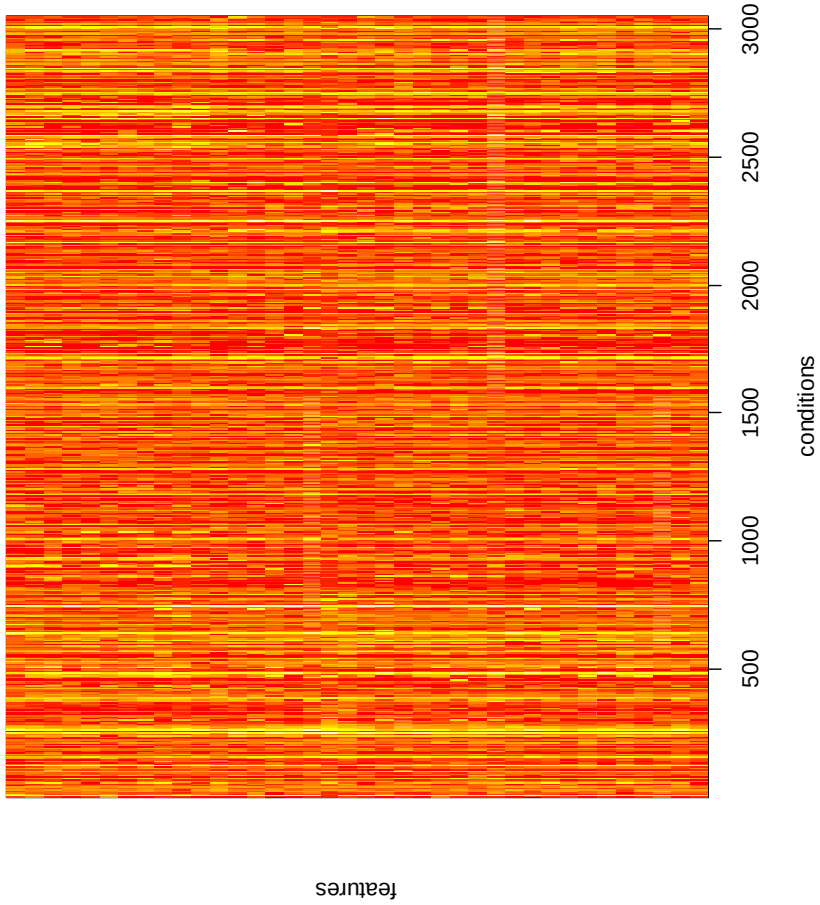
$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix}.$$

3051 features…

27 ALL 11 AML

Cluster samples based on similarity of the genes.

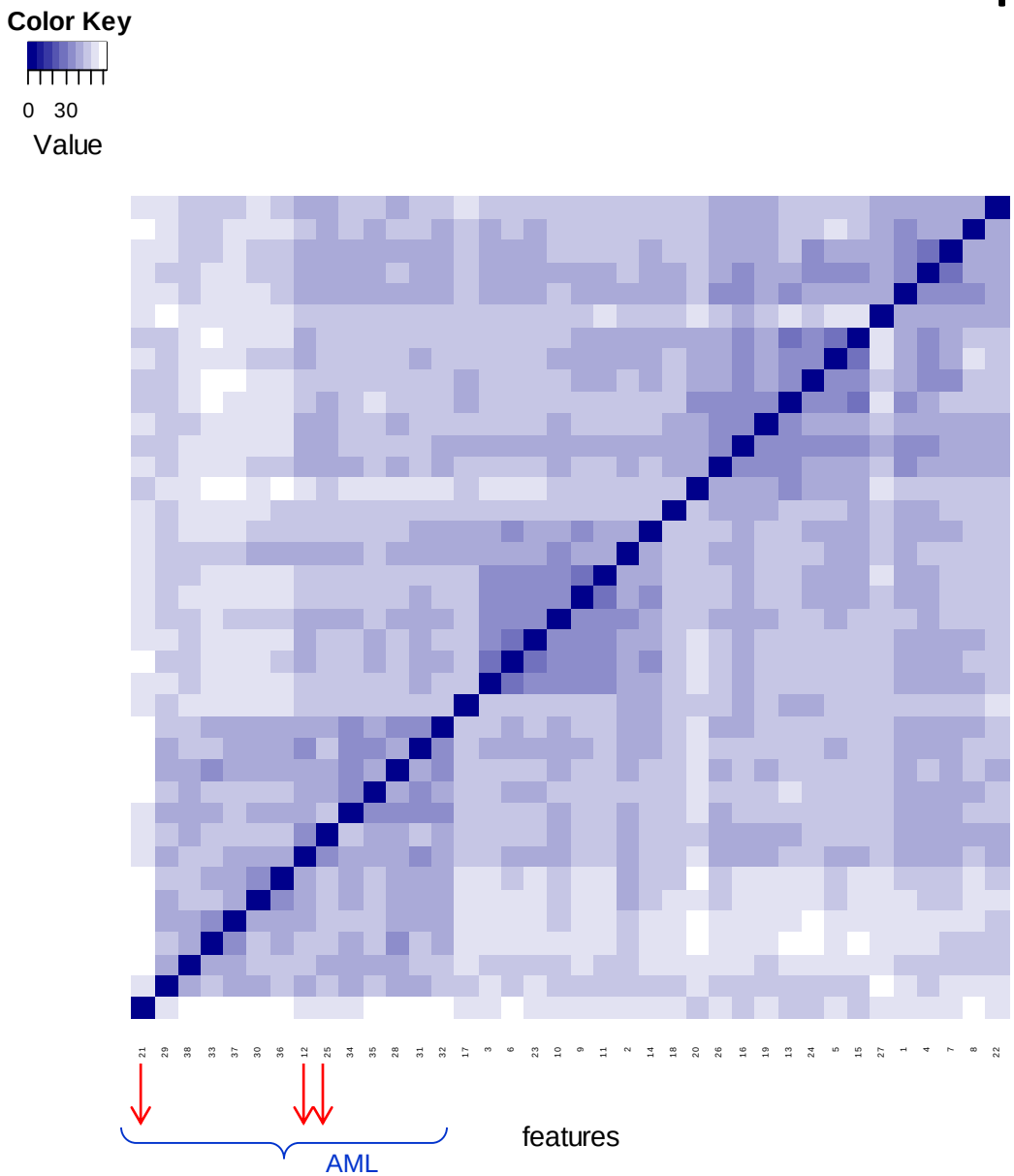
The golub data



3051 genes.
38 samples.

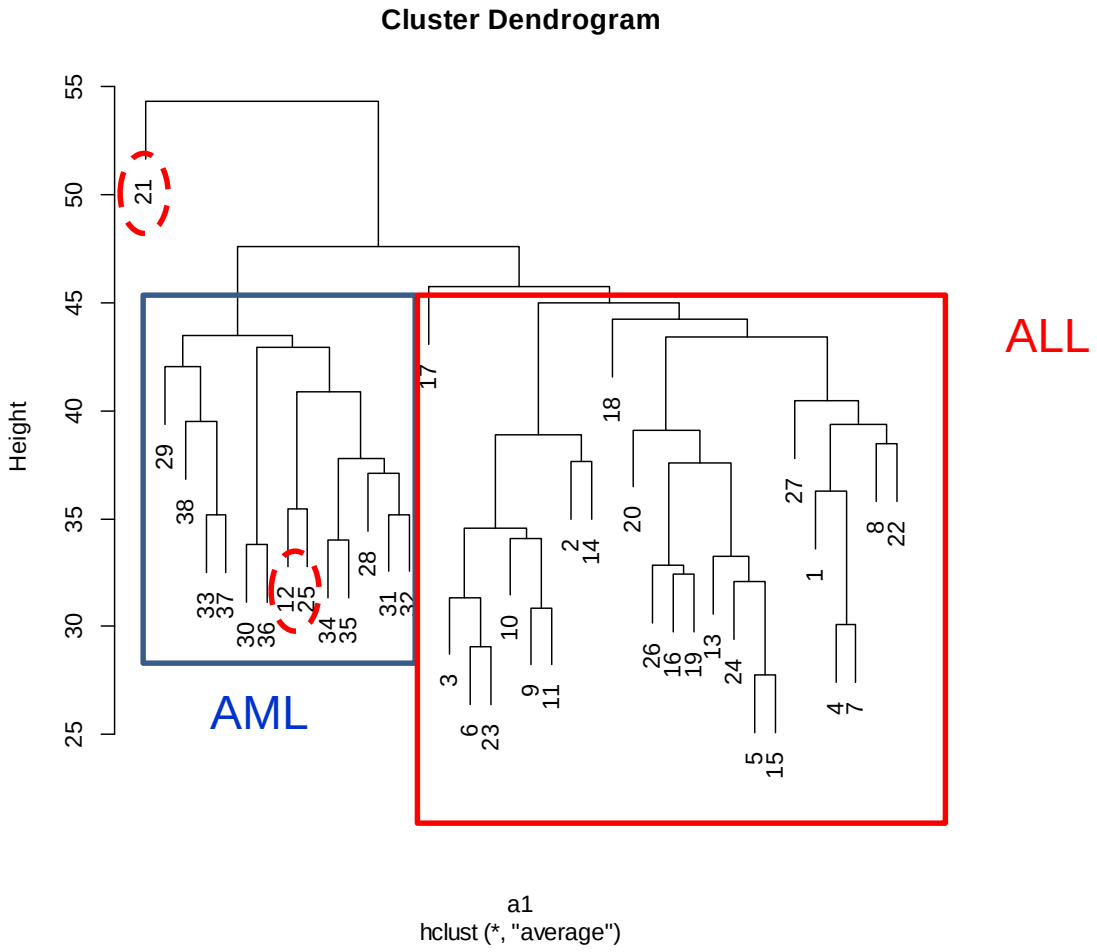
Which pattern we see here ?

Heatmap



Similarity matrix based on Euclidean distance.

Hierarchical clustering of the golub data



Gene and Sample Selection

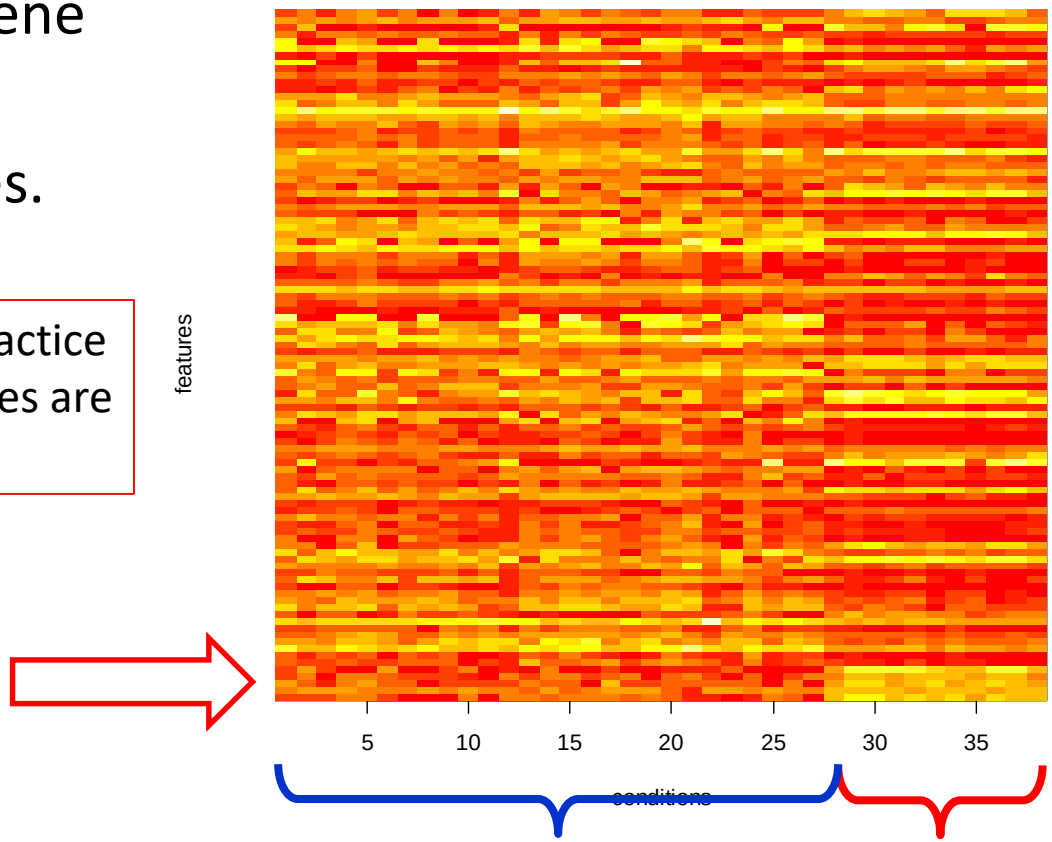
- Do you want all genes included?
- What to do about replicates from the same individual/tumor?
- Genes that contribute noise will affect your results.
- Including all genes: dendrogram can't all be seen at the same time.
- Perhaps screen the genes?

Illustration in the next example

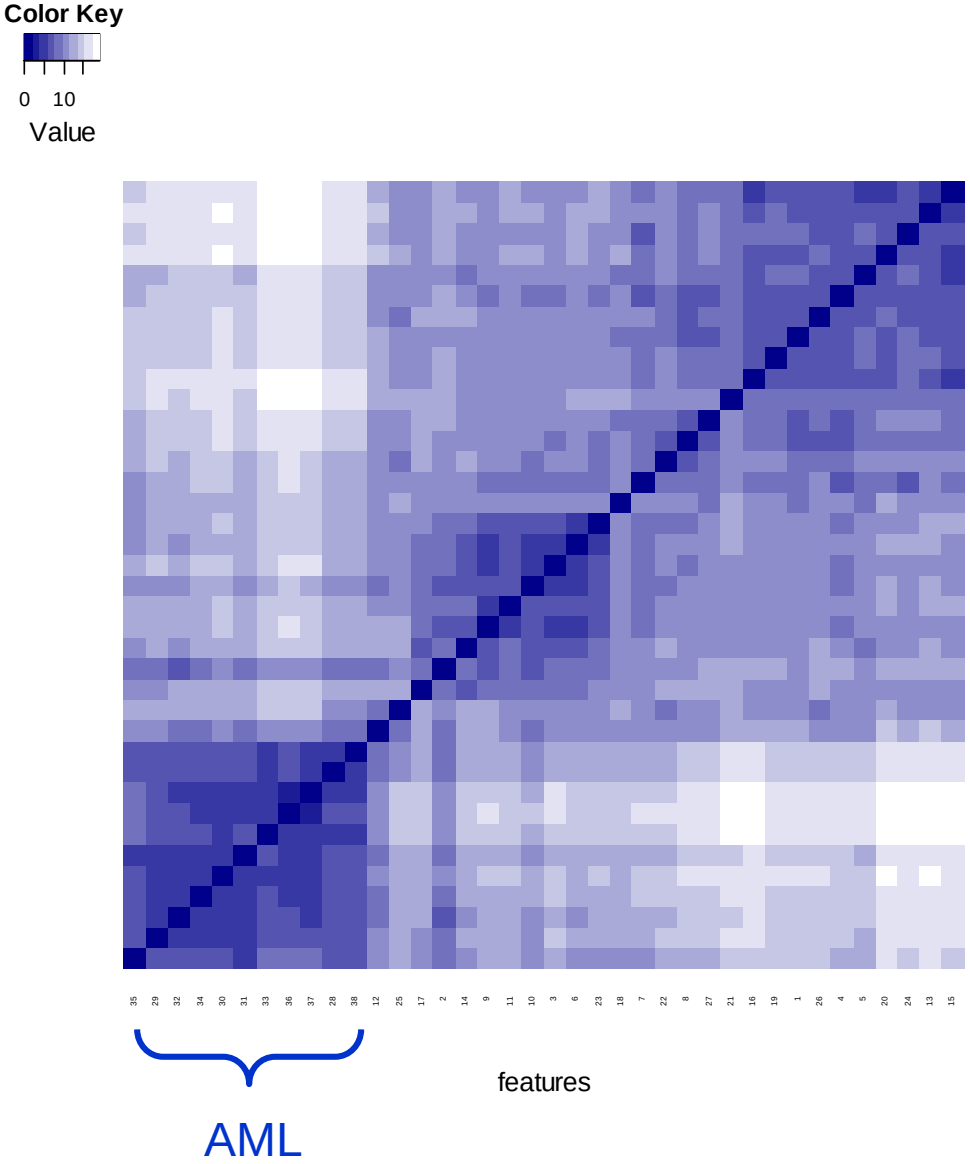
Reduced matrix

- Two sample t-test for gene selection.
- Selection: top 100 genes.

This is just an illustration, In practice cannot be done since the classes are unknown !!!

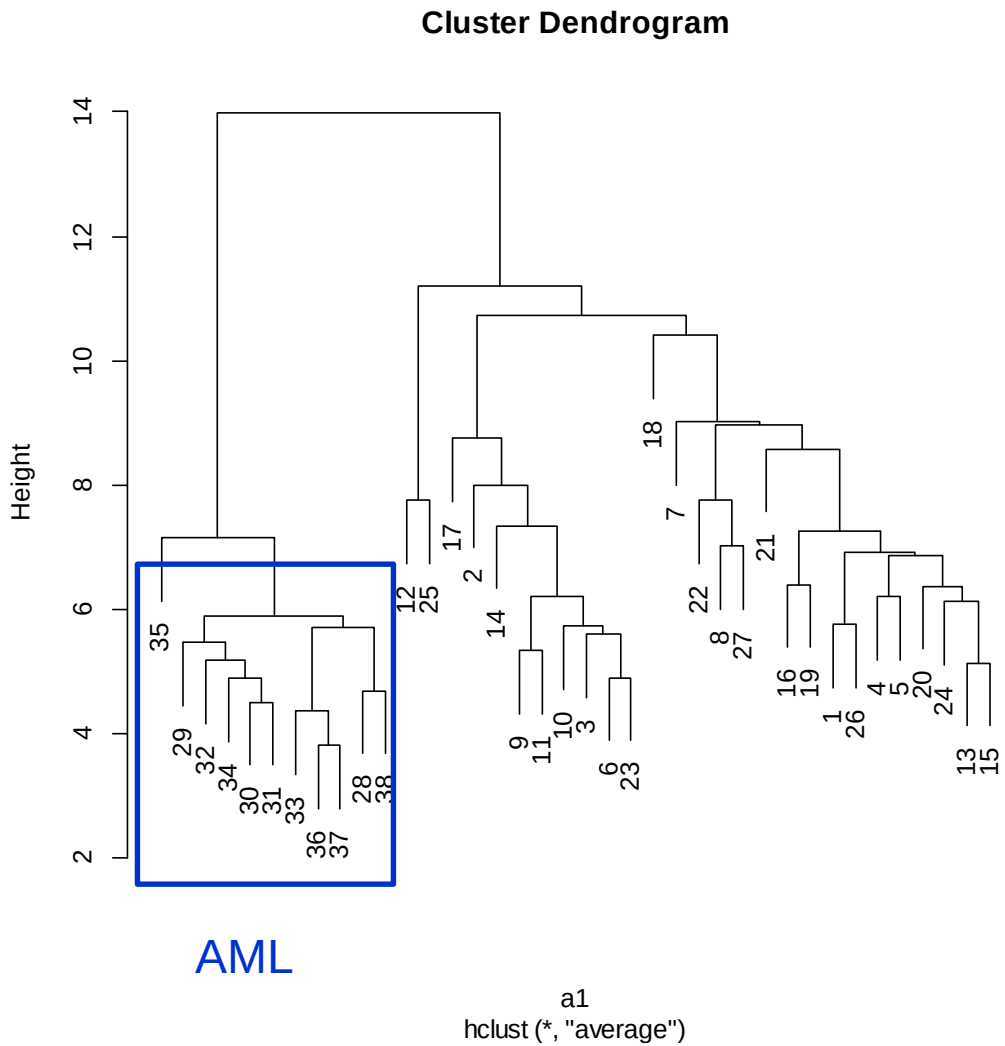


Heatmap

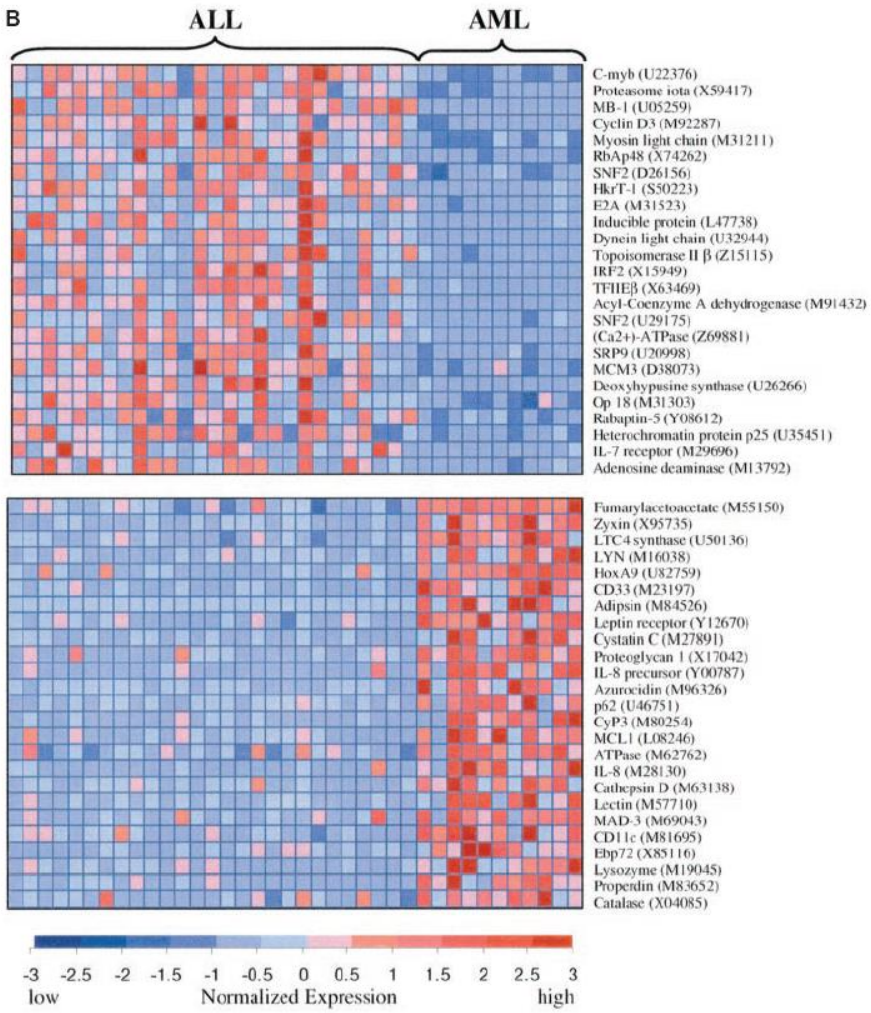


Clear two clusters ?

Hierarchical clustering for selected genes (I)

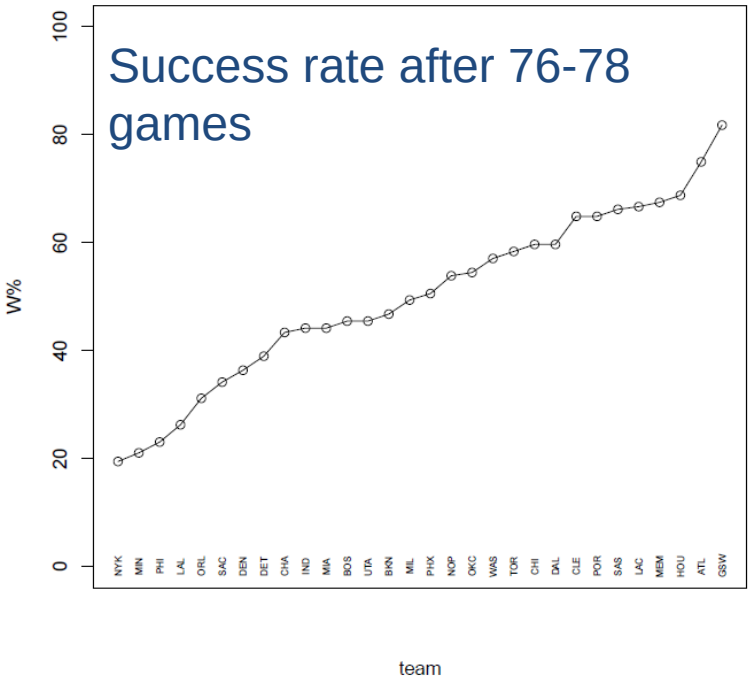


Hierarchical clustering for selected genes (II)



Example: cluster analysis for the NBA data

The NBA data (regular season of 2014/2015)



$$\%W = \frac{\#\,games \quad won}{\#\,games}$$

- 30 teams
- Regular season : 82 games per team.
- 16 teams go to the play-offs at the end of the regular season.
- Performance Statistics (teams and individuals) is well developed

Data structure

A team

A *30 X 12* matrix:

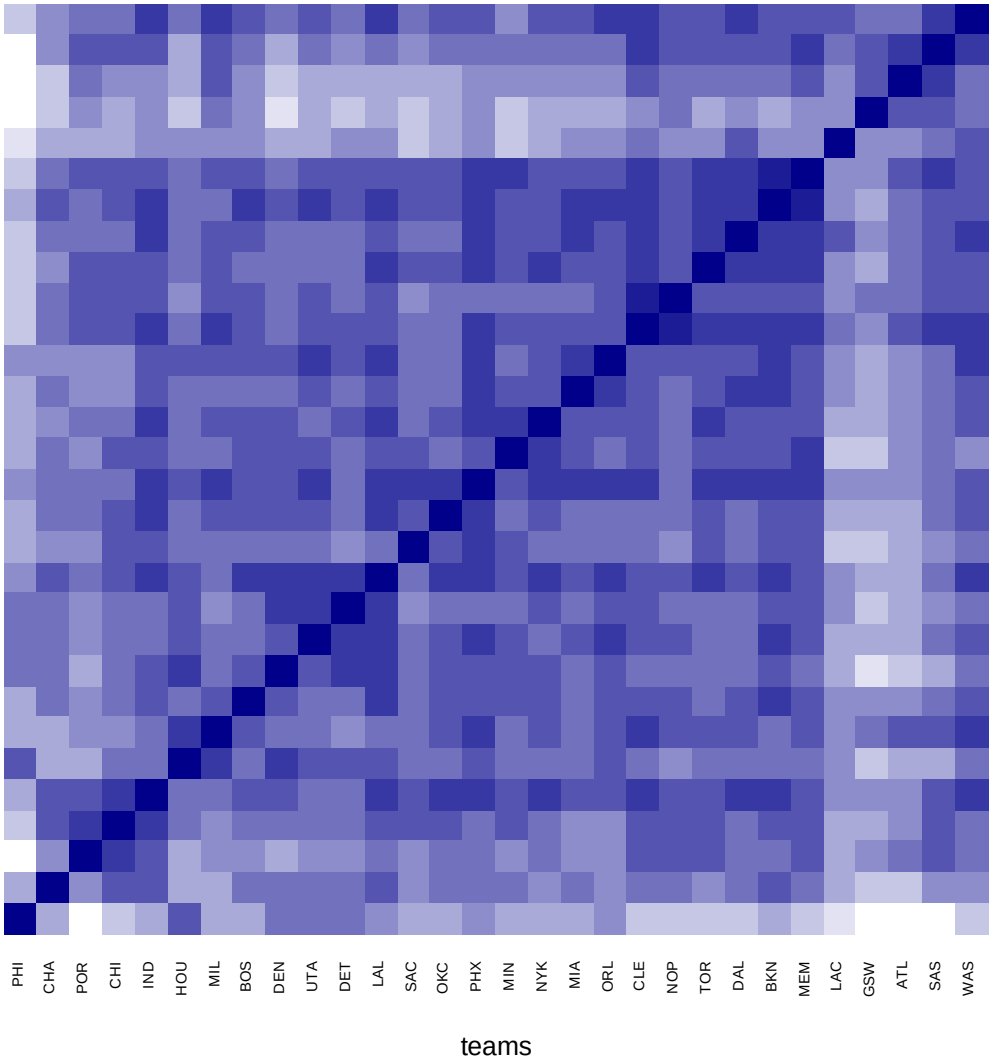
1. 2-pt & 3-pt Successful
2. 2-pt & 3-pt Unsuccessful
3. Free Throw Successful & Unsuccessful
4. Defensive & Offensive Rebounds
5. Assists
6. Turnovers
7. Steals
8. Dunks
9. Blocks Committed / Received
10. Fouls Committed / Received
11. …
12. …

$$\begin{bmatrix} x_{1,1} & x_{21} & \cdot & \cdot & x_{12,1} \\ x_{1,2} & x_{22} & \cdot & \cdot & x_{12,2} \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ x_{1,30} & x_{2,88} & \cdot & \cdot & x_{12,30} \end{bmatrix}$$

30 NBA teams

12 performance indicators

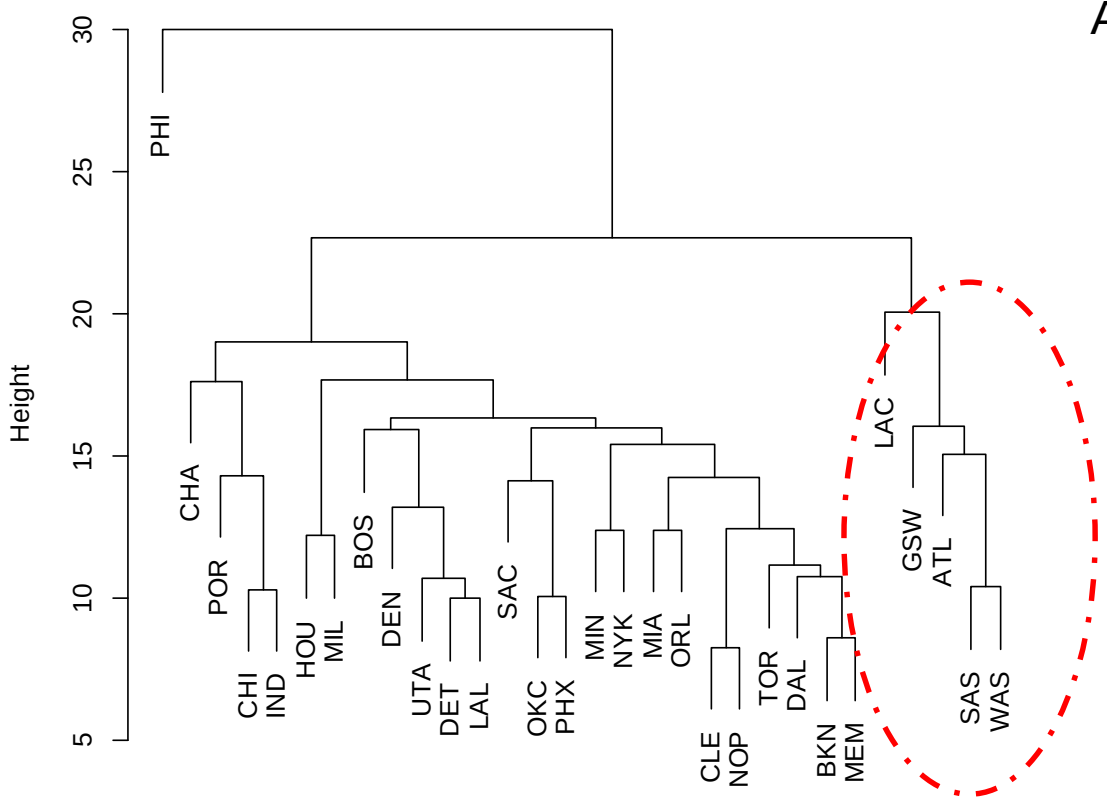
Similarity matrix



$$D_M\left(x_i, x_j\right) = \sum_{k=1}^{P} \left| x_{ik} - x_{jk} \right|$$

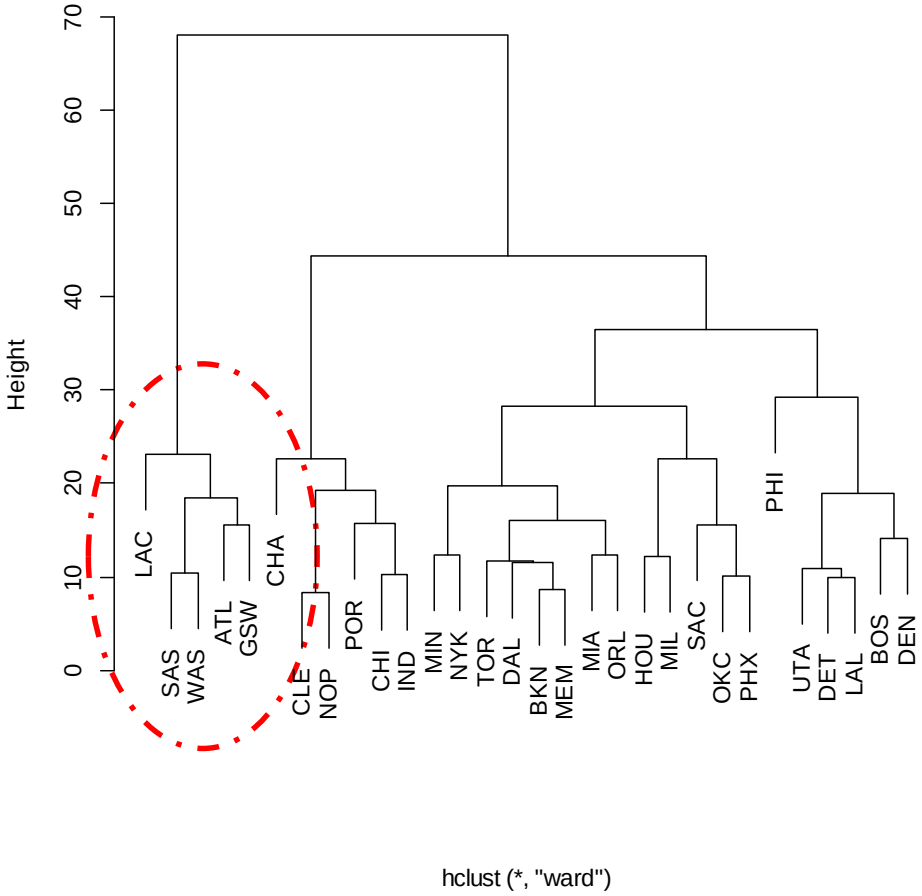
Clustering

Average linkage



hclust (*, "average")

Clustering



ward linkage

Analysis of standardized matrix

- Data matrix:
  - 12 variables, not all in the same scale.

$$\begin{bmatrix} x_{1,1} & x_{21} & . & . & x_{12,1} \\ x_{1,2} & x_{22} & . & . & x_{12,2} \\ . & . & . & . & \\ . & . & . & . & \\ x_{1,30} & x_{2,30} & . & . & x_{12,30} \end{bmatrix} \Rightarrow \begin{bmatrix} z_{1,1} & z_{21} & . & . & z_{12,1} \\ z_{1,2} & z_{22} & . & . & z_{12,2} \\ . & . & . & . & \\ . & . & . & . & \\ z_{1,30} & z_{2,30} & . & . & z_{12,30} \end{bmatrix}$$

Standardized variables:

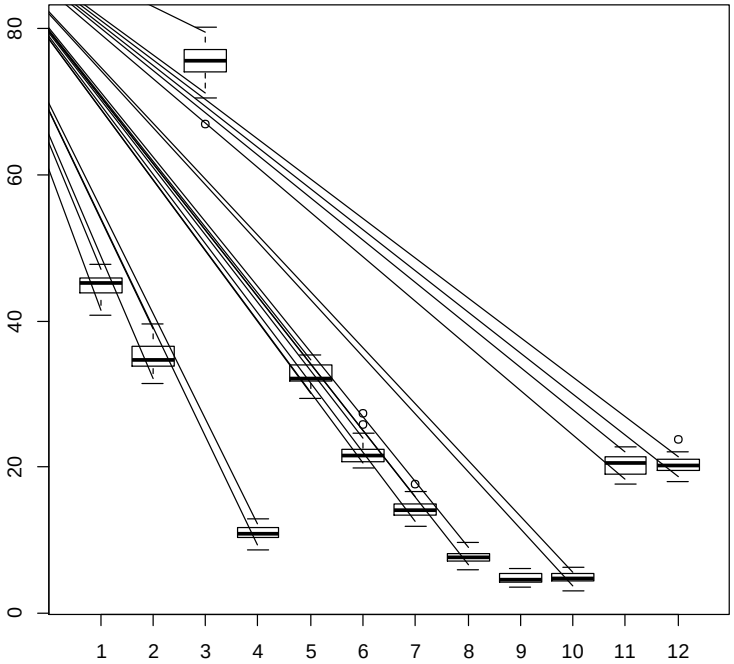
$$z_i = \frac{x_{ij} - \bar{x}_i}{\sqrt{\mathrm{var}(\bar{x}_i)}}$$

$$\bar{z}_i = 0$$

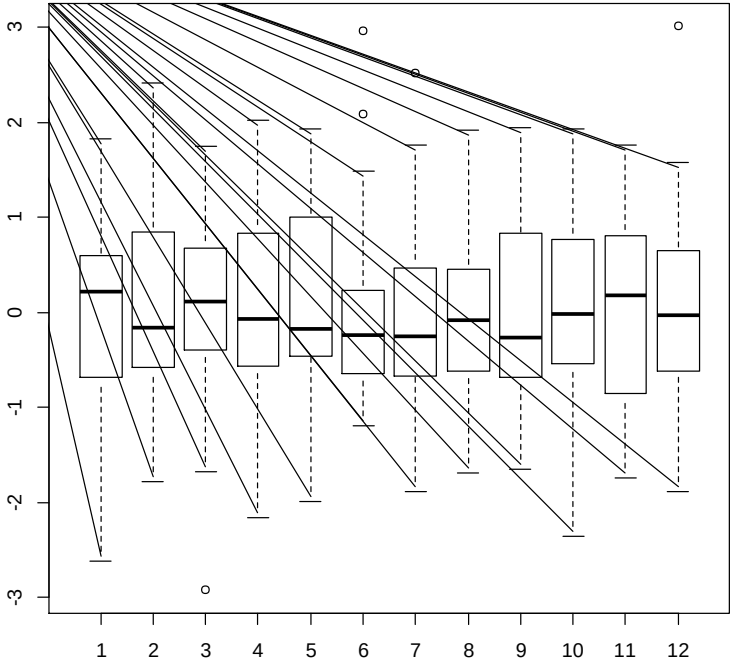
$$\mathrm{var}(z_i) = 1$$

Analysis of standardized matrix

Traditional indicators (original scale).



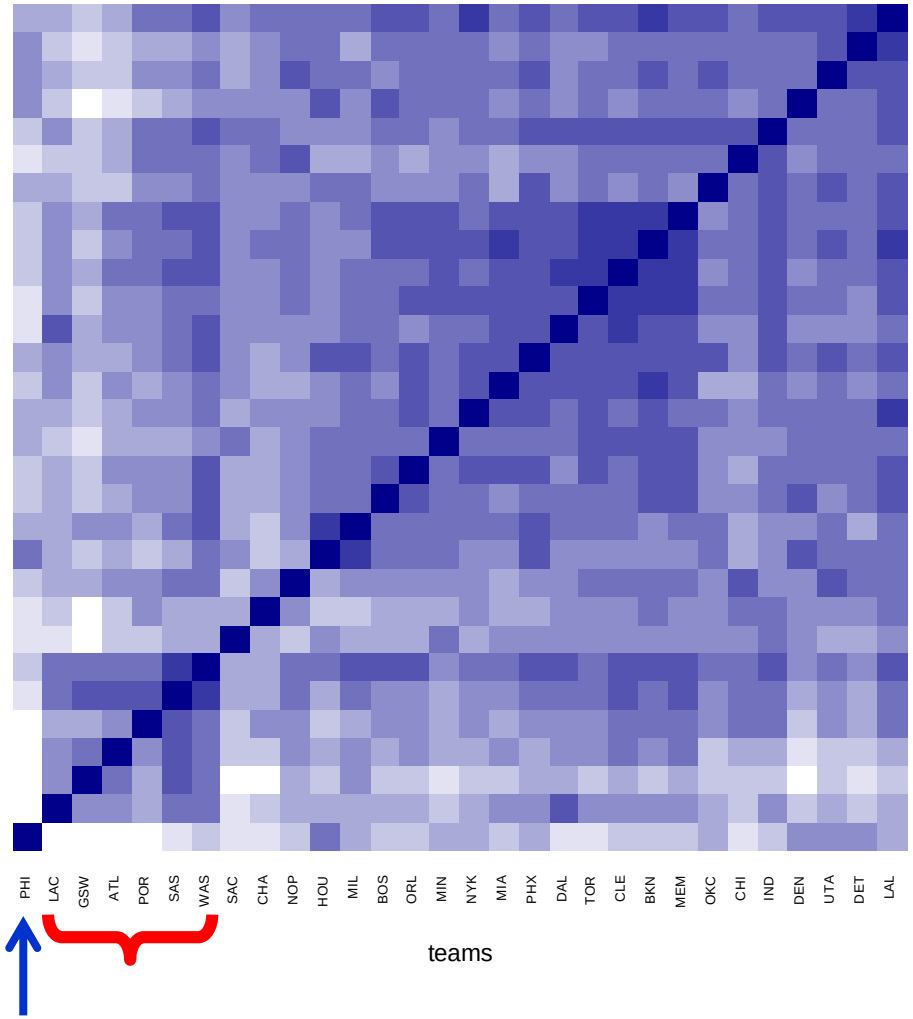
Traditional indicators (standardized scale).



$$\bar{z}_i = 0$$

$$\mathrm{var}(\bar{z}_i) = 1$$

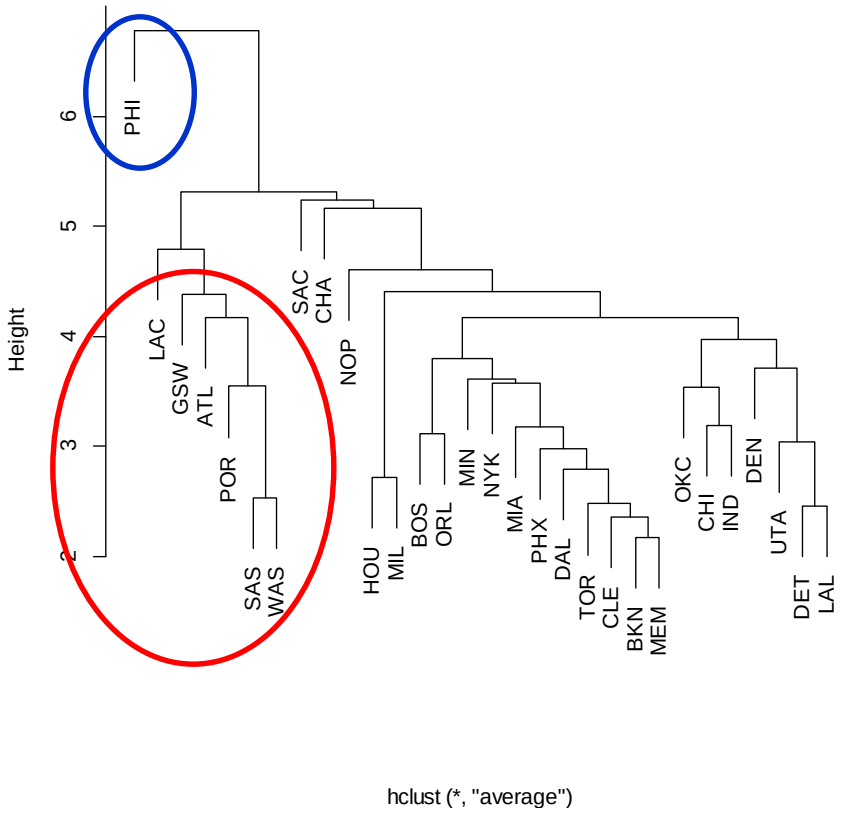
Similarity matrix



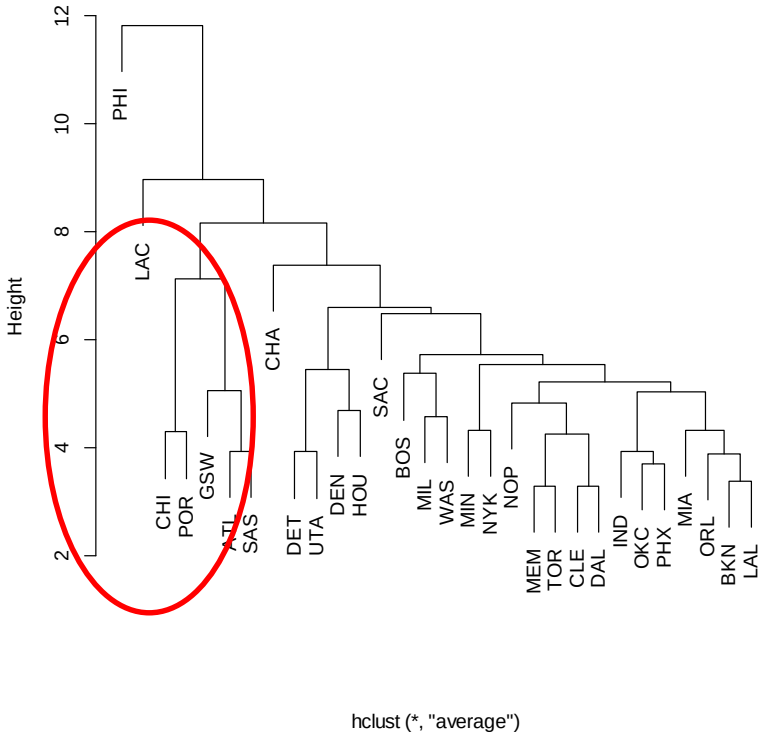
$$D_E\left(z_i, z_j\right) = \sqrt{\sum_{k=1}^{P}\left(z_{ik} - z_{jk}\right)^2}$$

Clustering (average linkage)

Standardized data



Original data



Example: cluster analysis for the Wine data

The wine data

- The wine dataset contains the results of a chemical analysis of wines grown in a specific area of Italy.
- Three types of wine:
  - 1 (59 observations).
  - 2(71 observations).
  - 3 (48 observations).

  For the analysis: the types are unknowns
- 178 samples.
- 13 chemical analyses recorded for each sample.
- Data : UCI Machien Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Wine

Data structure

178 wine samples

$$\begin{bmatrix} x_{11} & x_{21} & \cdot & \cdot & x_{n1} \\ x_{12} & x_{22} & \cdot & \cdot & x_{n2} \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ x_{1,178} & x_{2m} & \cdot & \cdot & x_{12,178} \end{bmatrix}$$

13 chemical tests

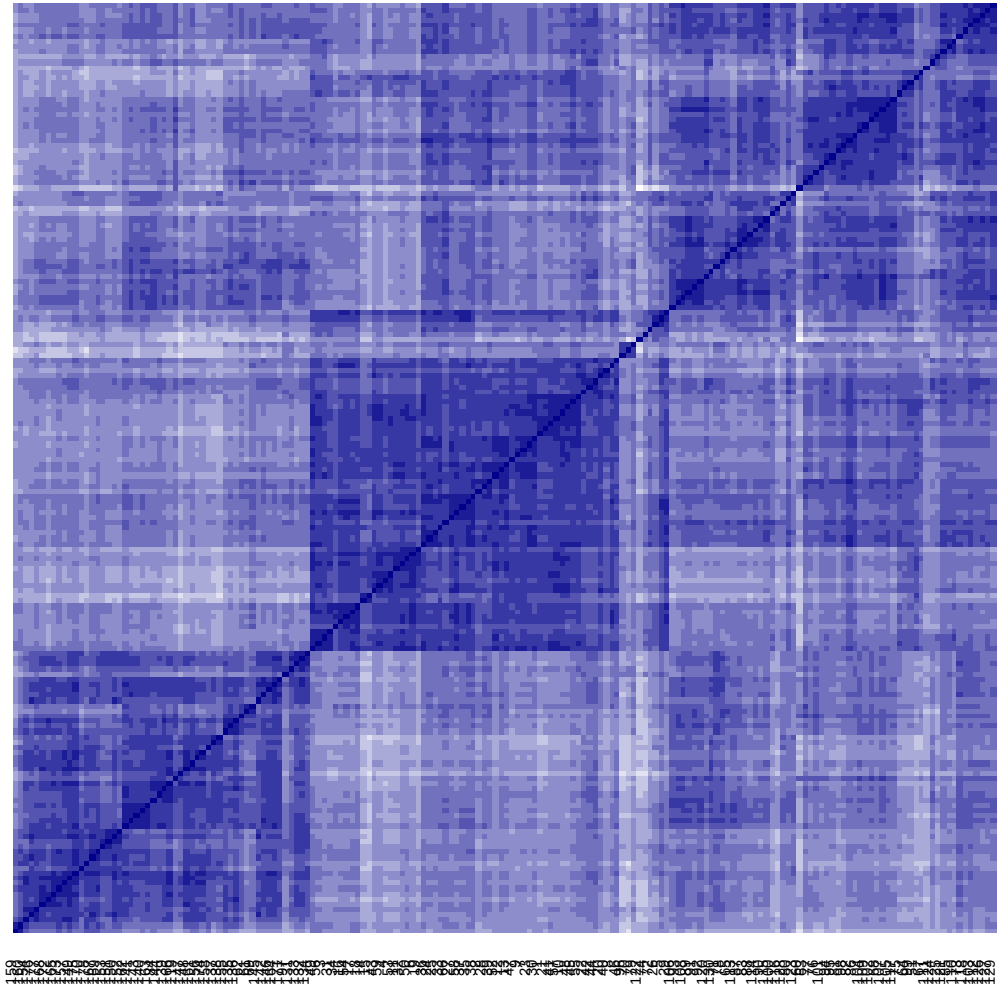
Membership:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} \qquad Y_i = \begin{cases} 1 & S_i \in A \\ 2 & S_i \in B \\ \cdot & \\ \cdot & \\ K & S_i \in K \end{cases}$$

Unsupervised analysis:

The membership is unobserved variable.

Similarity matrix

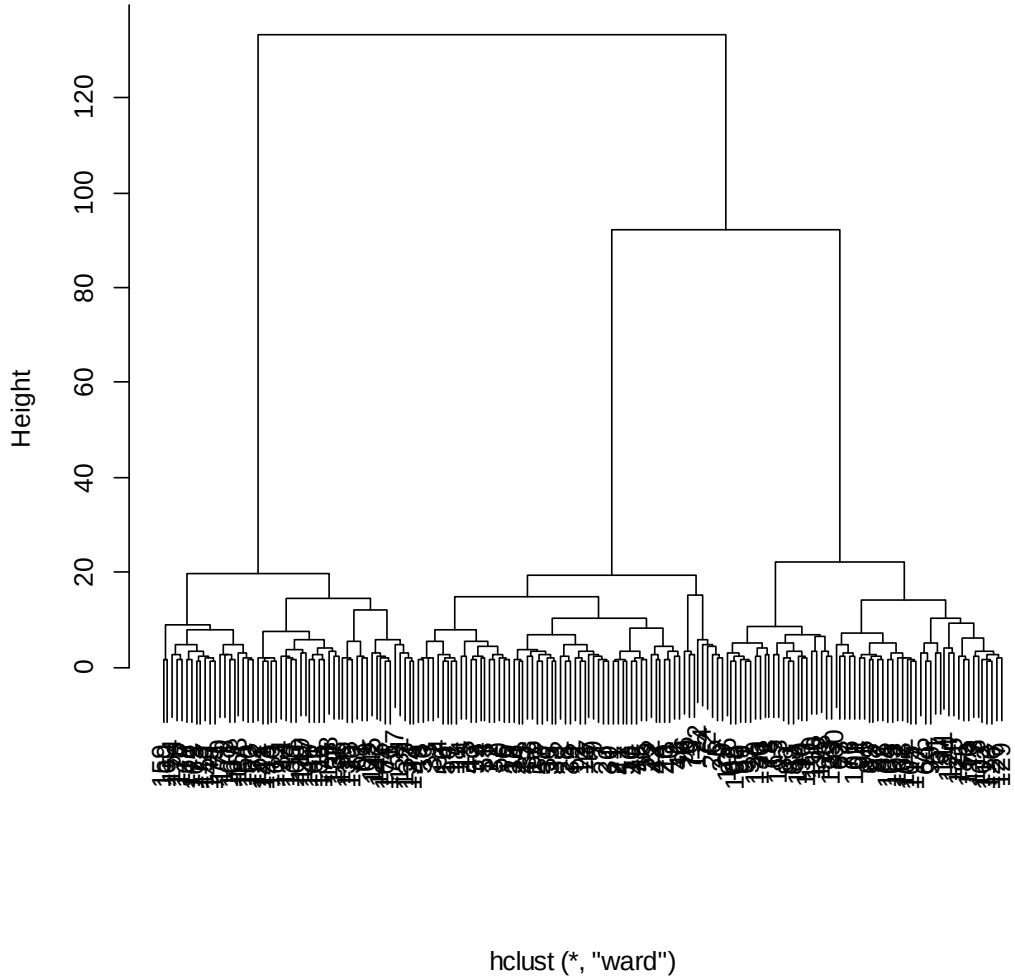


wine

Euclidean distance:

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

Clustering



- Ward linkage.
- Three clusters.

Summary

- Results of cluster analysis should be treated with CAUTION:
  - Results are exploratory.
  - Use cluster analysis to discover patterns.
  - ~~Many things can vary in a cluster analysis~~
  - ~~If covariates/group labels are known, then clustering is usually inefficient.~~

Partitioning

k-means

Partitioning or Hierarchical?

- Hierarchical
  - Advantages
    - Faster computation.
    - Visual.
  - Disadvantages
    - Unrelated genes are eventually joined
    - ~~Rigid, cannot correct later for erroneous decisions made earlier.~~
    - Hard to define number of cluster clusters.
- Partitioning:
  - Advantages
    - Optimal for certain criteria.
    - Genes automatically assigned to clusters.
  - Disadvantages
    - Need initial k;
    - Often require long computation times.
    - All genes are forced into a cluster.

Smallest distance from the center of the cluster.

Partitioning methods

- Partition the data into a prespecified number *k* of mutually exclusive and exhaustive groups.
- Iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares.
- Example: k-means.

K-means

- Euclidean distance most often used
- Can be hard to choose or figure out K (=number of clusters).
- Not unique solution: clustering can depend on initial partition.

K-means Algorithm

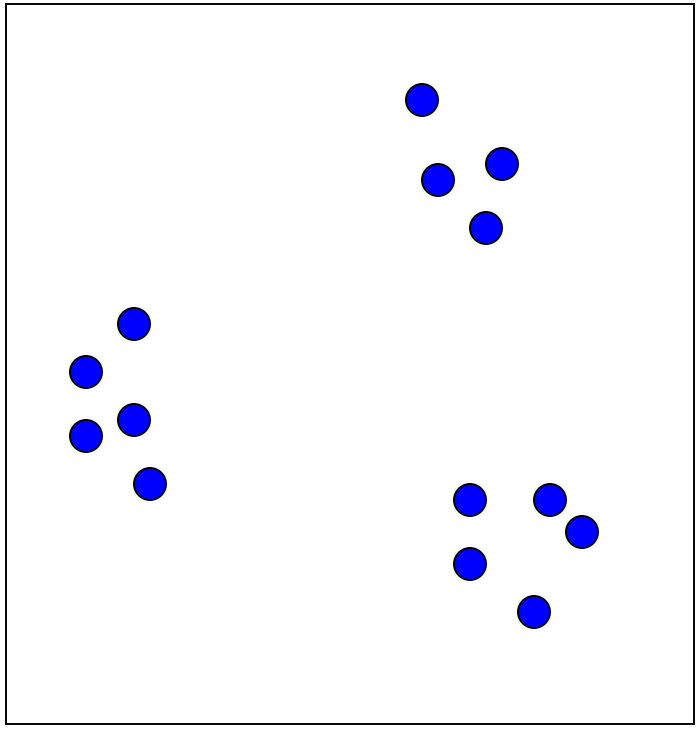
1. Choose K centroids at random
2. Make initial partition of objects into K clusters by assigning objects to closest centroid
3. Calculate the centroid (mean) of each of the k clusters.
4. a. For object i, calculate its distance to each of the centroids.

   b. Allocate object i to cluster with closest centroid.

   c. If object was reallocated, recalculate centroids based on new clusters.
4. Repeat 3 for object i = 1,….N.
5. Repeat 3 and 4 until no reallocations occur.
6. Assess cluster structure for fit and stability

K-means

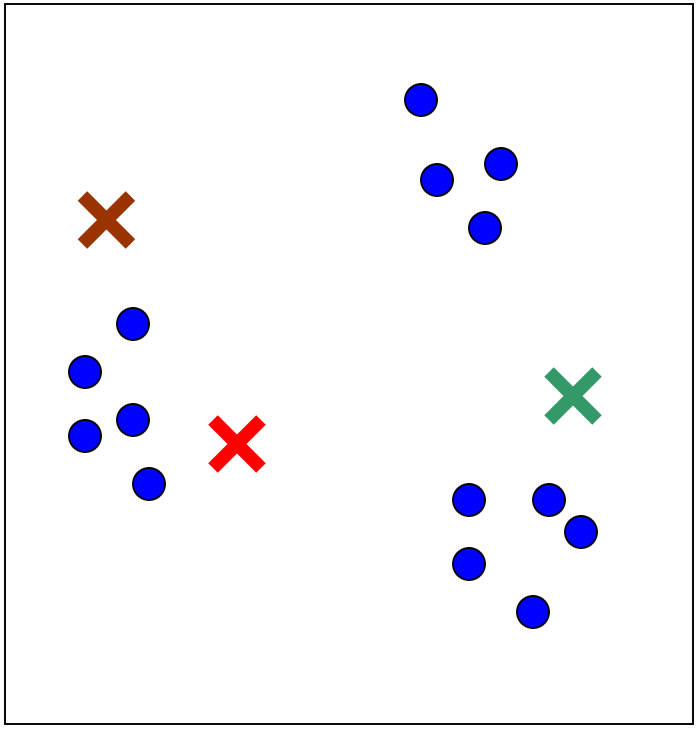
- Example:
  - 14 samples.
  - Two genes (=variables).



Iteration = 0

K-means

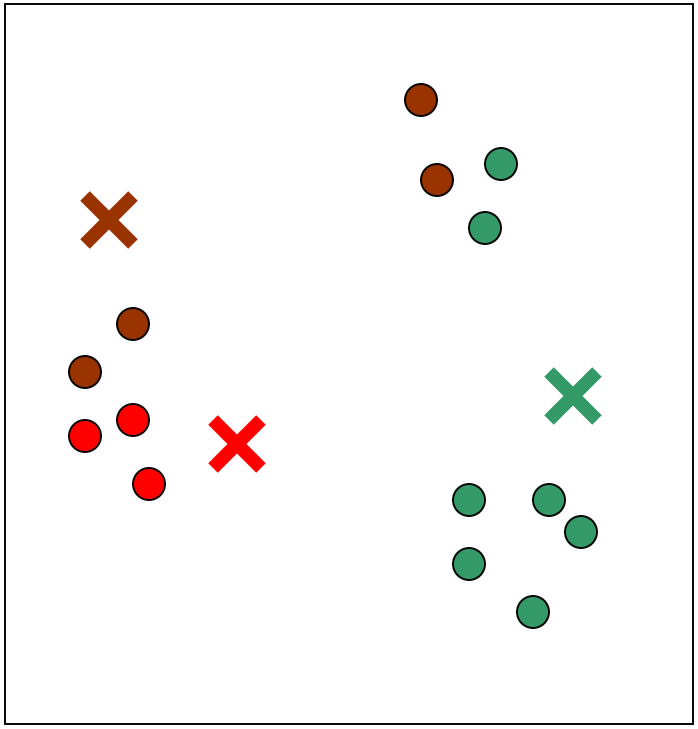
- K=3.
- Choose K *centroids.*
- These are starting values that the user picks.
- There are some data driven ways to do it.



Iteration = 0

K-means

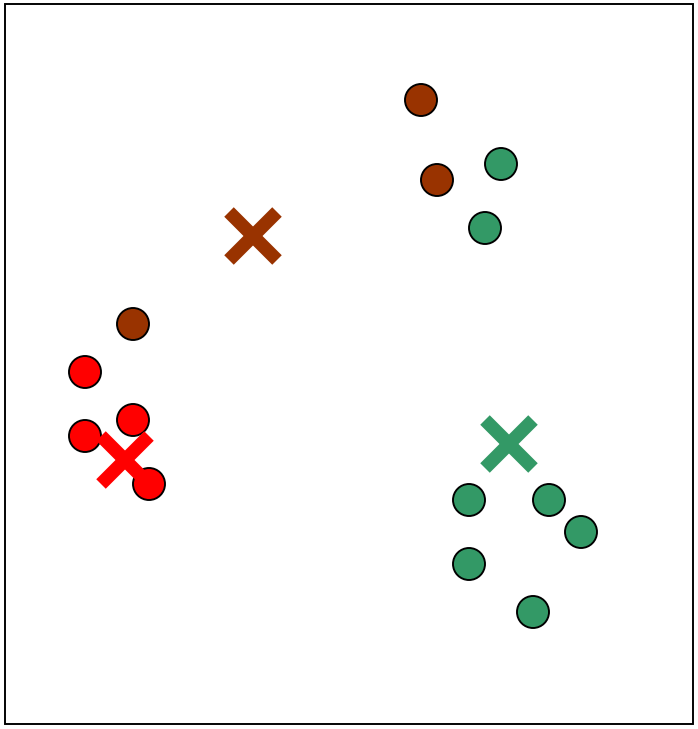
- 3 clusters.
- Make first *partition* by finding the closest centroid for each point.
- Choose a distance measure.



Iteration = 1

K-means

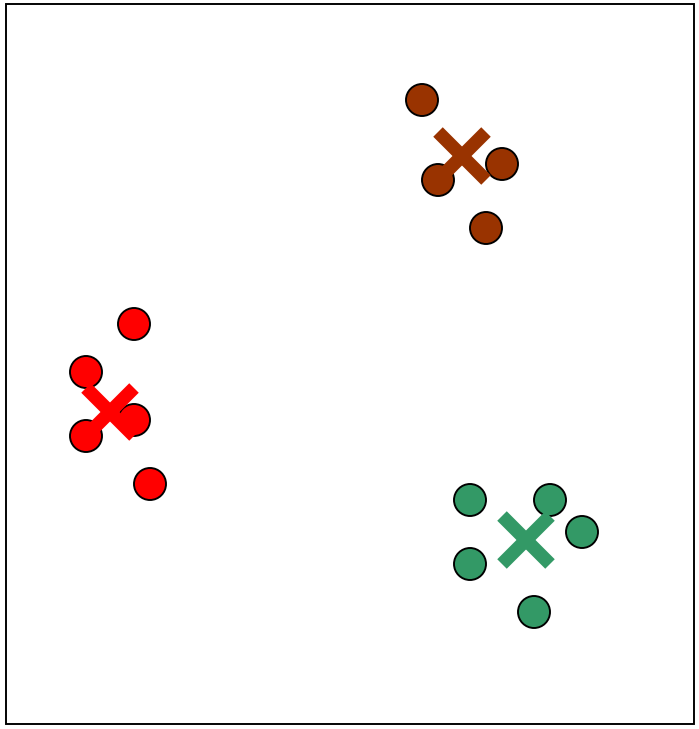
- Re-compute the centroids by taking the middle of each cluster.



Iteration = 2

K-means

- Repeat until the centroids stop moving or until you get tired of waiting.

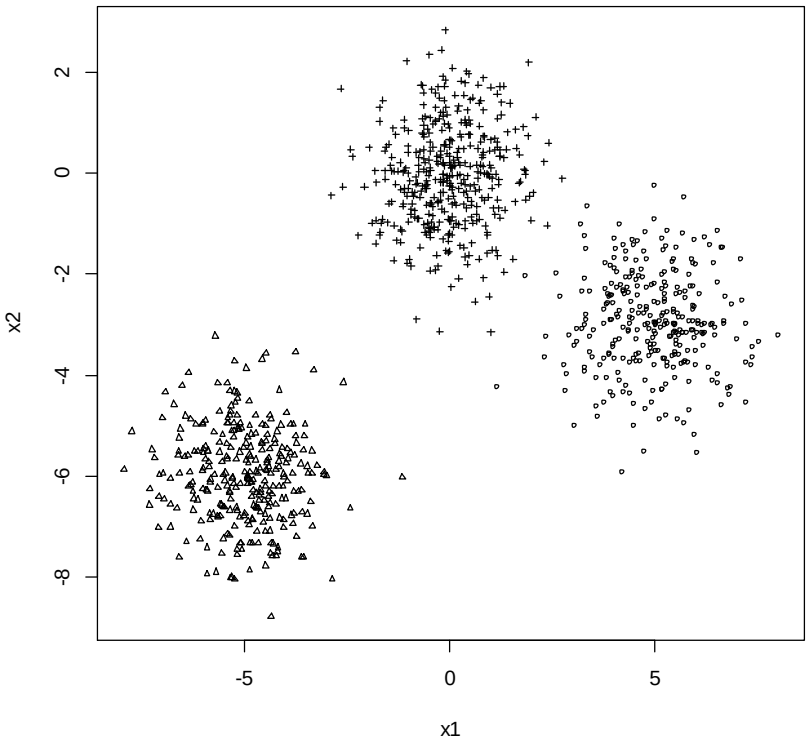


Iteration = 3

K-means Limitations

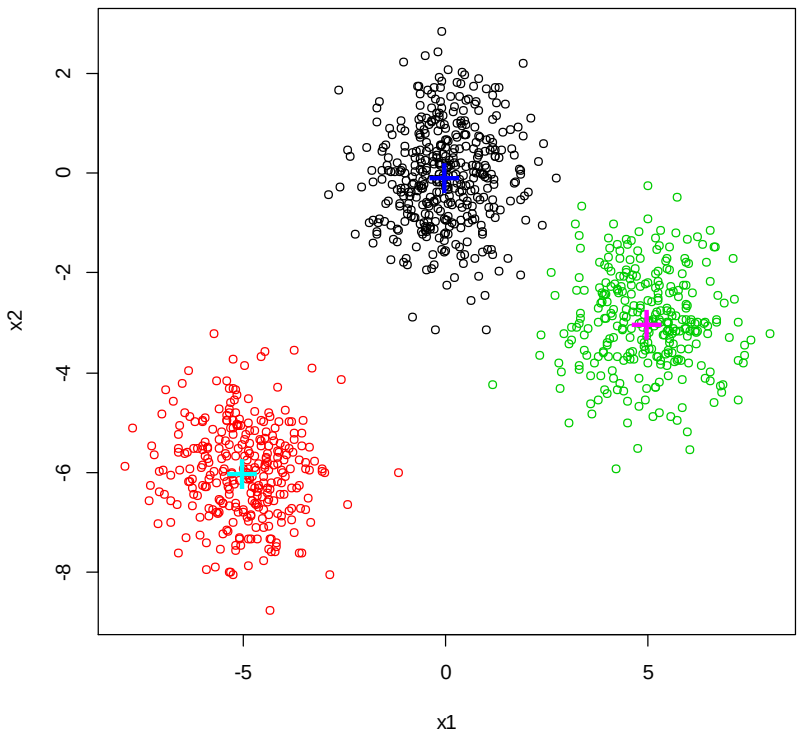
- Final results depend on starting values
- How do we chose K? There are methods , see example later.

Example 1



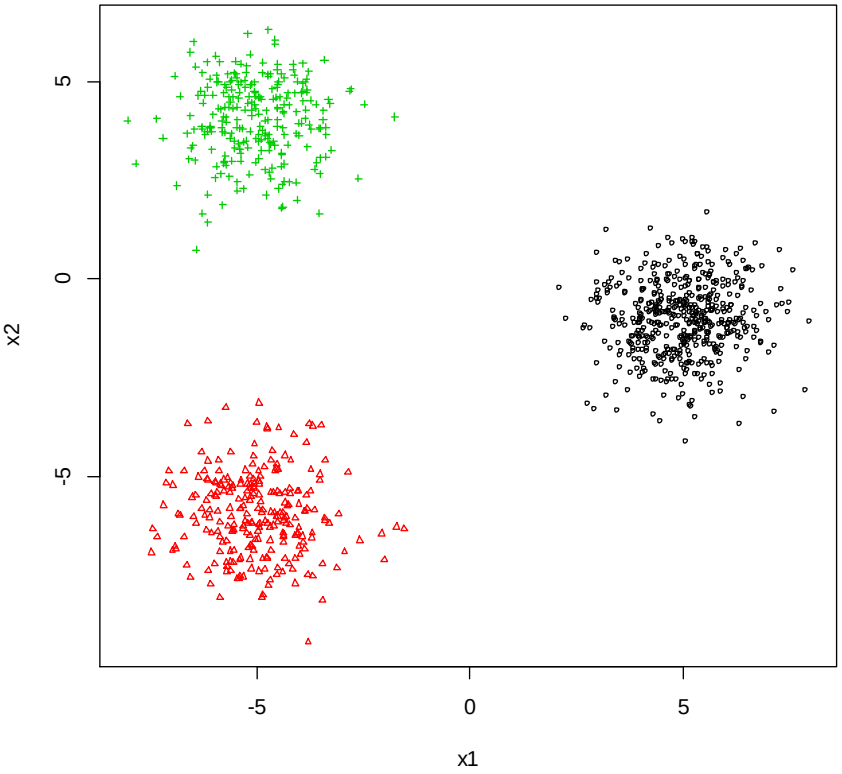
- Two variables.
- Three clusters (on both variables).

K-means



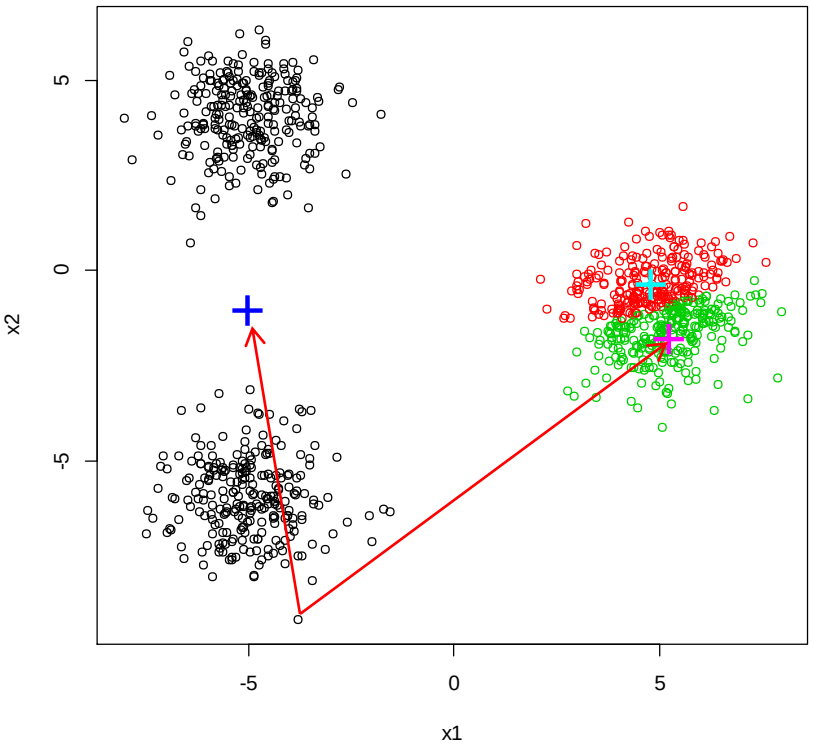
- Identification of three clusters.

Example 2



- Two variables.
- Three clusters based on one variable and two based on the second variable.

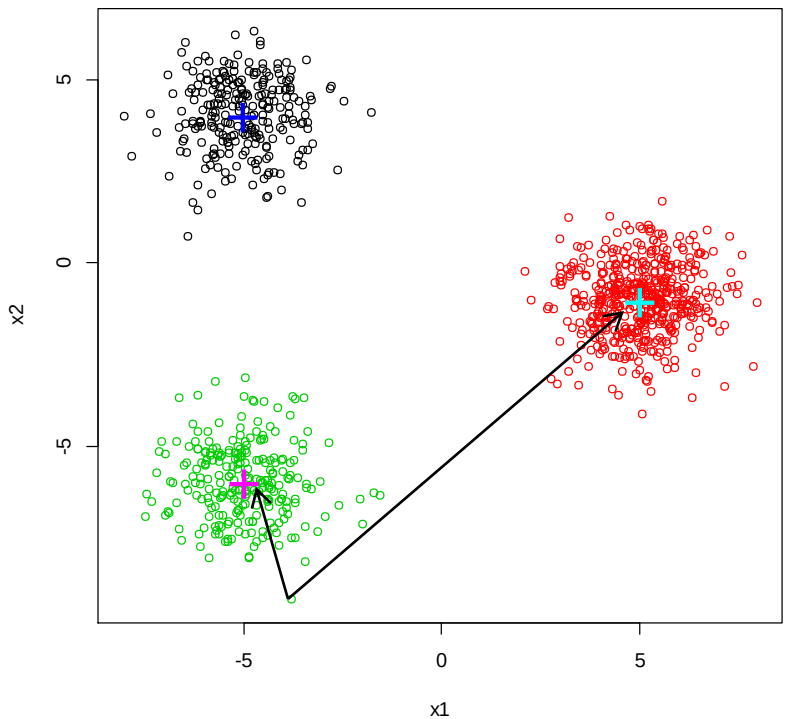
K-means: first run




```
Within cluster sum of squares by cluster:
[1] 13543.0295   311.9215   359.9933
 (between_SS / total_SS =  64.2 %)
```

- Identification of three clusters.
- Distance from the center.

K-means: second run




```
Within cluster sum of squares by cluster:
[1] 488.0388 955.3697 531.4910
 (between_SS / total_SS =  95.0 %)
```

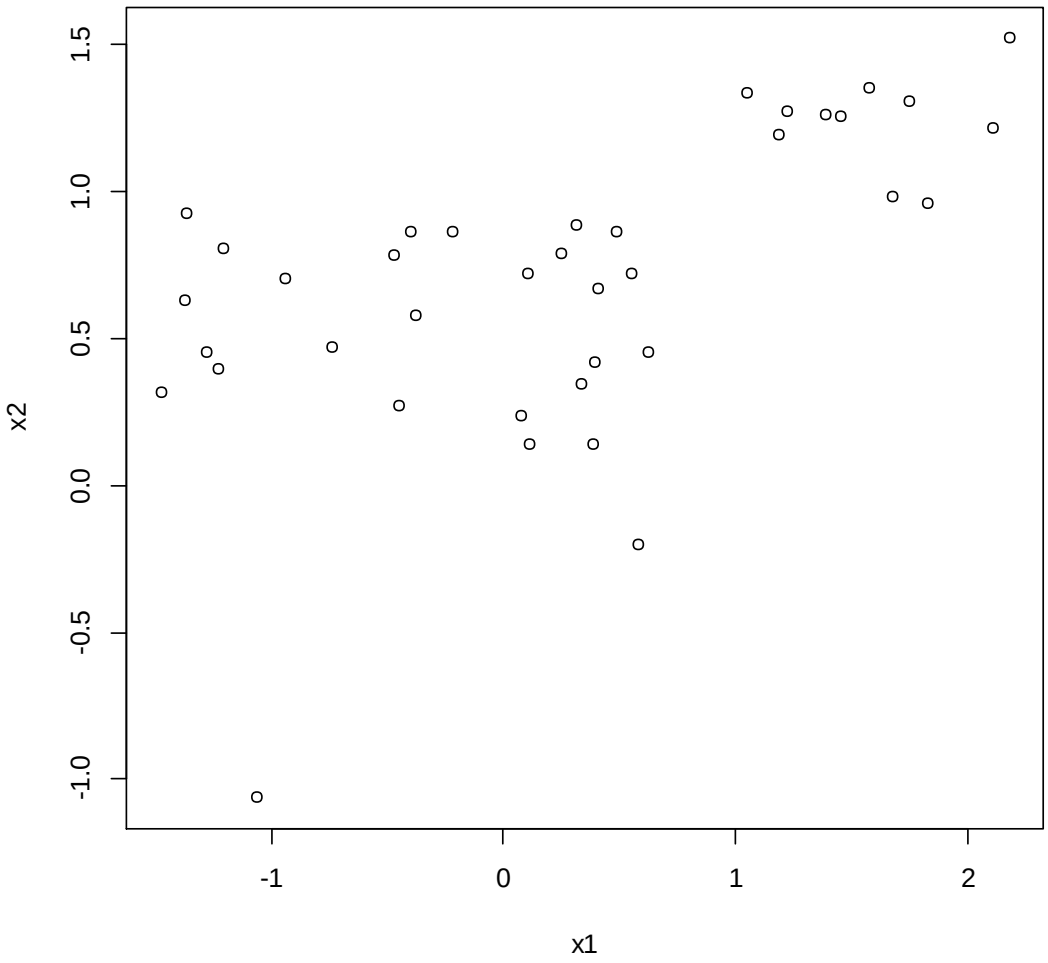
- Identification of three clusters.
- Distance from the center.

Example: K means for the Golub data

Clustering using SPCA

Example: the golub data

- Select the top 2 genes (by t-test)

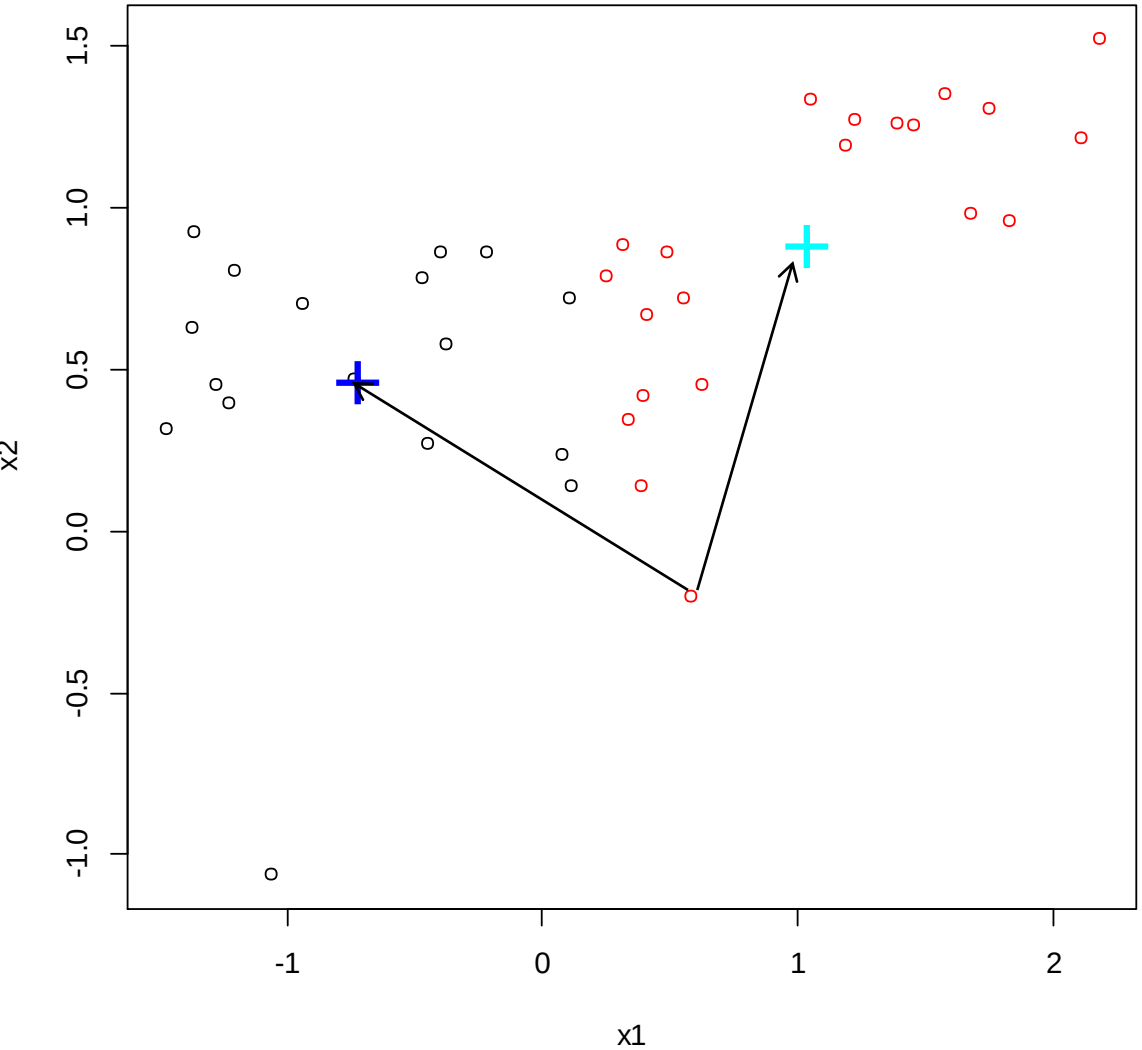


K-means for the golub data: first run

K mean clustering with two clusters.

```
> z<-cbind(x1,x2)
> k1<-kmeans(z, 2)
> k1
K-means clustering with 2 clusters of sizes 17, 21

Cluster means:
          x1        x2
1 -0.7234071 0.4763982
2  1.0377081 0.8933467

Clustering vector:
 [1] 1 1 1 2 2 1 1 1 1 2 1 2 2 1 2 2 2 1 2 1 1 1 1 1 2 1 1 2 2 2 2
2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1]  8.474071 12.625066
 (between_SS / total_SS =  59.3 %)
```

K-means for the golub data: first run

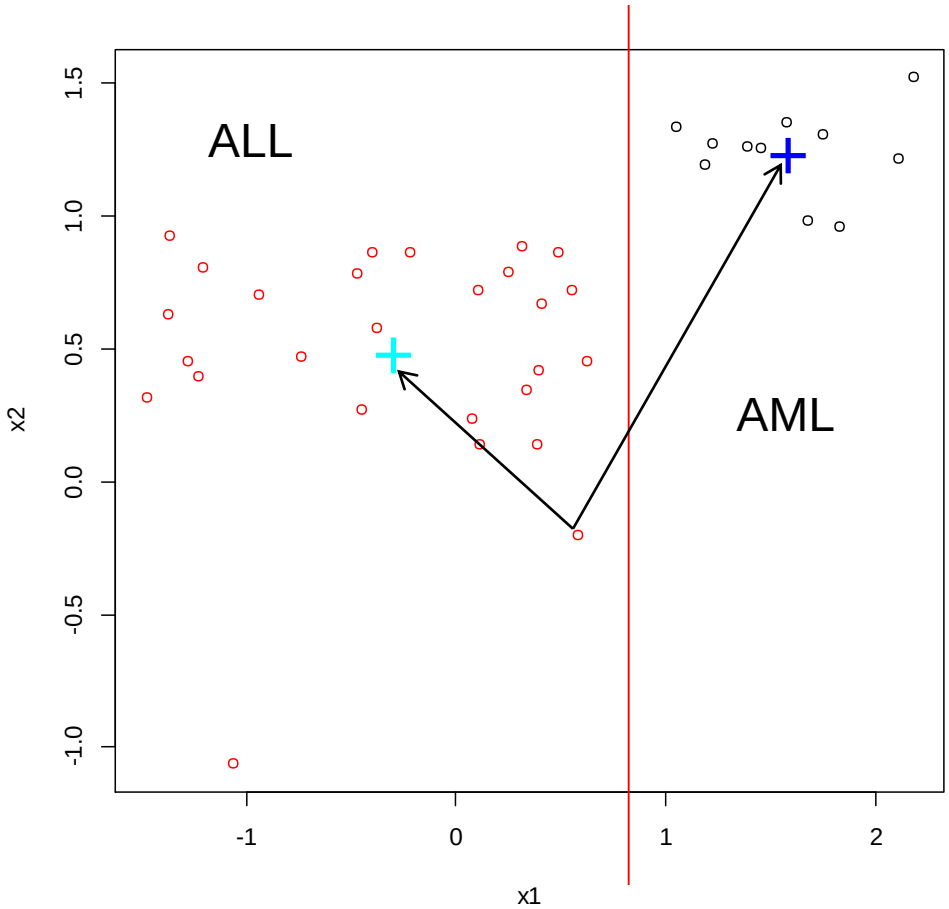


- Identification of two clusters.
- Distance from the center.

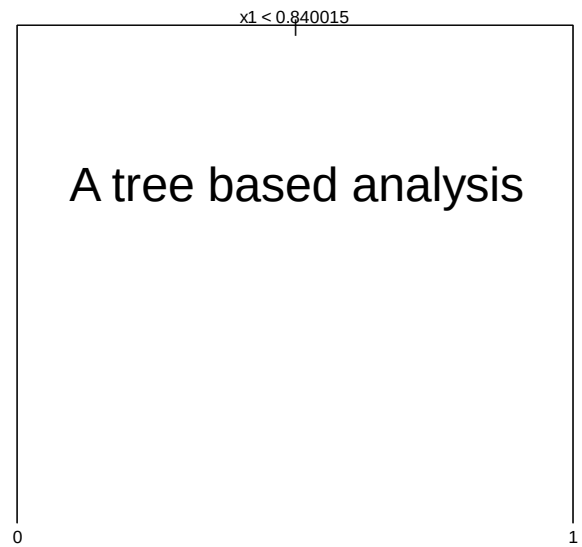
K-means for the golub data: second run

```
K-means clustering with 2 clusters of sizes 11, 27

Cluster means:
          x1        x2
1  1.5866682 1.2428364
2 -0.2947926 0.4884389

Clustering vector:
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1
1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1]  1.606477 18.148764
 (between_SS / total_SS =  61.9 %)
```

K-means for the golub data: second run



- Identification of two clusters (AML and ALL).
- Distance from the center.




```
1) root 38 45.73 0 ( 0.7105 0.2895 )
  2) x1 < 0.840015 27  0.00 0 ( 1.0000 0.0000 ) *
  3) x1 > 0.840015 11  0.00 1 ( 0.0000 1.0000 ) *
```

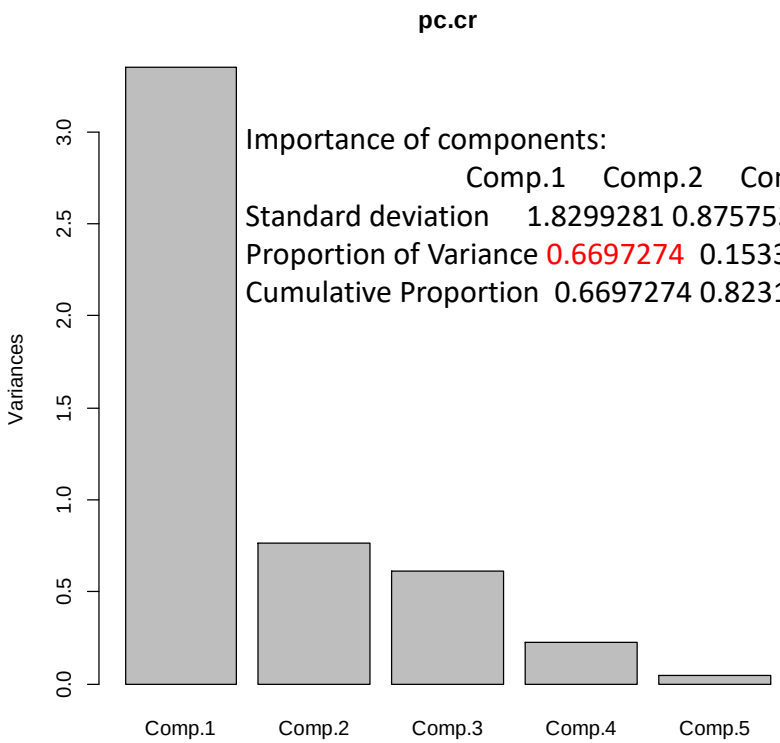
Clustering using SPCA

- Supervise PCA for the golub data.
- Form a reduced matrix.
- Calculate first PC for the reduce matrix.

$$\underbrace{\begin{bmatrix} x_{11} & x_{21} & . & . & x_{n1} \\ x_{12} & x_{22} & . & . & x_{n2} \\ . & . & . & . & \\ . & . & . & . & \\ x_{15} & x_{25} & . & . & x_{n5} \end{bmatrix}}_{\text{38 samples}} \left.\vphantom{\begin{bmatrix} x \\ x \\ . \\ . \\ x \end{bmatrix}}\right\} \text{Top 5 genes} \Rightarrow \text{Calculate PCs}$$

SPCA: top 5 genes

Variance of the components



Importance of components:

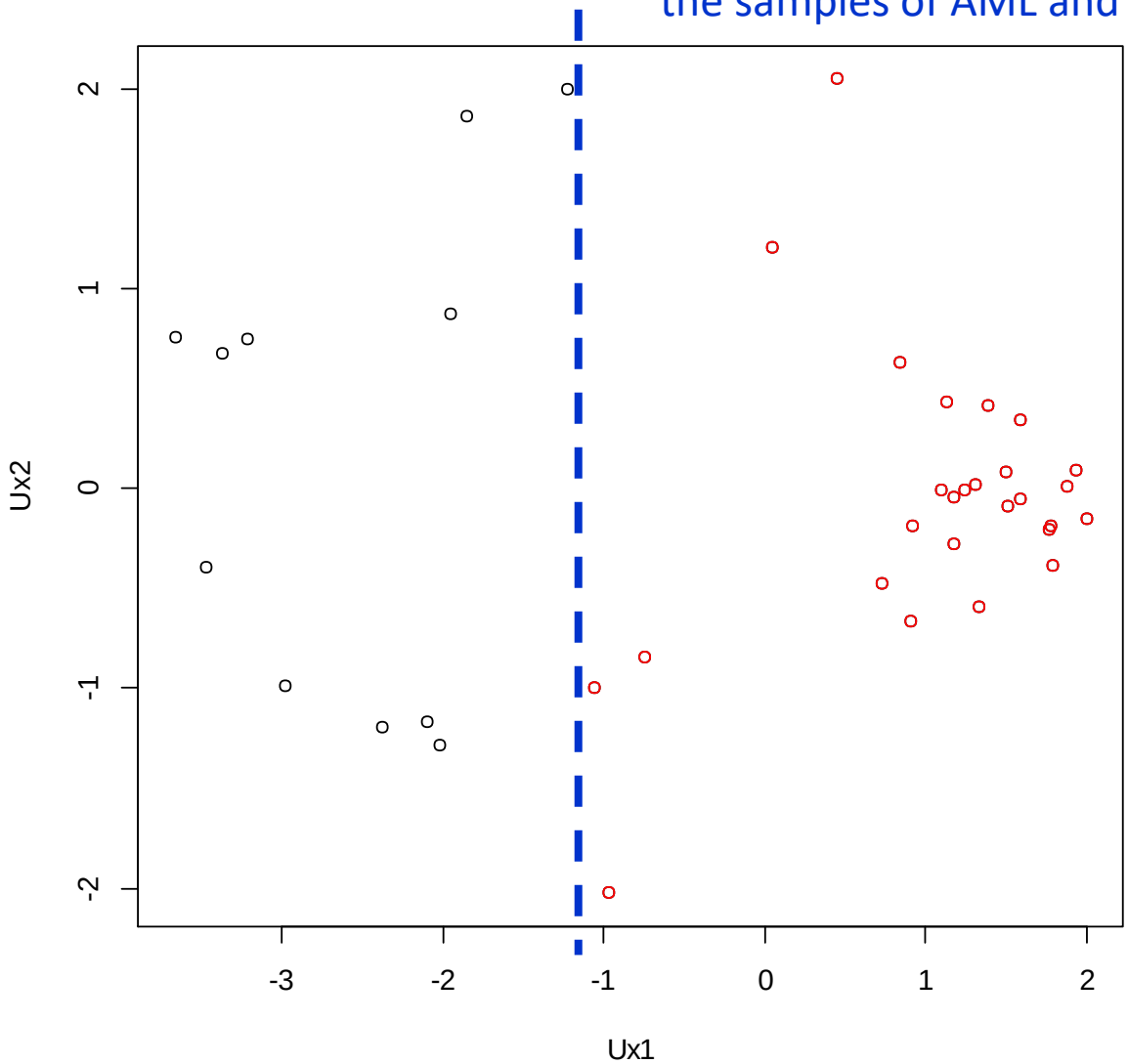
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.8299281	0.8757530	0.7840983	0.4730951	0.213987221
Proportion of Variance	0.6697274	0.1533887	0.1229620	0.0447638	0.009158106
Cumulative Proportion	0.6697274	0.8231161	0.9460781	0.9908419	1.000000000

First PC explain 66.97% of the total variability

PC1 Vs. PC2

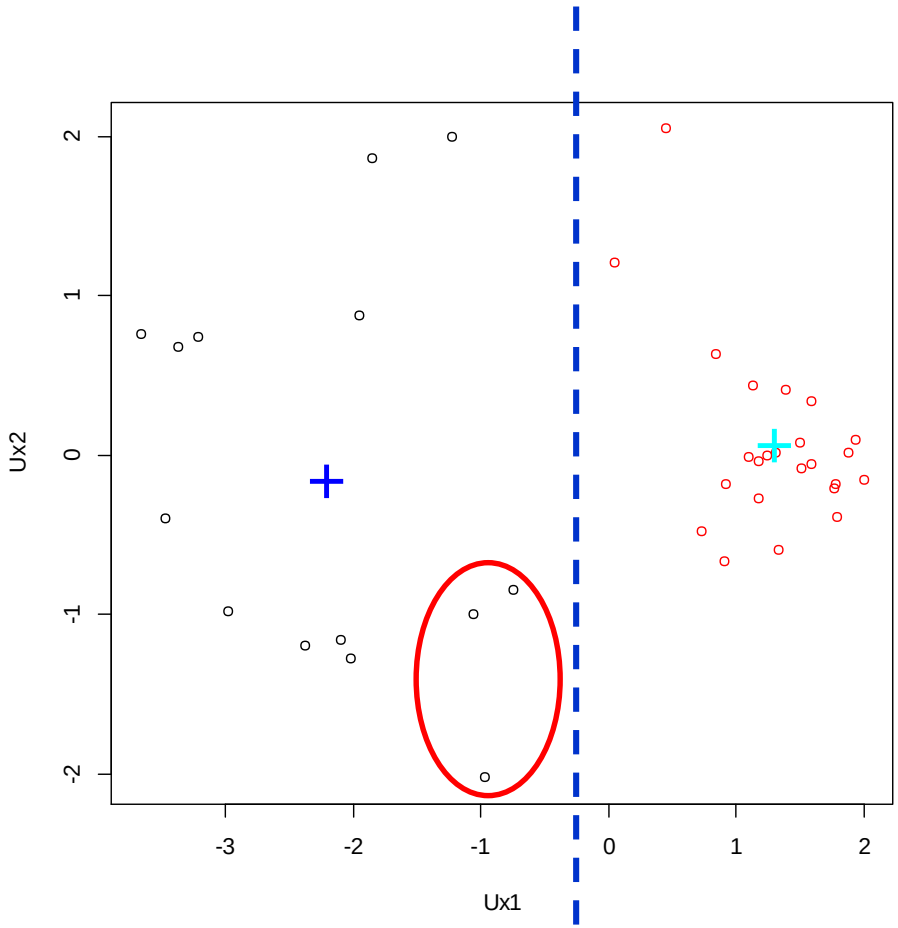
Based on PC1 we can separate the samples of AML and ALL

15.33%



66.97%

K-means for the golub data: first run




```
Within cluster sum of squares by cluster
[1] 33.87450 13.21048
 (between_SS / total_SS =  69.9 %)
```

Example: K means clustering for the NBA data

Data structure

A *30 X 12* matrix:

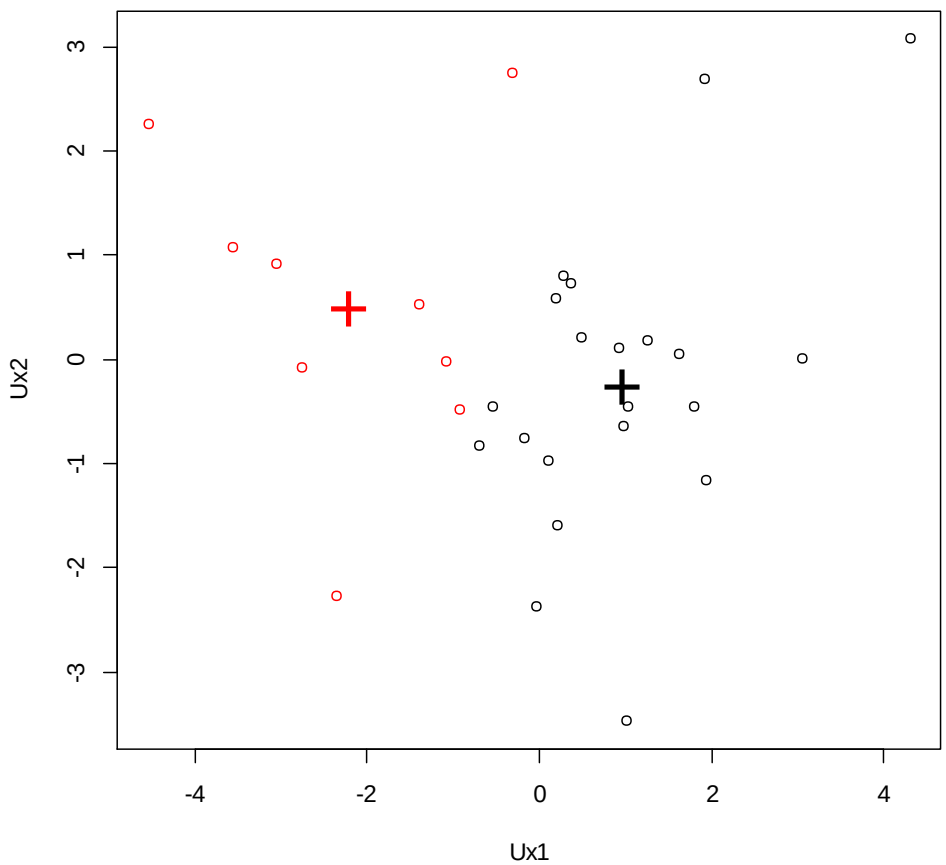
- 2-pt & 3-pt Successful
- 2-pt & 3-pt Unsuccessful
- Free Throw Successful & Unsuccessful
- Defensive & Offensive Rebounds
- Assists
- Turnovers
- Steals
- Dunks
- Blocks Committed / Received
- Fouls Committed / Received

$$\underbrace{\begin{bmatrix} x_{1,1} & x_{21} & \cdot & \cdot & x_{12,1} \\ x_{1,2} & x_{22} & \cdot & \cdot & x_{12,2} \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ x_{1,30} & x_{2,88} & \cdot & \cdot & x_{12,30} \end{bmatrix}}_{\text{12 performance indicators}} \Bigg\} \text{30 NBA teams}$$

PCA

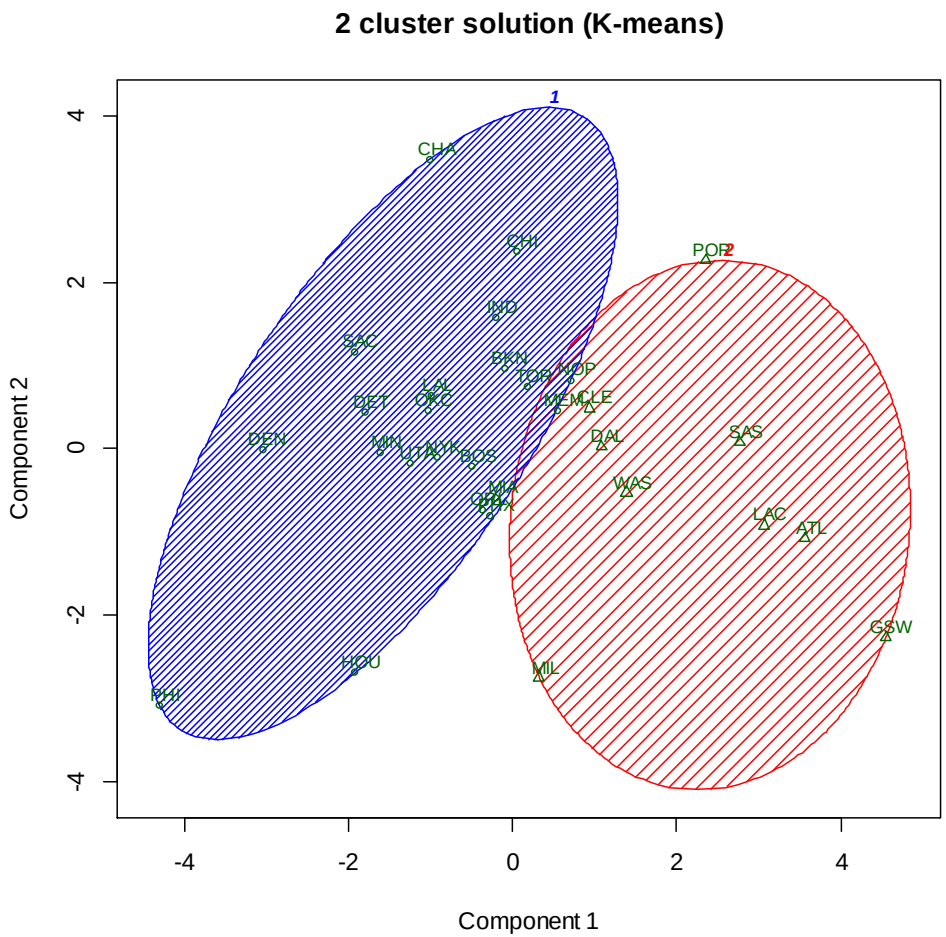
$$\begin{bmatrix} U(x_1)_1 & U(x_2)_1 \\ U(x_1)_2 & U(x_2)_2 \\ & \\ & \\ U(x_1)_{30} & U(x_{30})_2 \end{bmatrix}$$

K-means using PC1 & PC2



- Two cluster solution.
- Separation based on PC1.

K-means with original 12 indicators



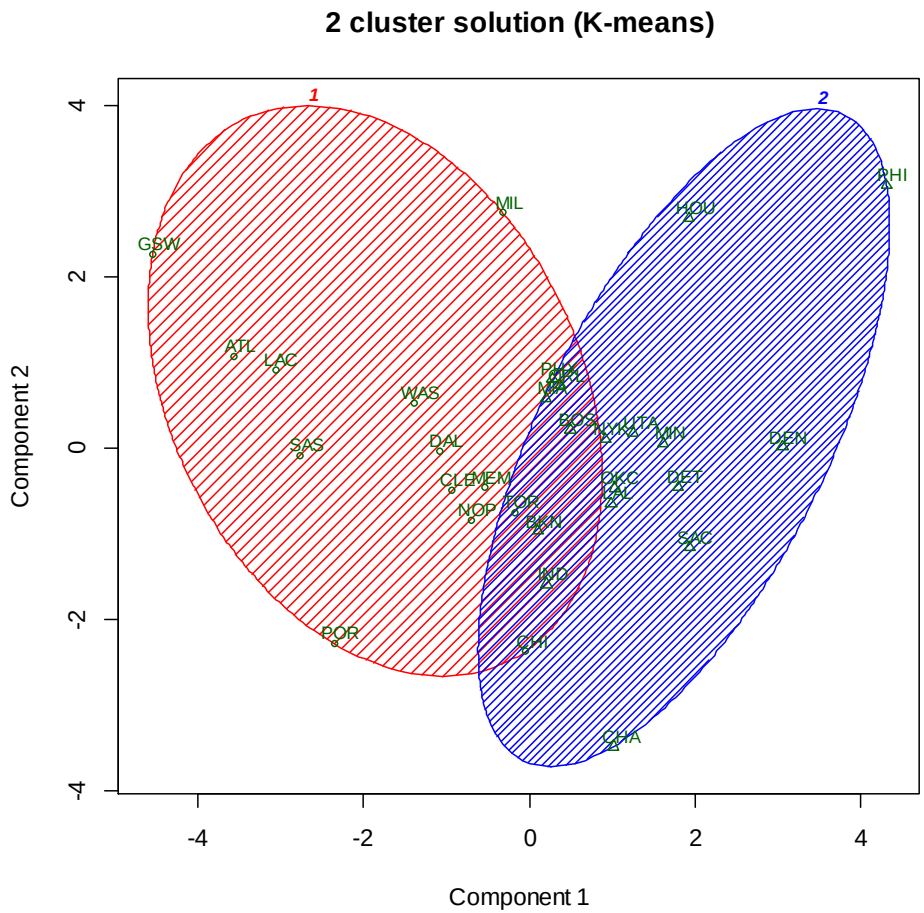
These two components explain 100 % of the point variability.

$$\hat{Y}_i = \begin{cases} 1 & team_i \in Cluster_1 \\ 2 & team_i \in Cluster_2 \end{cases}$$

	1	2	
1	14	0	Not PlayOff teams
2	7	9	PlayOff teams

$\hat{Y}_i$

K-means using all performance indicators



These two components explain 47.62 % of the point variability.

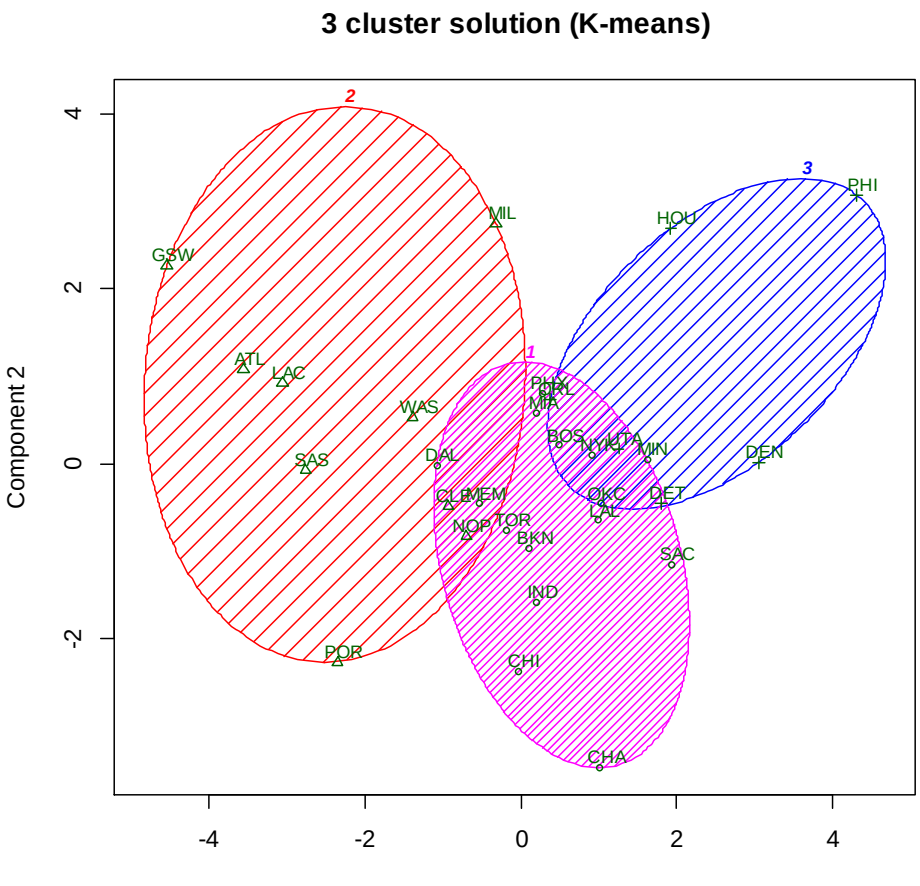
A two clusters solution.

	1	2
1	0	14
2	13	3

PlayOff teams

$\hat{Y}_i$ The "PlayOff"cluster

K-means using all performance indicators



These two components explain 47.62 % of the point variability.

A three clusters solution.

1	2	3
1	9	0 5
2	6	9 1

PlayOff teams

Example: k means for the for the Wine data

The wine data

- The wine dataset contains the results of a chemical analysis of wines grown in a specific area of Italy.
- Three types of wine:
  - 1 (59 observations).
  - 2(71 observations).
  - 3 (48 observations).

  For the analysis: the types are unknowns
- 178 samples.
- 13 chemical analyses recorded for each sample.
- Data : UCI Machien Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Wine

Data structure

178 wine samples

$$\begin{bmatrix} x_{11} & x_{21} & \cdot & \cdot & x_{n1} \\ x_{12} & x_{22} & \cdot & \cdot & x_{n2} \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ x_{1,178} & x_{2m} & \cdot & \cdot & x_{12,178} \end{bmatrix}$$

13 chemical tests

Membership:

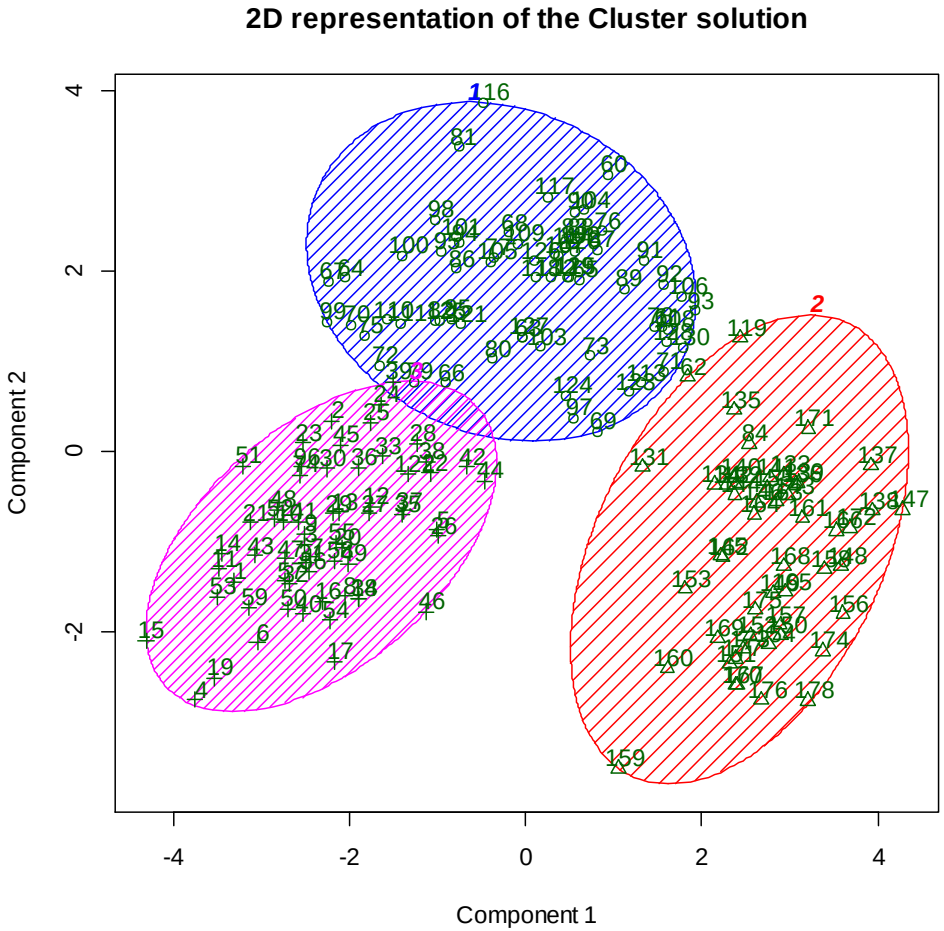
$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}$$

$$Y_i = \begin{cases} 1 & S_i \in A \\ 2 & S_i \in B \\ 3 & S_i \in C \end{cases}$$

The membership is unobserved variable.

- Three types of wine:
  - 1 (59 observations).
  - 2(71 observations).
  - 3 (48 observations).

Three cluster solution



These two components explain 55.41 % of the point variability.

Within cluster sum of square

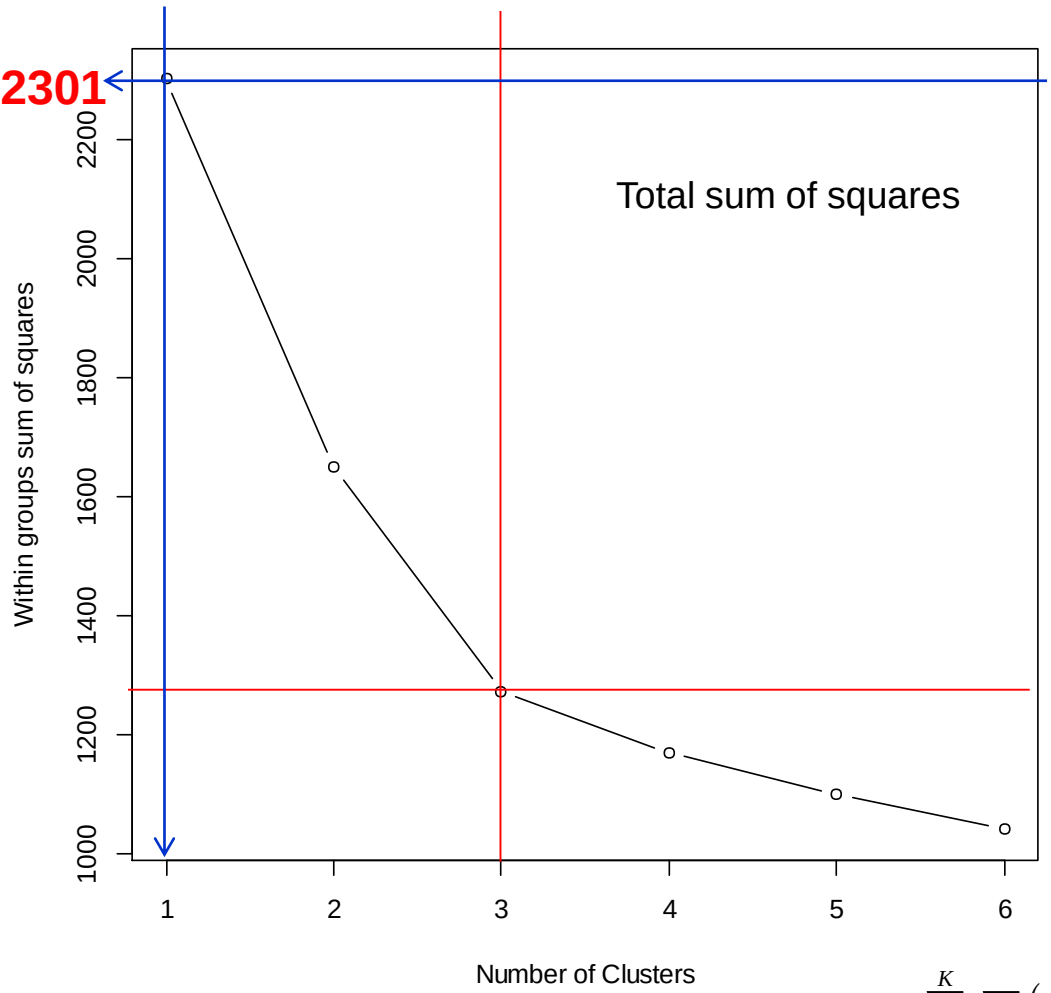
Sum of squares within a cluster S_k:

$$\sum_{x_j \in S_k} \left(x_j - \hat{\mu}_k \right)^2$$

Within cluster sum of squares (over all clusters):

$$\sum_{k=1}^{K} \sum_{x_j \in S_k} \left(x_j - \hat{\mu}_k \right)^2$$

Within cluster sum of squares



Within cluster sum of squares by cluster:
[1] 558.6971 326.3537 385.6983
(between_SS / total_SS = 44.8 %)

$$1-\frac{1270.74}{2301}=0.4477$$

1270.749

The cluster structure explains 44.77% of the total variability.

$$\sum_{k=1}^{K}\sum_{x_j\in S_k}\left(x_j-\mu_k\right)^2=558.6971+326.3537+385.6983=1270.749$$

Cluster membership

$$\hat{Y}_i = \begin{cases} 1 & S_i \in A \\ 2 & S_i \in B \\ 3 & S_i \in C \end{cases}$$

```
> table(as.numeric(wine.type),k.means.fit$cluster)

   1  2  3
1  0  0 59
2 65  3  3
3  0 48  0
```

True type (only for illustration, in practice the "real type" is unknown).

$\hat{Y}_i$

K-means in R

```
> k1 <- kmeans(wine.stand, 3)
```

data

number of clusters

R output:

```
> k1$size
[1] 51 65 62
> k1$totss
[1] 2301
> k1$withinss
[1] 326.3537 558.6971 385.6983
> k1$tot.withinss
[1] 1270.749
```

K-means in R

Cluster means for two variables: alcohol and Malic:

```
> k1$centers[,1:2]
    Alcohol      Malic
1  0.1644436  0.8690954
2 -0.9234669 -0.3929331
3  0.8328826 -0.3029551
```

Aggregating Bundles of Clusters (the ABC method)

Aggregating Bundles of Clusters

- An ensemble method for cluster analysis
  - Usual (partial) solution: Filtering the genes based on variance or coefficient of variation reduces the error rates.
  - Ensemble approach: Filter genes repeatedly and apply an ensemble technique.
- Amaratunga, Cabrera and Kovtun(*Biostatistics, 2008)*

ABC

Gene expression matrix

	S1	S2	S3	S4	S5	S6
G8521	1003	1306	713	1628	1268	1629
G8522	890	705	566	975	883	1005
G8523	680	749	811	669	724	643
G8524	262	311	336	1677	1286	1486
G8525	254	383	258	1652	1799	1645
G8526	81	140	288	298	241	342
G8527	4077	2557	2600	3394	2926	2755
G8528	2571	1929	1406	2439	1613	5074
G8529	55	73	121	22	141	44
G8530	1640	1693	1517	1731	1861	1550
G8531	168	229	284	220	310	315
G8532	323	258	359	345	308	315
G8533	12131	11199	14859	11544	11352	11506
G8534	11544	11352	12131	11199	14859	12529
G8535	1929	1406	2439	254	383	258
G8536	191	140	288	298	241	342
G8537	4077	2557	2600	3394	2926	2755
G8538	2571	1613	5074	1652	1799	1645
G8539	55	73	121	22	91	24
G8540	1640	1693	1517	1731	1861	1750
G8541	168	229	284	220	312	335
G8542	323	258	359	345	298	325
G8543	2007	1878	1502	1758	2480	1731
G8544	2480	1731	2007	1878	1502	1758
G8545	1652	1799	1645	254	383	258
G8546	298	241	342	81	150	298
G8547	2607	3394	2926	2755	3077	2227
G8548	2571	1929	1406	2439	1613	5074
G8549	121	22	55	730	201	35
G8550	1640	1693	1517	1731	1861	1550

Select n samples and g genes

	S1	S2	S4	S5	S6
G8523	680	749	669	724	643
G8524	262	311	1677	1286	1486
G8528	2571	1929	2439	1613	5074
G8530	1640	1693	1731	1861	1550
G8537	4077	2557	3394	2926	2755
G8545	1652	1799	254	383	258
G8547	2607	3394	2755	3077	2227

Compute similarity

Similarity	S1	S2	S3	S4	S5	S6
S1	0	6	7	7	0	0
S2	6	0	5	5	1	1
S3	7	5	0	8	0	0
S4	7	5	8	0	2	2
S5	0	2	0	2	0	10
S6	0	2	0	2	10	0

Final Clusters

ABC

1. Draw a random sample of N samples with replacement; discard replicates.
2. Rank the variances of the genes from 1 (most variable) to G and use the ranks to determine weights for the genes: $W_g=1/(R_g+c)$.
   - c is such that the 1% of genes with the highest variance have a combined probability of 20% of being selected
3. Draw a weighted random sample of √G genes without replacement.
4. Run Ward's clustering procedure on the resulting matrix to cluster the samples into √N clusters.
5. Repeat these steps many times.

ABC

- Collate the results: P_{ij} = proportion of runs in which the ith and jth samples cluster together.
- Interpretation:
  - P_{ij} large ⇒ ith and jth samples close
  - P_{ij} small ⇒ ith and jth samples far

  → P_{ij} : measure of sample similarity
- ABC dissimilarity measure: $D_{ij} = 1 - P_{ij}$
- ABC-based clustering: use $\{D_{ij}\}$ in a standard clustering procedure (such as Ward's).

Limitations

Cluster analyses:

- Usually outside the normal framework of statistical inference;
- less appropriate when only a few genes are likely to change.
- Needs lots of experiments
- Always possible to cluster even if there is nothing going on.
- Useful for learning about the data, but does not provide biological truth.

Single gene tests:

- may be too noisy in general to show much
- may not reveal coordinated effects of positively correlated genes.
- hard to relate to pathways.

Not relevant to our course

Software

- Spectral map
  - library(mpm)
- k-means
  - library(amap)
  - kmeans(stats)
- Hierarchical clustering
  - hclust (stats)
  - library(cluster)
  - library(clValid)
- ABC
  - http://www.geocities.com/damaratung/

The number of clusters

The GAP statistic

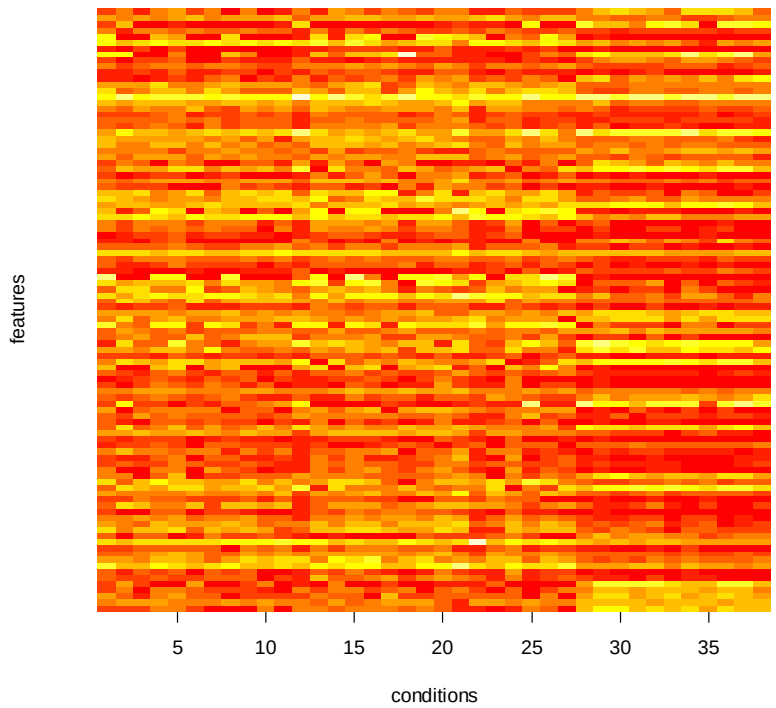
Tibshirani, Walther and Hastie, 2001

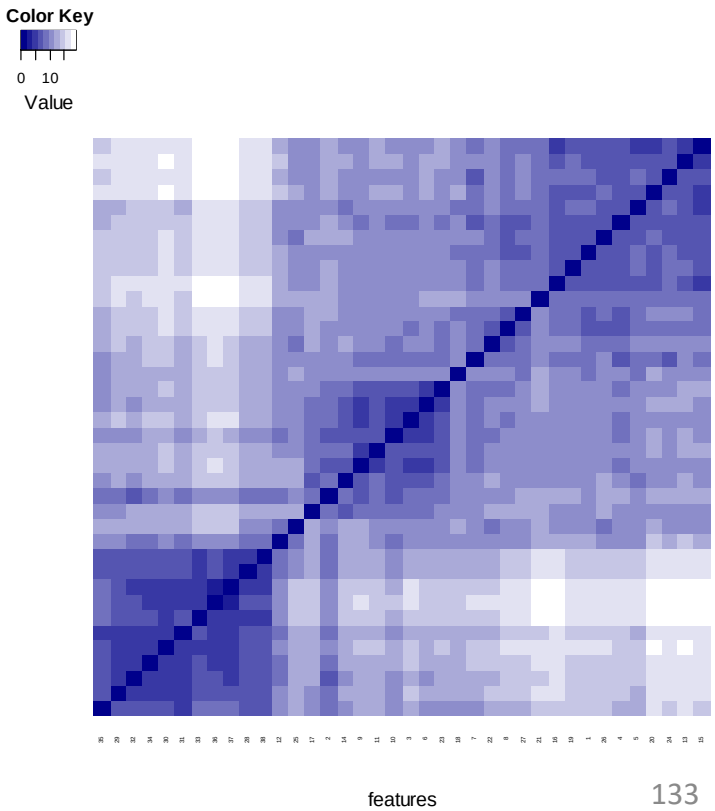
Software

- Paper:
  - Estimating the number of cluster in a dataset via the GAP statistic (Tibshirani et al. 2001)
- The R package:
  - `clusterGenomics`

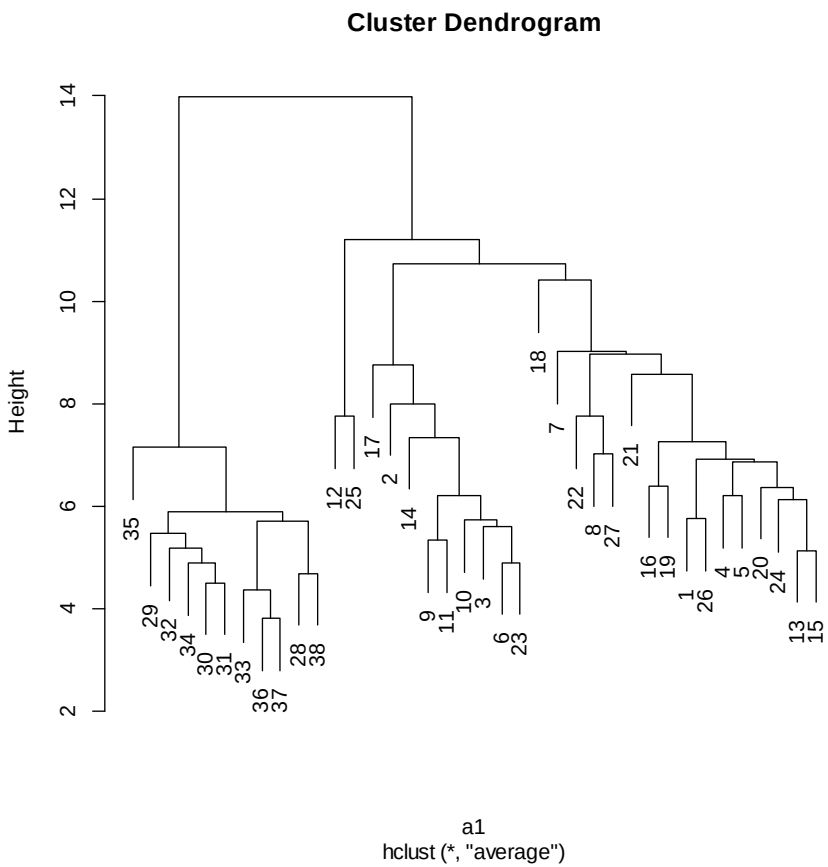
Example: the golub data

- For illustration: use top 100 genes.





Hierarchical clustering



- Euclidean distance.
- Average linkage.
- Three clusters ??

```
> a1 <- dist(t(data.g), method="euclidean",diag=TRUE, upper=TRUE)
> hc1 <- hclust(a1, method="ave")
> par(mfrow=c(1,1))
> plot(hc1,cex=0.35)
```

The GAP statistic

Assume that the data is clustered into k clusters

$$C_1, C_2, \ldots, C_k$$

C_r: the indices of the observations in cluster r.

Number of observations in the r'th cluster:

$$n_r = |C_r|$$

The GAP statistic

The sum of pairwise distances within cluster r:

$$D_r = \sum_{i,i^* \in C_r} d_{ii^*}$$

For d=squared Euclidean distance, for k clusters, the pooled within cluster sum of squares is

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

W_k decreases as the number of clusters increases.

The Gap statistic

$$Gap(k) = E\left(\log(W_k)\right) - \log(W_k)$$

The expected value of $\log(W_k)$ for sample size n from a reference distribution.

The observed value for the data.

$\hat{k}$:the estimate number of clusters is the value of k that maximize Gap(k).

The reference destruction

- A distribution with one cluster (i.e. no structure).
  - Normal distribution.
  - Uniform distribution.

$$Gap(k) = E\left(\log(W_k)\right) - \log(W_k)$$

The expected value of $\log(W_k)$ for sample size n from a reference distribution.

$E(\log(W_k))$ will be large for the reference distribution since there are no clusters associated with this distribution.

Calculation of the Gap statistics

- For each feature, generate a sample from the reference destruction (over the range of the observed data).
- Repeat B times.
- For a given number of clusters, k, for each generated datasets calculate

$$\log(W_{kb}^{*})$$

- For B datasets:

$$\log(W_{k1}^{*}), \log(W_{k2}^{*}), \ldots, \log(W_{kB}^{*})$$

Calculation of the Gap statistics

- For each feature, generate a sample from the reference destruction (over the range of the observed data).

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix}.$$

$$X_{\min} = \min\{X_{21}, \dots, X_{2m}\}$$

$$X_{\max} = \min\{X_{21}, \dots, X_{2m}\}$$

$$X_{2i}^{*} \sim U(X_{\min}, X_{\max})$$

The reference distribution

Calculation of the Gap statistics

- Estimate E(log(W_k))

$$\hat{E}\left(\log(W_k^*)\right) = \frac{1}{B}\sum_{b=1}^{B}\log(W_{kb}^*)$$

- Estimate the standard error of E(log(W_k))

$$sd_k = \left[\frac{1}{B}\sum_{b=1}^{B}\left\{\log(W_{kb}^*) - \hat{E}\left(\log(W_k^*)\right)\right\}^2\right]^{0.5}$$

Calculation of the Gap statistics

- Estimate the Gap statistic for a given value of k:

$$Gap(k) = \hat{E}\left(\log(W_k^*)\right) - \log(W_k)$$

The choice of k

- For k=1,2,3,…,K calculate

$$Gap(k) = \hat{E}\left(\log(W_k^*)\right) - \log(W_k)$$

and

$$s_k = sd_k \sqrt{\left(1 + \frac{1}{B}\right)}$$

- Choose the value first value of k for which

$$Gap(k) \geq Gap(k+1) - s_{k+1} \Rightarrow Gap(k) - Gap(k+1) > s_{k+1}$$

The Gap statistic in R

```
library(clusterGenomics)
```

```
res <- gap(t(data.g),
                  cl.method="hclust",
                  dist.method="euclidean",
                  linkage="ave",
                  Kmax=15,
                  B=100)
```

- Hierarchical clustering
- Euclidean distance.
- Average linkage.
- Explore maximum 15 clusters.
- Generate 100 datasets for each number of clusters.

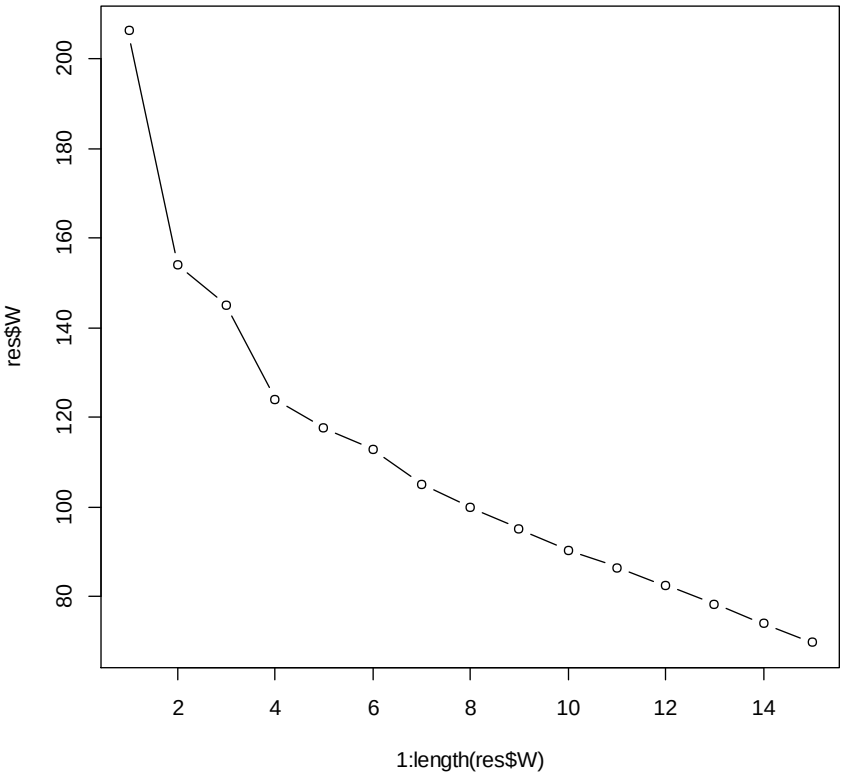
Output

```
> res
$hatK
[1] 2
```

$\hat{k} = 2$

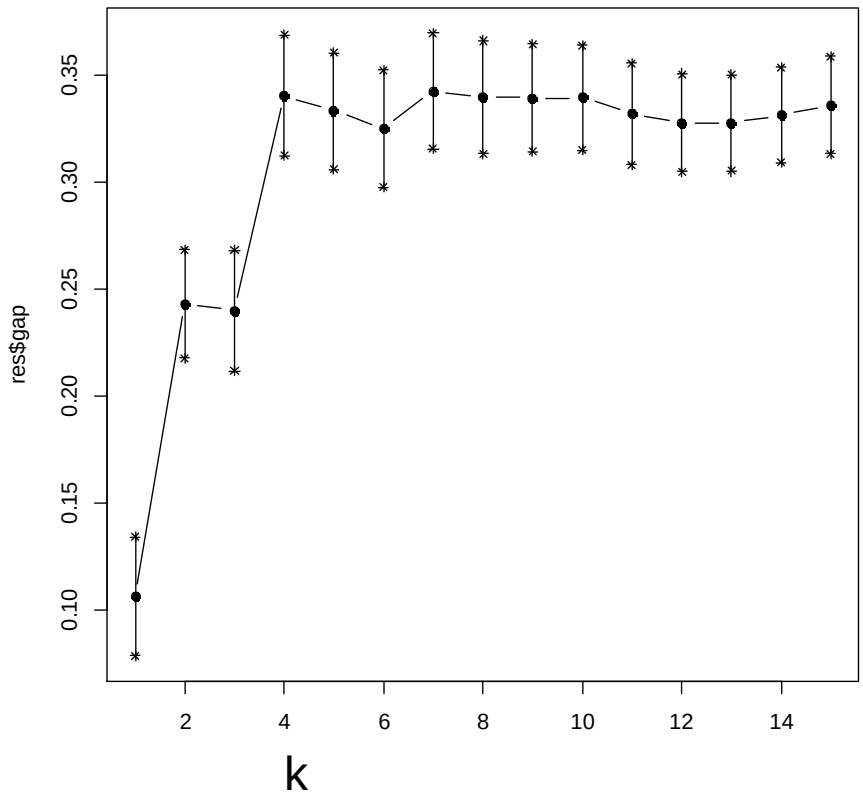
```
$lab.hatK
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2

$gap
 [1] 0.1061658 0.2428449 0.2397402 0.3403534 0.3329261 0.3247911 0.3424079
 [8] 0.3395133 0.3392024 0.3392275 0.3316401 0.3276111 0.3275321 0.3312073
[15] 0.3358039

$sk
 [1] 0.02774142 0.02535067 0.02814635 0.02819704 0.02713994 0.02731280
 [7] 0.02701233 0.02631137 0.02501484 0.02453088 0.02377241 0.02279634
[13] 0.02258421 0.02215460 0.02284534

$W
 [1] 206.29826 153.95149 145.07900 124.01425 117.76156 112.73660 105.18057
 [8] 100.02671  95.03616  90.39992  86.52438  82.53053  78.33906  73.94802
[15]  69.72588
```

Within cluster sum of squares



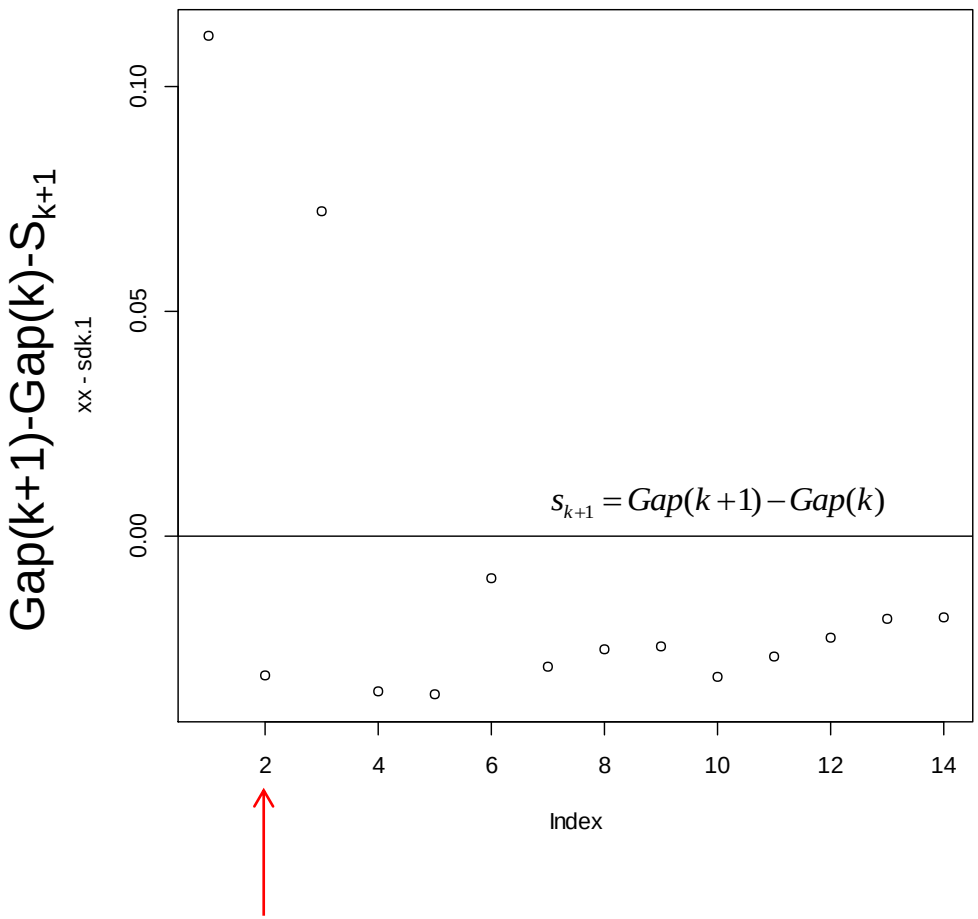
- The pooled within sum of squares decreases as the number of clusters increases.

The GAP statistic



$$Gap(k) = \hat{E}\left(\log(W_k^*)\right) - \log(W_k)$$

How many clusters ?



The number of clusters is the smallest number of k such that:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

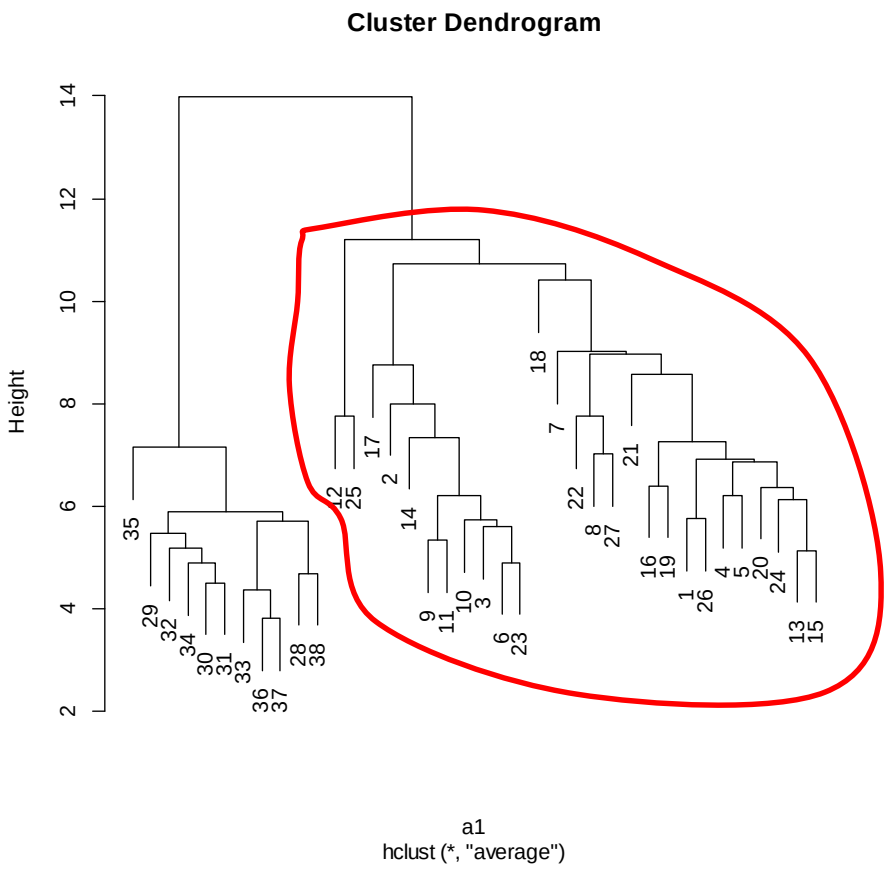
or

$$s_{k+1} \geq Gap(k+1) - Gap(k)$$

or

$$0 \geq Gap(k+1) - Gap(k) - s_{k+1}$$

Hierarchical clustering



- Euclidean distance.
- Average linkage.
- Three clusters ??

$$\hat{k} = 2$$