Module 19 Multivariate Analysis for Genetic data
Session 08 Cluster Analysis

Jan Graffelman

jan.graffelman@upc.edu

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

28th Summer Institute in Statistical Genetics (SISG 2023)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Contents

Objectives

Goals:

- Discover "natural" groups of cases (or variables) in the data.
- Data reduction: from n cases to $m \ll n$ clusters

Considerations:

- The number of clusters may a priori be unkown.
- There is no categorical variable that defines the grouping.
- Cluster analysis is an exploratory tool.

Ingredients

- Distance measure
  - In order to cluster item or variables we need a measure of similarity (proximity) or distance (dissimilarity).
- Algorithm
  - We cannot consider all possible groupings and need algorithms to produce the grouping.

Algorithms

- Hierarchical methods: cases can not change group
  - agglomerative (most common)
  - divisive
- Partitioning methods: cases can change group
  - K-means
- Model based methods
- Other

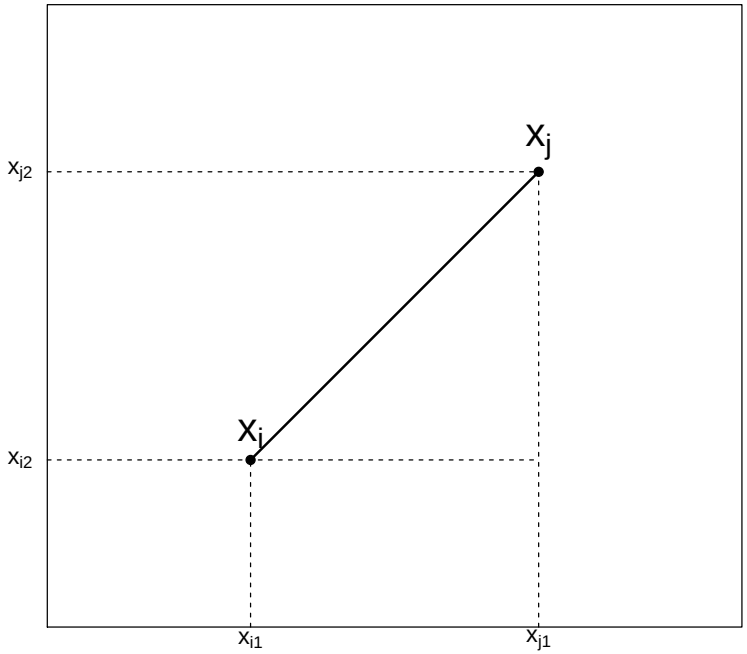
Distance measure for quantitative variables

$$\mathbf{x}' = (x_1, x_2, \ldots, x_p) \quad \mathbf{y}' = (y_1, y_2, \ldots, y_p)$$

Euclidian distance:	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$
Weighted Euclidian distance:	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{A}(\mathbf{x}-\mathbf{y})}$
Mahalanobis distance:	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$
Manhattan distance:	$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \lvert x_i - y_i \rvert$
Minkowski distance:	$d(\mathbf{x}, \mathbf{y}, \lambda) = \left(\sum_{i=1}^{p} \lvert x_i - y_i \rvert^{\lambda} \right)^{1/\lambda}$
Canberra distance:	$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \frac{\lvert x_i - y_i \rvert}{x_i + y_i}$
Bray-Curtis distance:	$d(\mathbf{x}, \mathbf{y}) = \frac{1}{P} \frac{\sum_{i=1}^{P} \lvert x_i - y_i \rvert}{\sum_{i=1}^{P} (x_i + y_i)}$

...

Euclidean distance



In two dimensions:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{j1} - x_{i1})^2 + (x_{j2} - x_{i2})^2} = \sqrt{(\mathbf{x}_j - \mathbf{x}_i)'(\mathbf{x}_j - \mathbf{x}_i)}$$

In p dimensions:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{j1} - x_{i1})^2 + (x_{j2} - x_{i2})^2 + \cdots + (x_{jp} - x_{ip})^2} = \sqrt{(\mathbf{x}_j - \mathbf{x}_i)'(\mathbf{x}_j - \mathbf{x}_i)}$$

Similarity measures for qualitative variables

		case j		
		1	0	
case i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	p=a+b+c+d

$\frac{a+d}{p}$ simple matching coefficient

$\frac{a}{p}$ only one-one matches

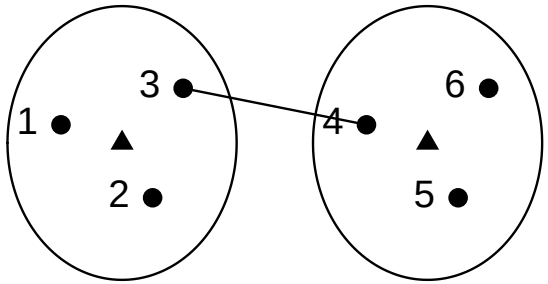
$\frac{a}{a+b+c}$ Jaccard's coefficient (0-0 irrelevant)

Example

	indicators					
case 1	1	1	0	0	1	1
case 2	0	1	1	0	0	1

- Compute the squared Euclidean distance between the cases
- What does this distance represent?

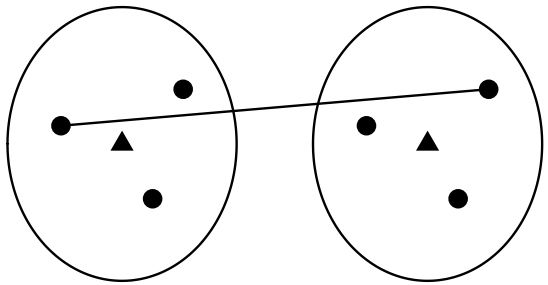
Cluster distance



Single linkage

Nearest neighbour

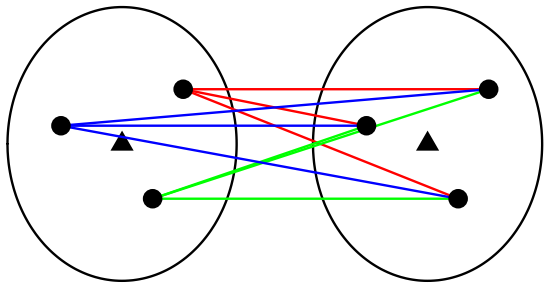
d_{34}



Complete linkage

Farthest neighbour

d_{16}



Average linkage

$$\frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I} d_{ij}$$

Criteria for joining clusters

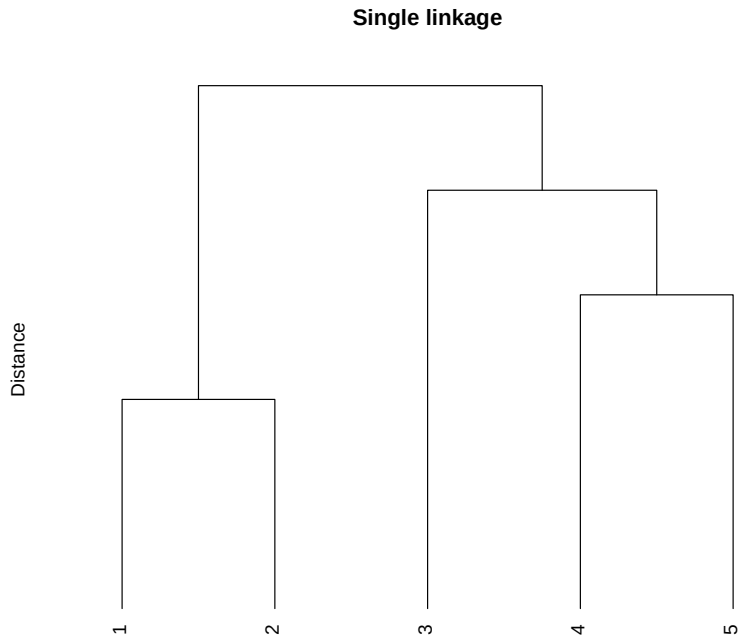
- single linkage
- complete linkage
- average linkage
- centroid distance $d_{rs}^2 = \sum_{j=1}^{p}(\bar{x}_{rj} - \bar{x}_{sj})^2$
  (UPGMA, Unweighted Pair Group Method using Averages)
- Ward's incremental sum-of-squares

Miniature example: hierarchical agglomerative

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

Distance	Clusters
0	1,2,3,4,5
2	(1,2),3,4,5
3	(1,2),3,(4,5)
4	(1,2),(3,4,5)
5	(1,2,3,4,5)

Dendrogram



Miniature example: continuation

D_0	1	2	3	4	5
1	0	2	6	10	9
2	2	0	5	9	8
3	6	5	0	4	5
4	10	9	4	0	3
5	9	8	5	3	0

D_1	(1,2)	3	4	5
(1,2)	0	5	9	8
3	5	0	4	5
4	9	4	0	3
5	8	5	3	0

D_2	(1,2)	3	(4,5)
(1,2)	0	5	8
3	5	0	4
(4,5)	8	4	0

D3	(1,2)	(3,4,5)
(1,2)	0	5
(3,4,5)	5	0

D_4	(1,2,3,4,5)
(1,2,3,4,5)	0

Exercise

Given the distance matrix

	1	2	3	4	5
1	0	10	27	15	19
2	10	0	18	6	8
3	27	18	0	16	12
4	15	6	16	0	7
5	19	8	12	7	0

Write down the successive formation of clusters according to the complete linkage criterion

Ward's criterion

Given two clusters r and s we have the within-group sums-of-squares

$$WSS_r = \sum_{j=1}^{p} \sum_{i=1}^{n_r} (x_{ij} - \bar{x}_j)^2 \qquad WSS_s = \sum_{j=1}^{p} \sum_{i=1}^{n_s} (x_{ij} - \bar{x}_j)^2$$

On joining, an new cluster t is obtained with a new WSS:

$$WSS_t = \sum_{j=1}^{p} \sum_{i=1}^{n_r+n_s} (x_{ij} - \bar{x}_j)^2$$

This gives an increase in WSS:

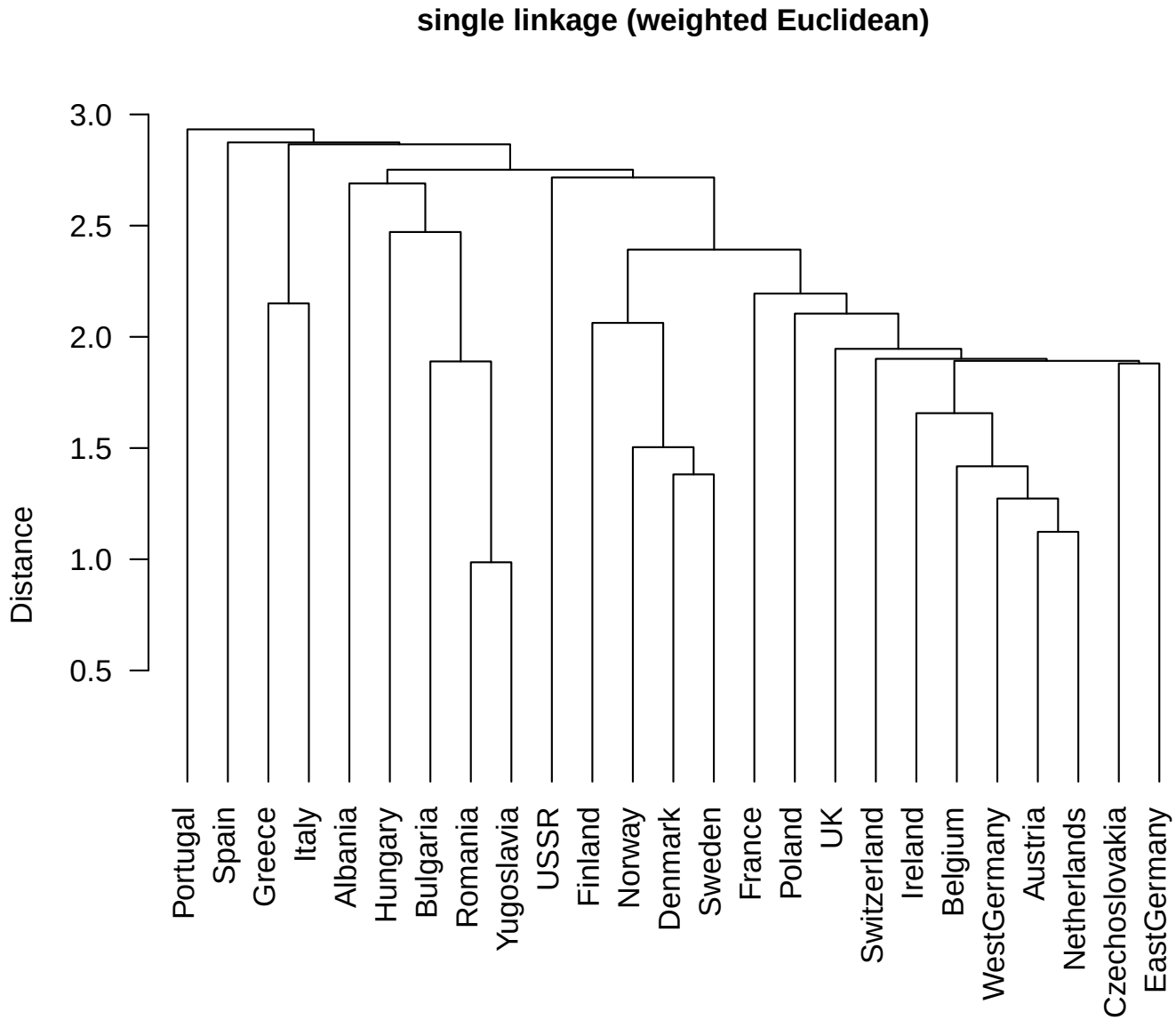
$$\Delta = WSS_t - (WSS_r + WSS_s) = \frac{n_r n_s}{n_r + n_s} d_{rs}^2$$

Join those two clusters for which Δ is minimal.

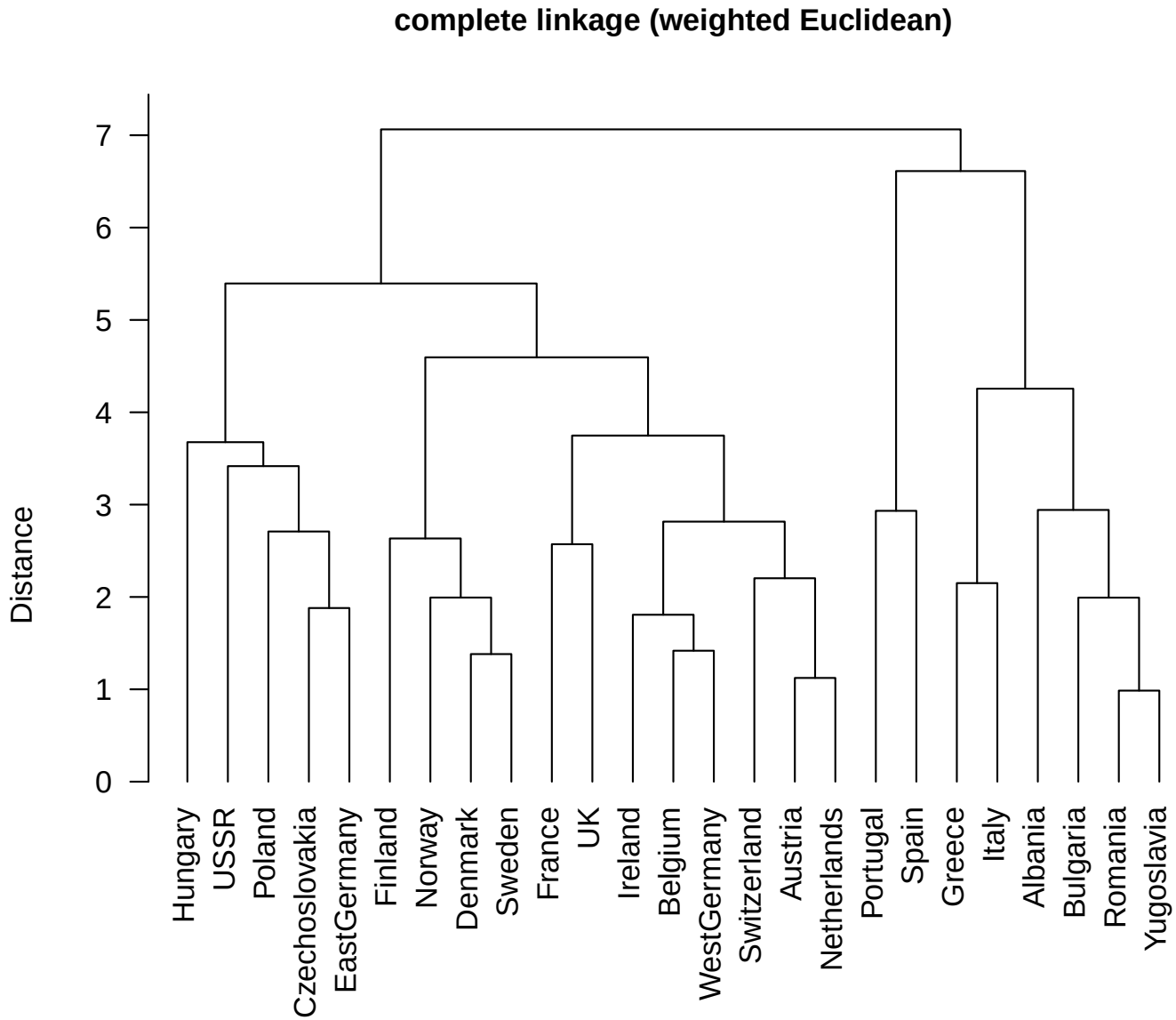
Example: protein consumption data

Country	Red meat	White meat	Eggs	Milk	Fish	Cereals	Starchy foods	Pulses, Nuts Oilseeds	Fruit Vegetables
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Denmark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
EastGermany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Ireland	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italy	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
WestGermany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

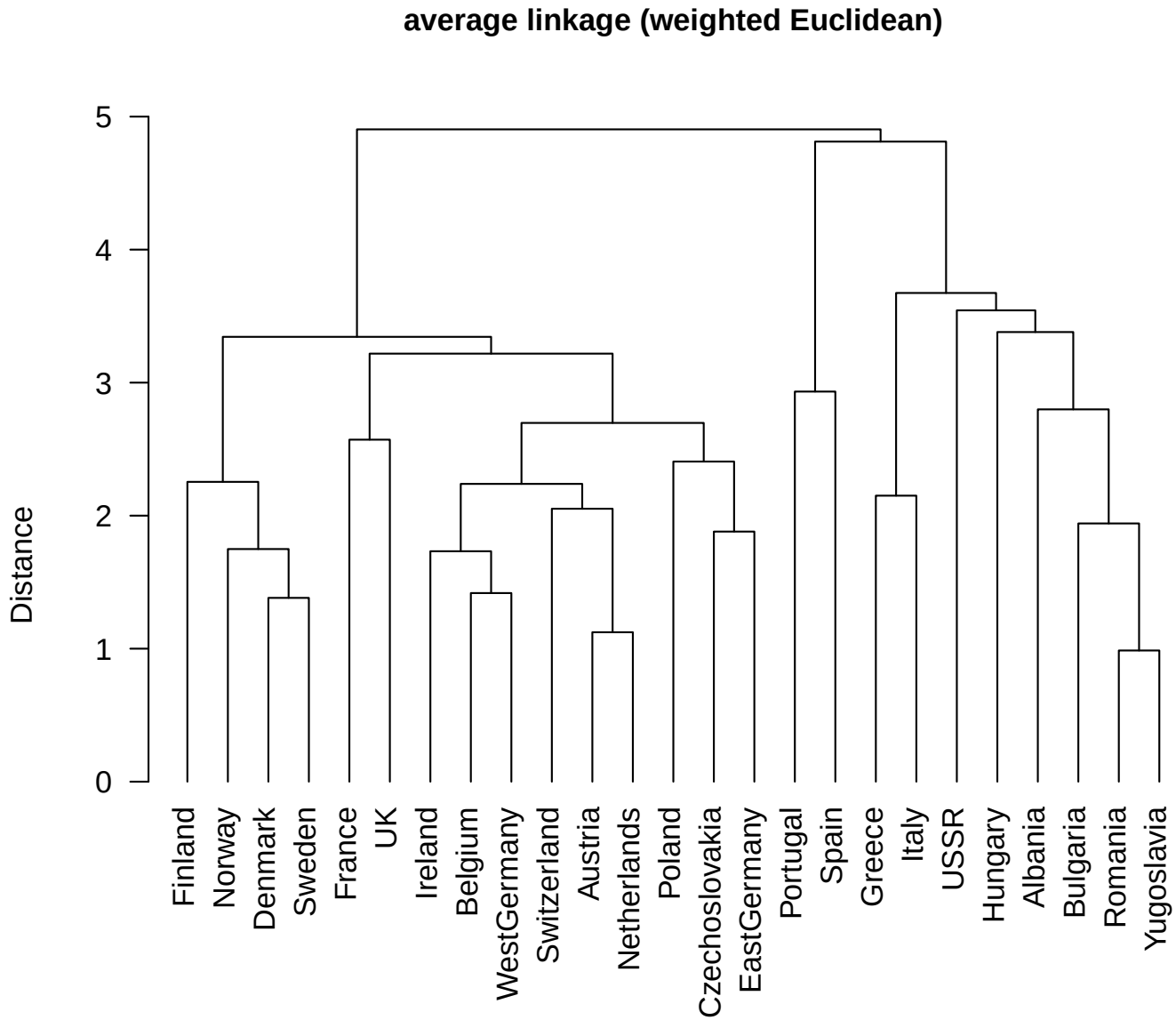
Dendrogram


single linkage (weighted Euclidean)
Distance
3.0
2.5
2.0
1.5
1.0
0.5
Portugal
Spain
Greece
Italy
Albania
Hungary
Bulgaria
Romania
Yugoslavia
USSR
Finland
Norway
Denmark
Sweden
France
Poland
UK
Switzerland
Ireland
Belgium
WestGermany
Austria
Netherlands
Czechoslovakia
EastGermany

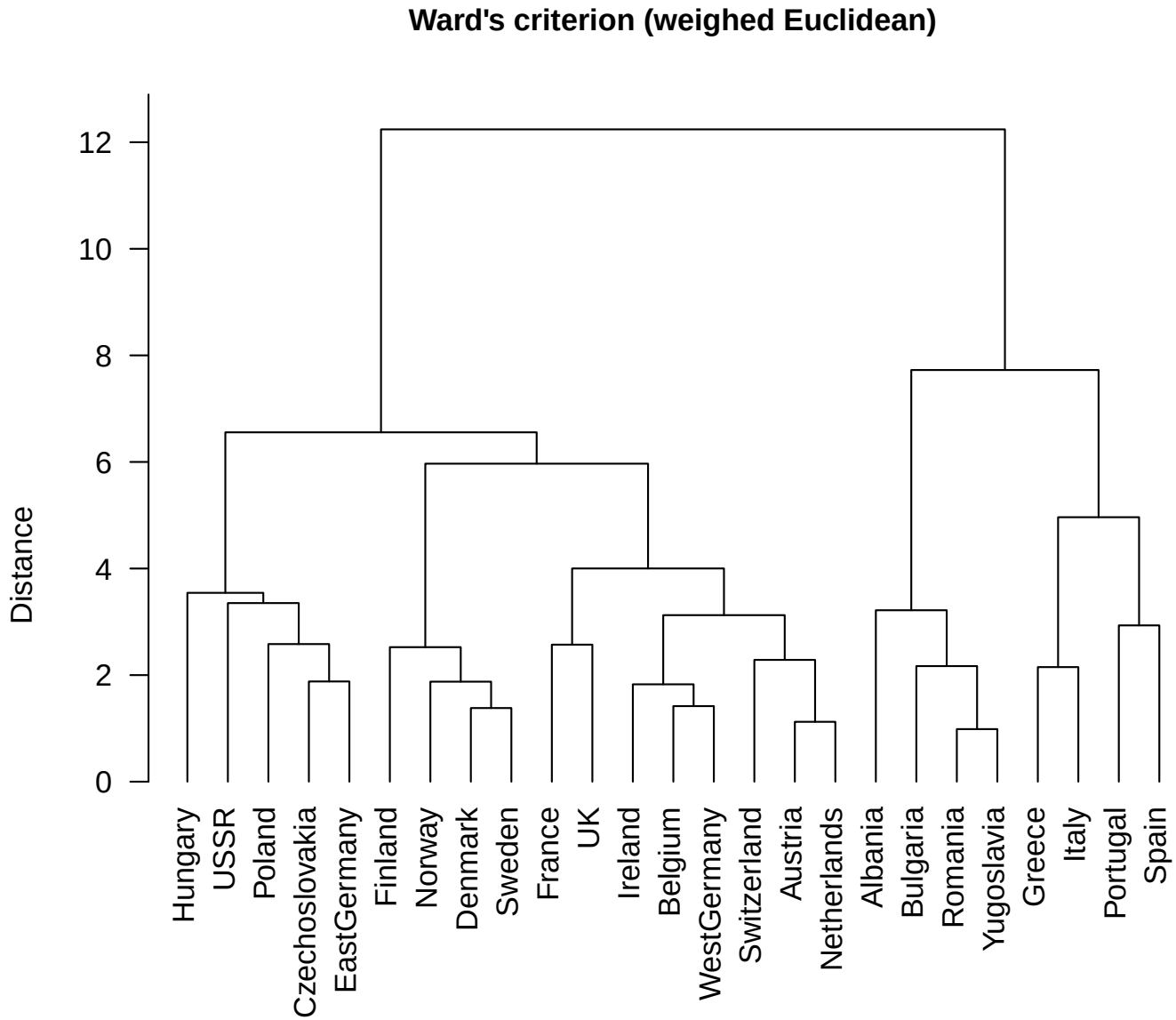
Dendrogram


complete linkage (weighted Euclidean)
Distance
0
1
2
3
4
5
6
7
Hungary
USSR
Poland
Czechoslovakia
EastGermany
Finland
Norway
Denmark
Sweden
France
UK
Ireland
Belgium
WestGermany
Switzerland
Austria
Netherlands
Portugal
Spain
Greece
Italy
Albania
Bulgaria
Romania
Yugoslavia

Dendrogram


average linkage (weighted Euclidean)
Distance
0
1
2
3
4
5
Finland
Norway
Denmark
Sweden
France
UK
Ireland
Belgium
WestGermany
Switzerland
Austria
Netherlands
Poland
Czechoslovakia
EastGermany
Portugal
Spain
Greece
Italy
USSR
Hungary
Albania
Bulgaria
Romania
Yugoslavia

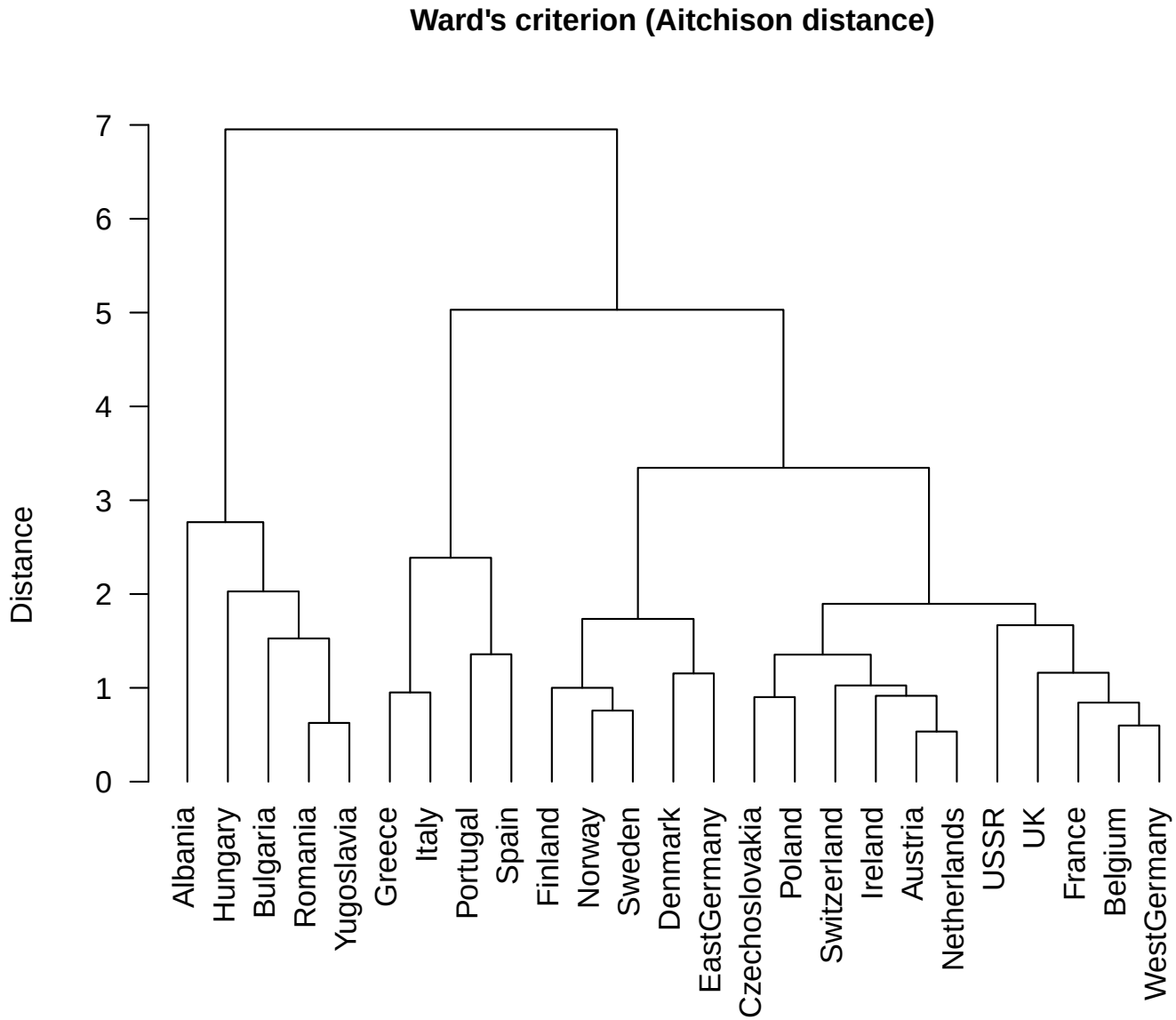
Dendrogram


Ward's criterion (weighed Euclidean)
Distance
0
2
4
6
8
10
12
Hungary
USSR
Poland
Czechoslovakia
EastGermany
Finland
Norway
Denmark
Sweden
France
UK
Ireland
Belgium
WestGermany
Switzerland
Austria
Netherlands
Albania
Bulgaria
Romania
Yugoslavia
Greece
Italy
Portugal
Spain

A compositional note

- The protein data set can be seen as a compositional data set
- Apply closure by dividing each row by its sum.
- Transform by taking log-ratios
- Compute the Euclidean distance of the transformed data (Aitchison distance)
- Cluster with this new distance matrix.

Dendrogram


Ward's criterion (Aitchison distance)
Distance
0
1
2
3
4
5
6
7
Albania
Hungary
Bulgaria
Romania
Yugoslavia
Greece
Italy
Portugal
Spain
Finland
Norway
Sweden
Denmark
EastGermany
Czechoslovakia
Poland
Switzerland
Ireland
Austria
Netherlands
USSR
UK
France
Belgium
WestGermany

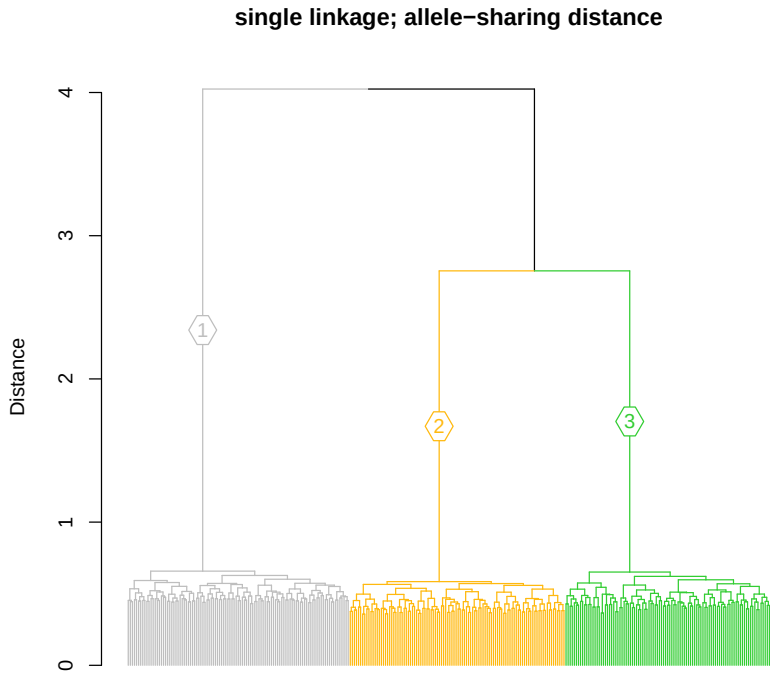
Some considerations on cluster distance

single linkage	late inclusion of outliers
	can identify chain-like clusters
	sensitive to outliers
complete linkage	fast inclusion of outliers
	sensitive to outliers
average linkage	less sensitive to outliers
centroid distance	less sensitive to outliers
Ward's criterion	less sensitive to outliers
	tends to form equally sized clusters

Genetic examples: 1,000 genomes populations

- Subset of 1,000 Genomes individuals (CEU, JPT and YRI).
- 500,000 SNPs from CHR 20; MAF > 0.05
- Allele sharing distance, Ward's criterion

	CEU	JPT	YRI
1	0	0	107
2	0	104	0
3	99	0	0

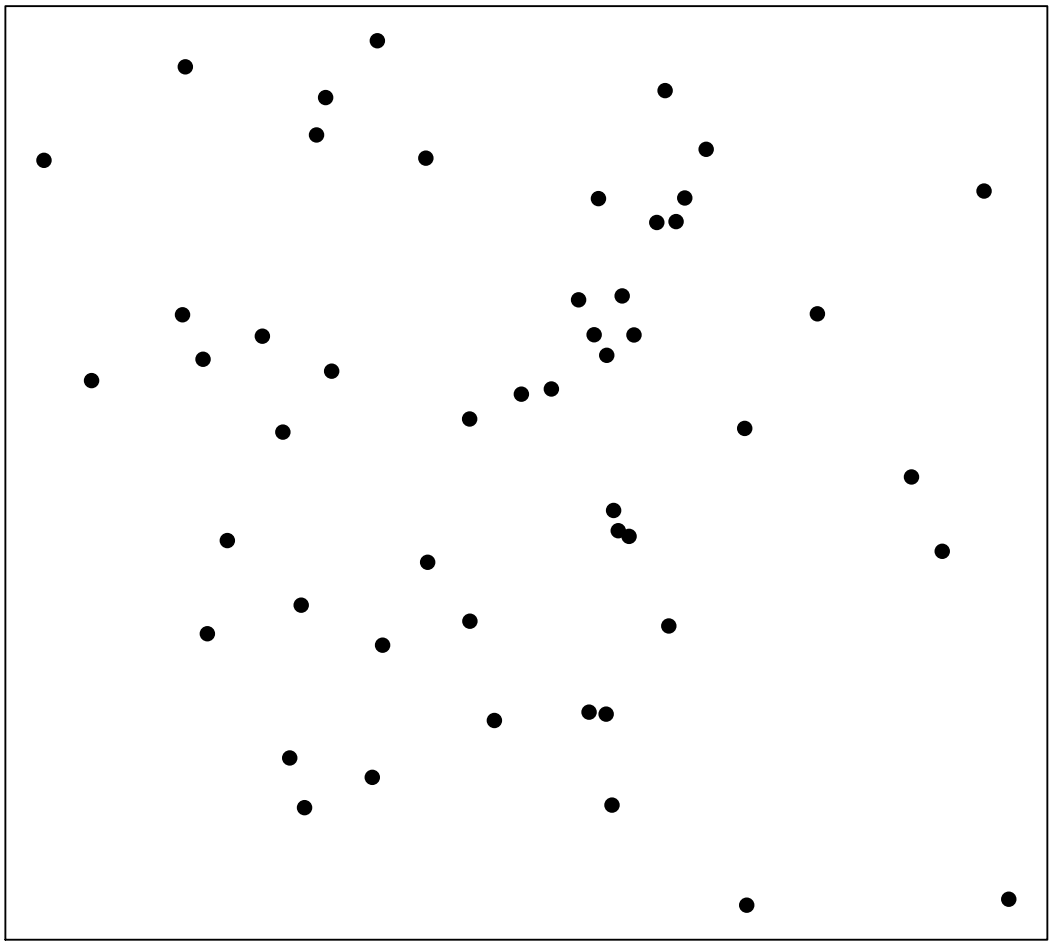


Non-hierarchical Clustering: K-means method

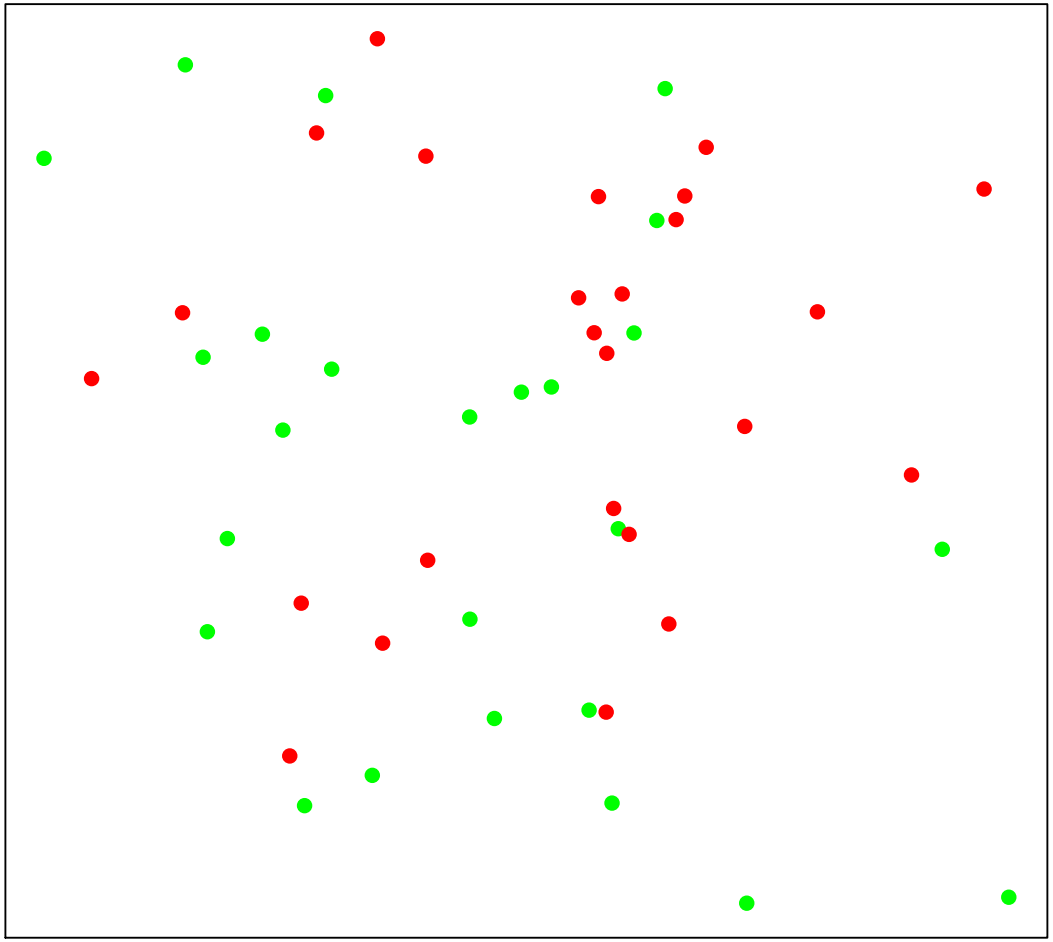
Algorithm:

1. Choose a value for the number of clusters K.
2. Partition all items into K initial clusters (at random or using seeds).
3. Compute the centroids of each cluster.
4. Assign each item to the cluster whose centroid is nearest.
5. Go back to 3, until there are no re-assignments.

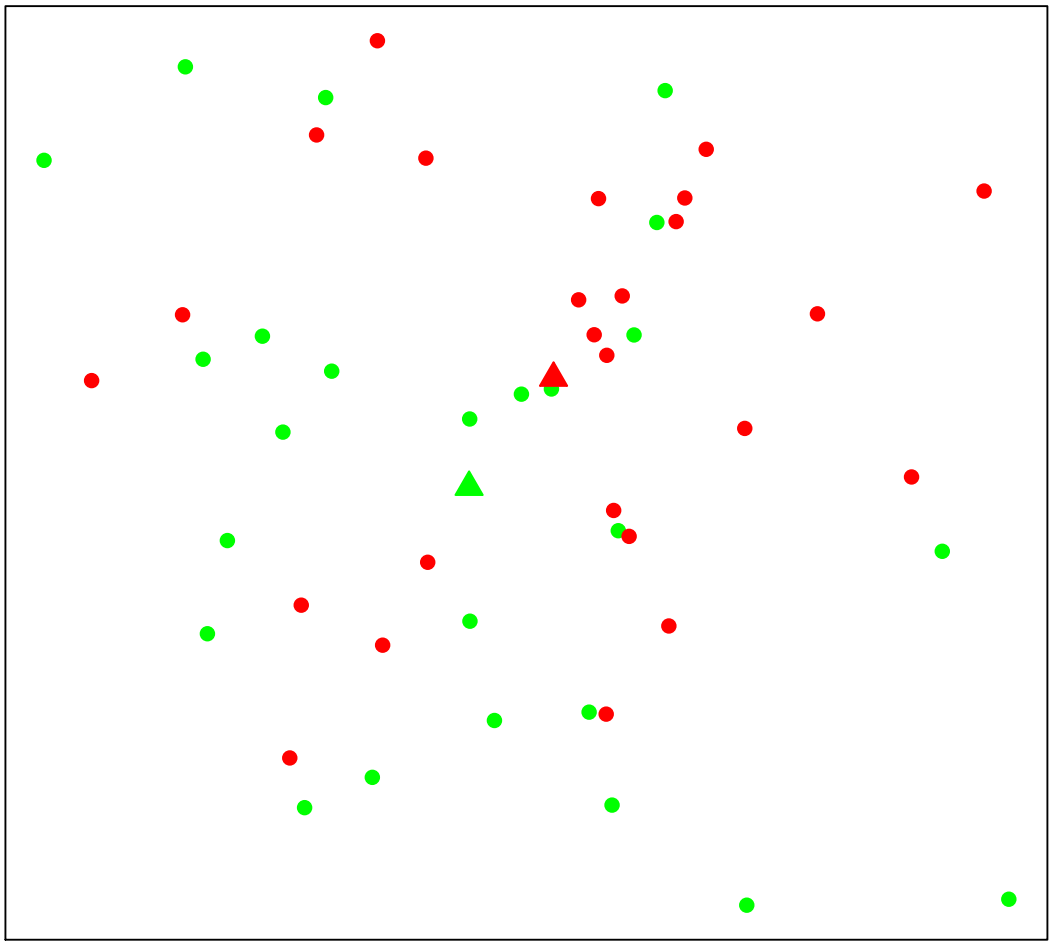
K means graphical illustration



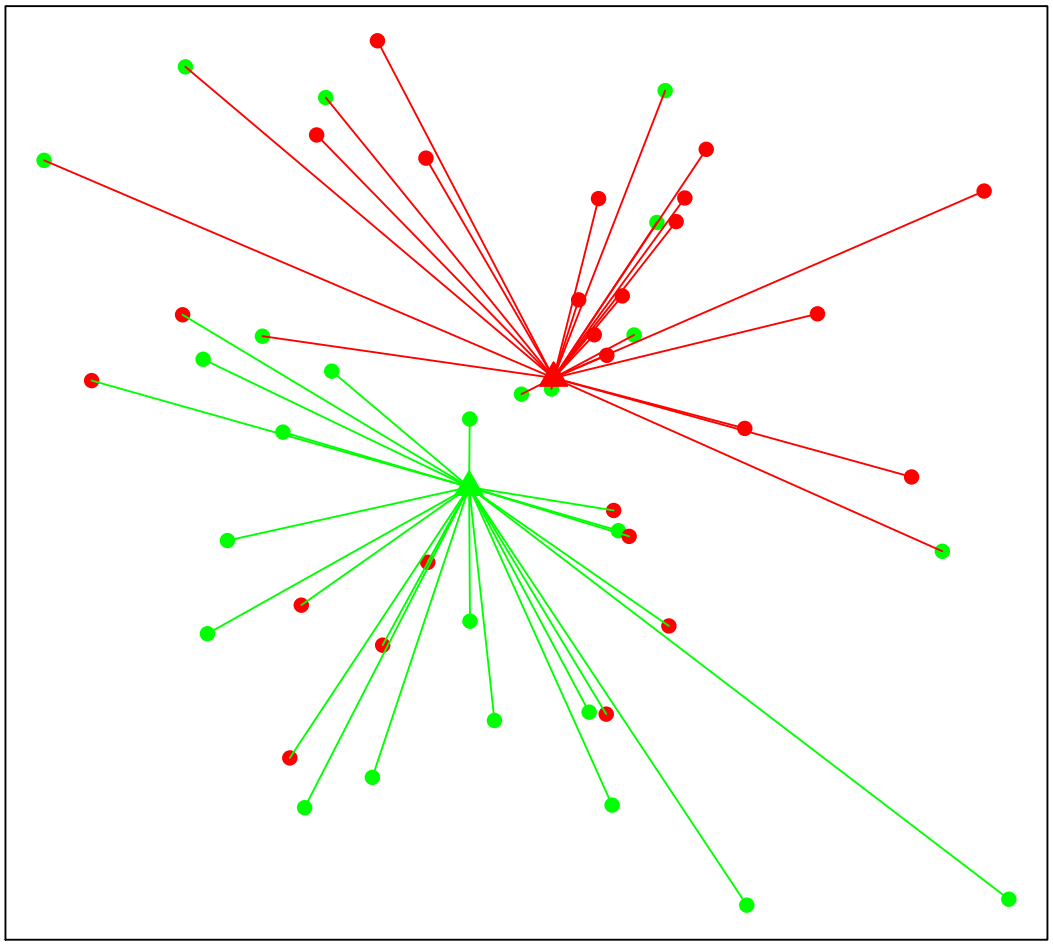
K means graphical illustration



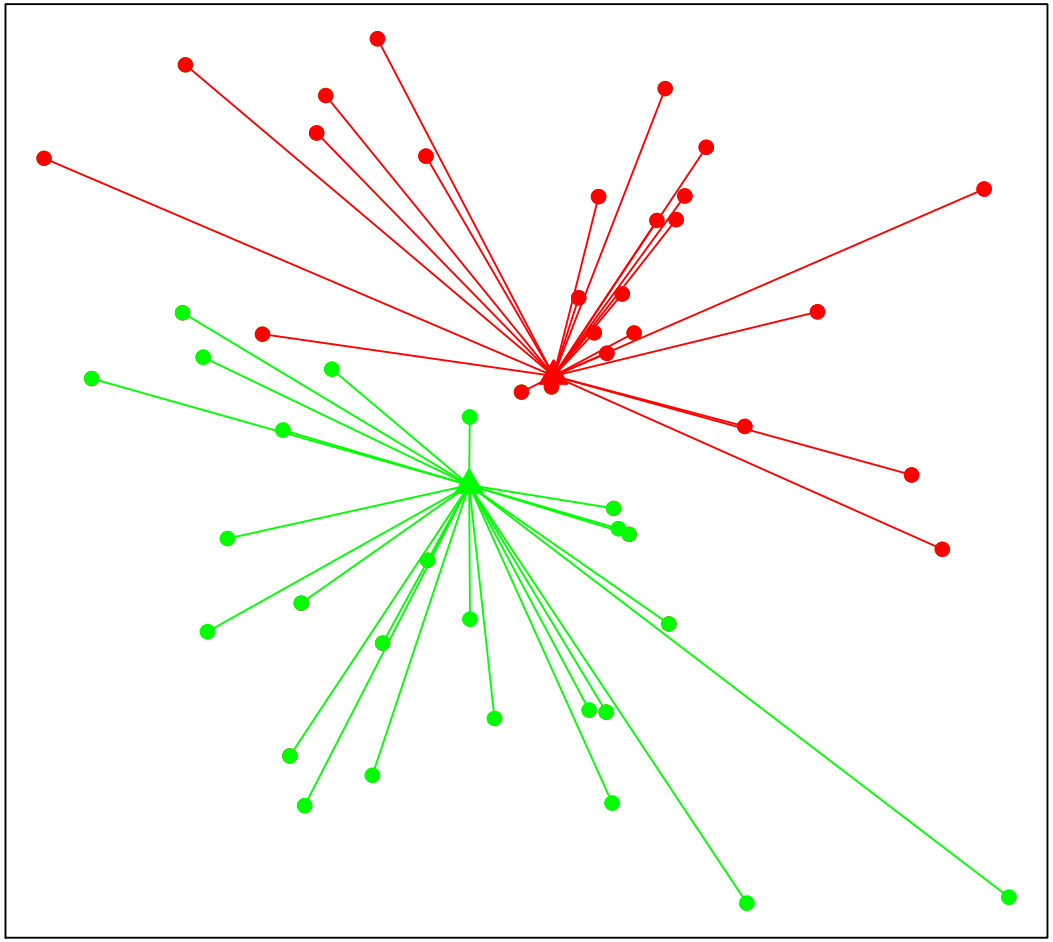
K means graphical illustration



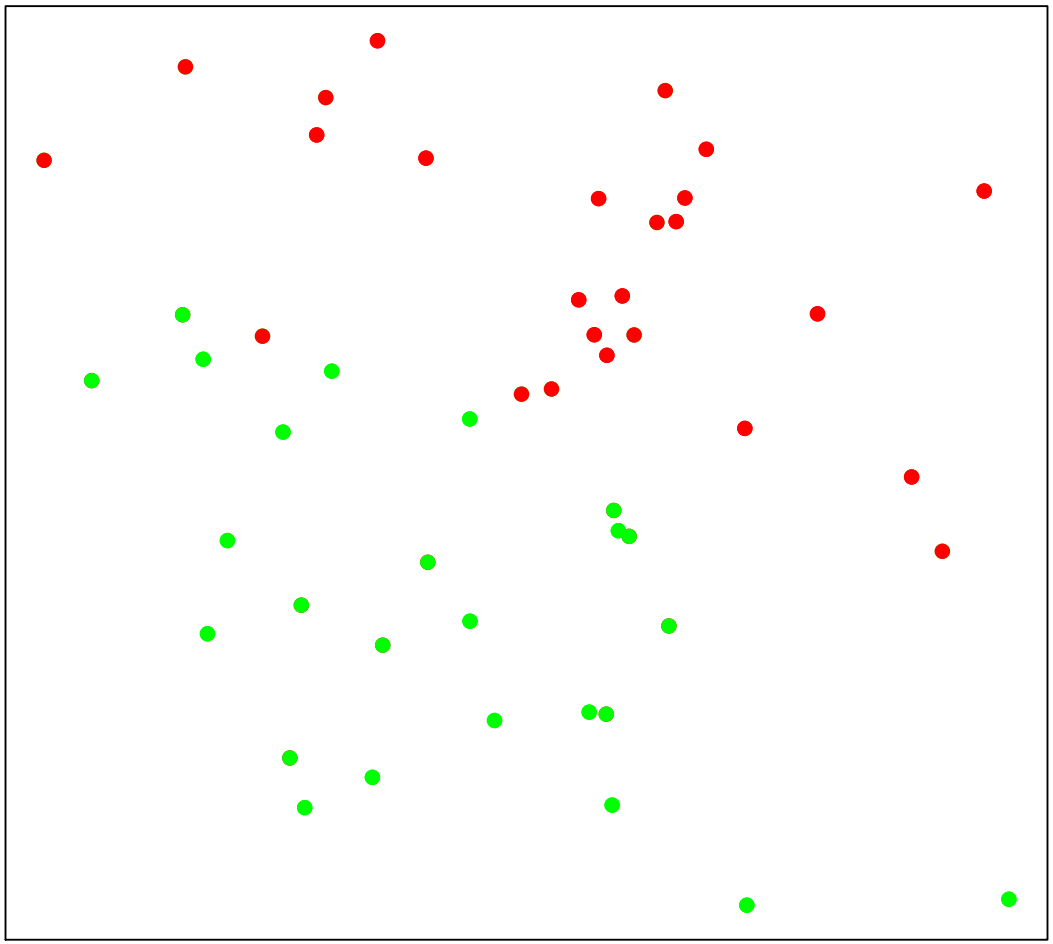
K means graphical illustration



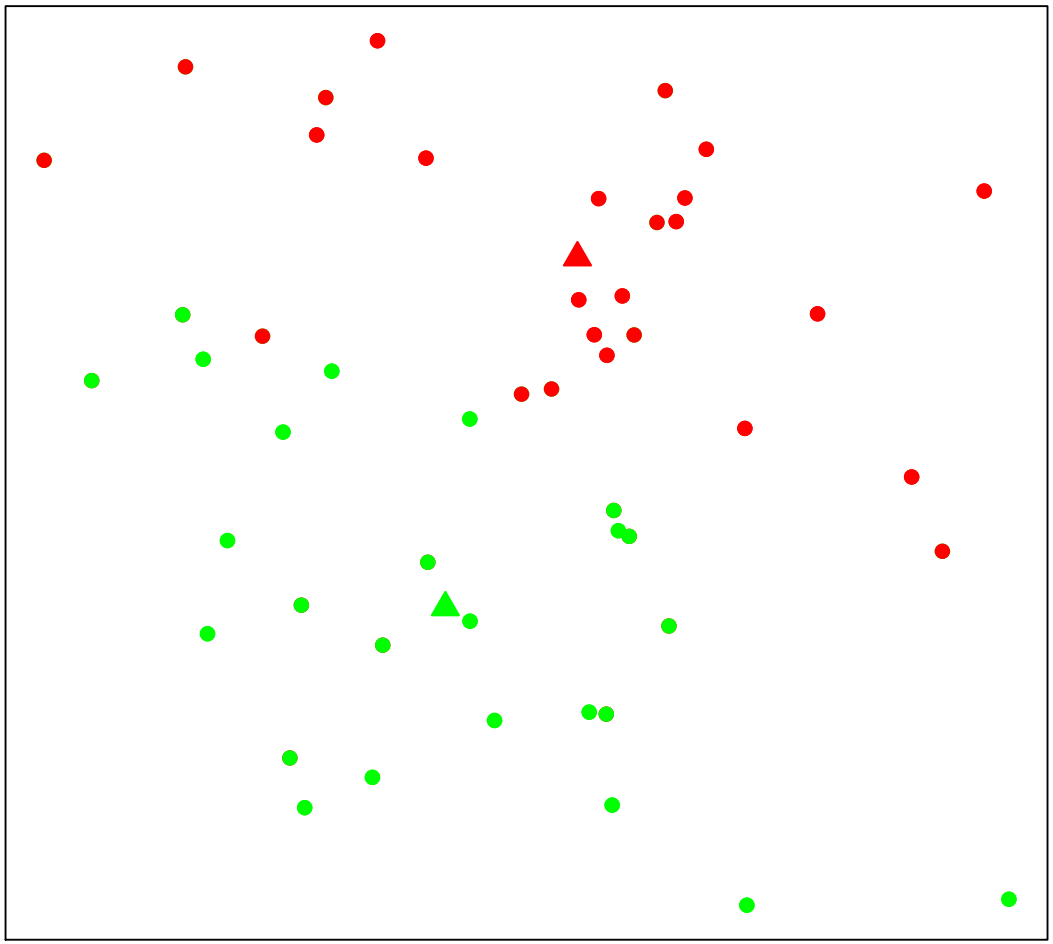
K means graphical illustration



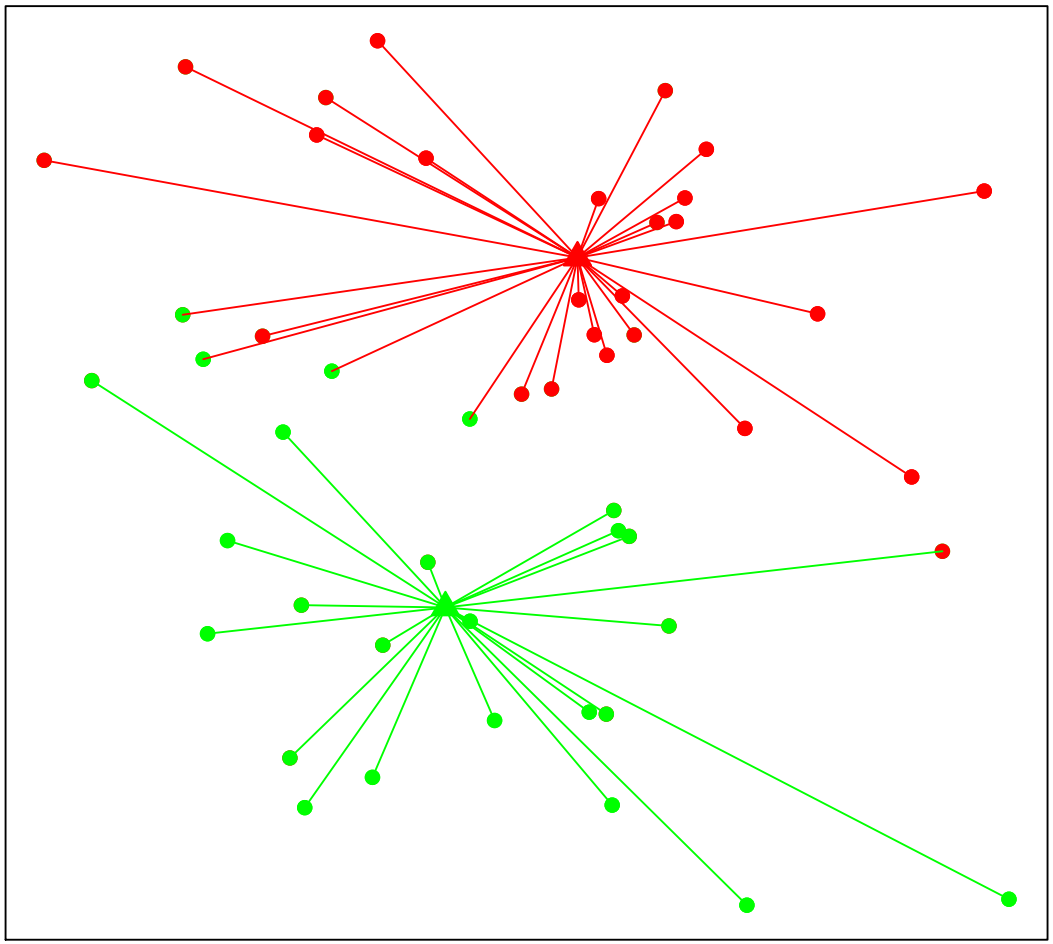
K means graphical illustration



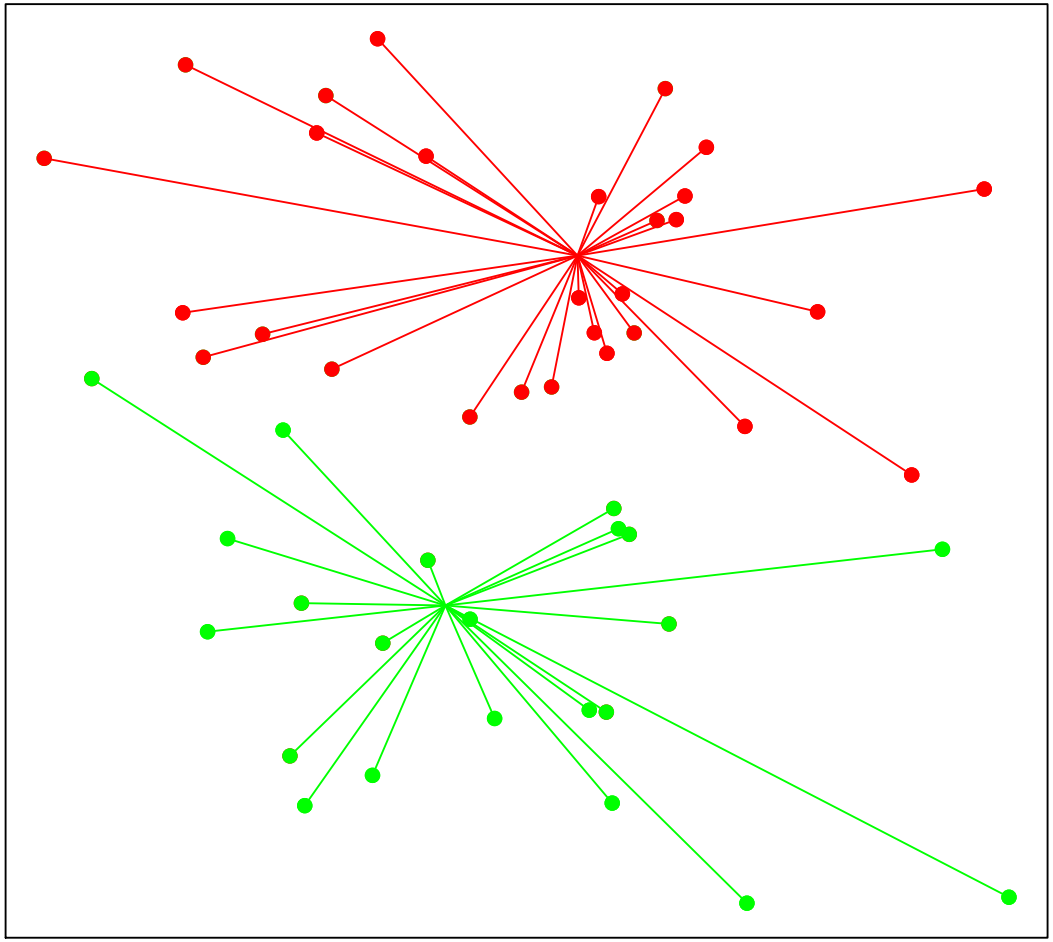
K means graphical illustration



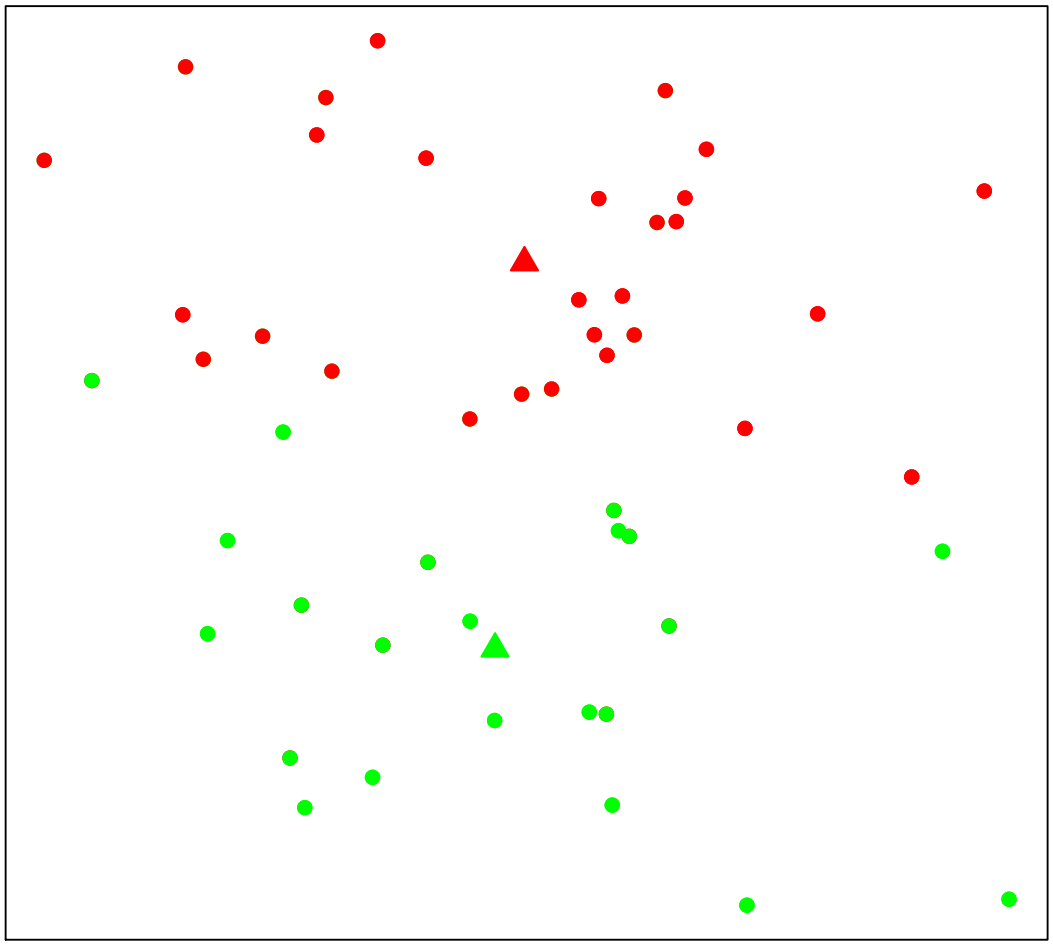
K means graphical illustration



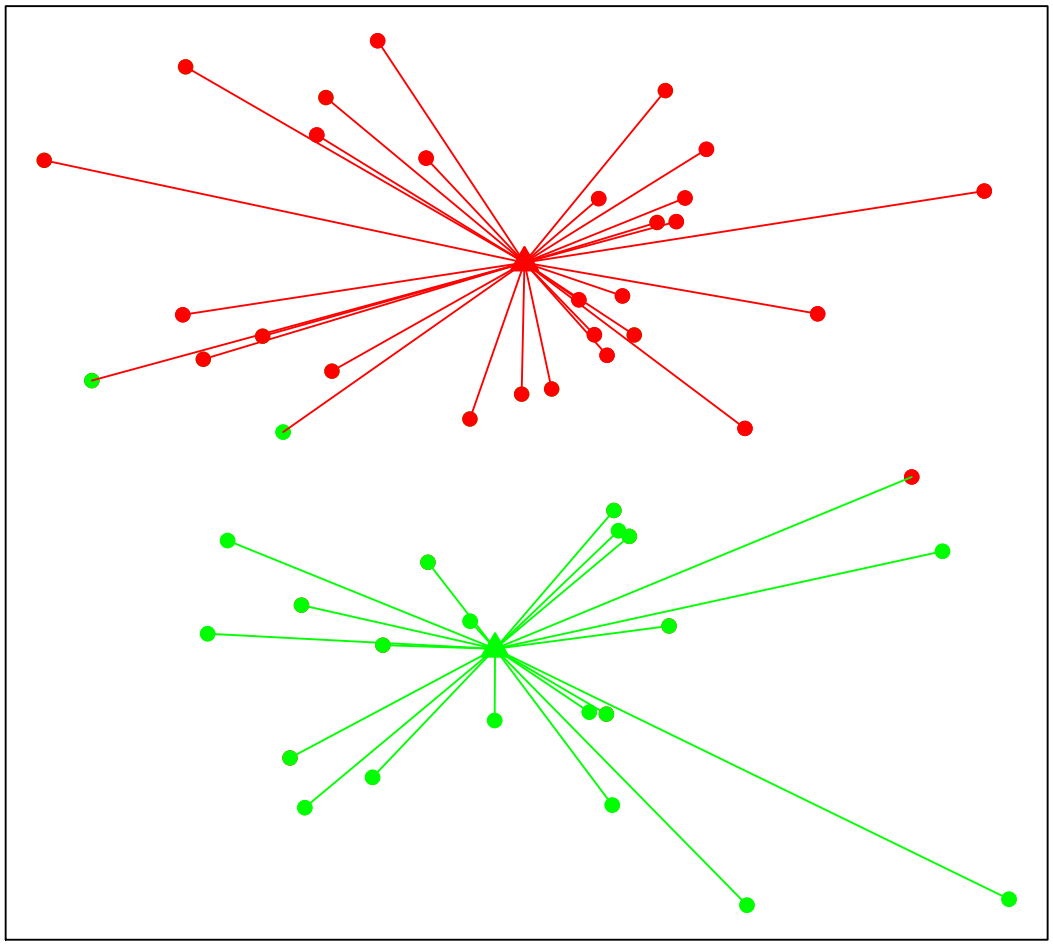
K means graphical illustration



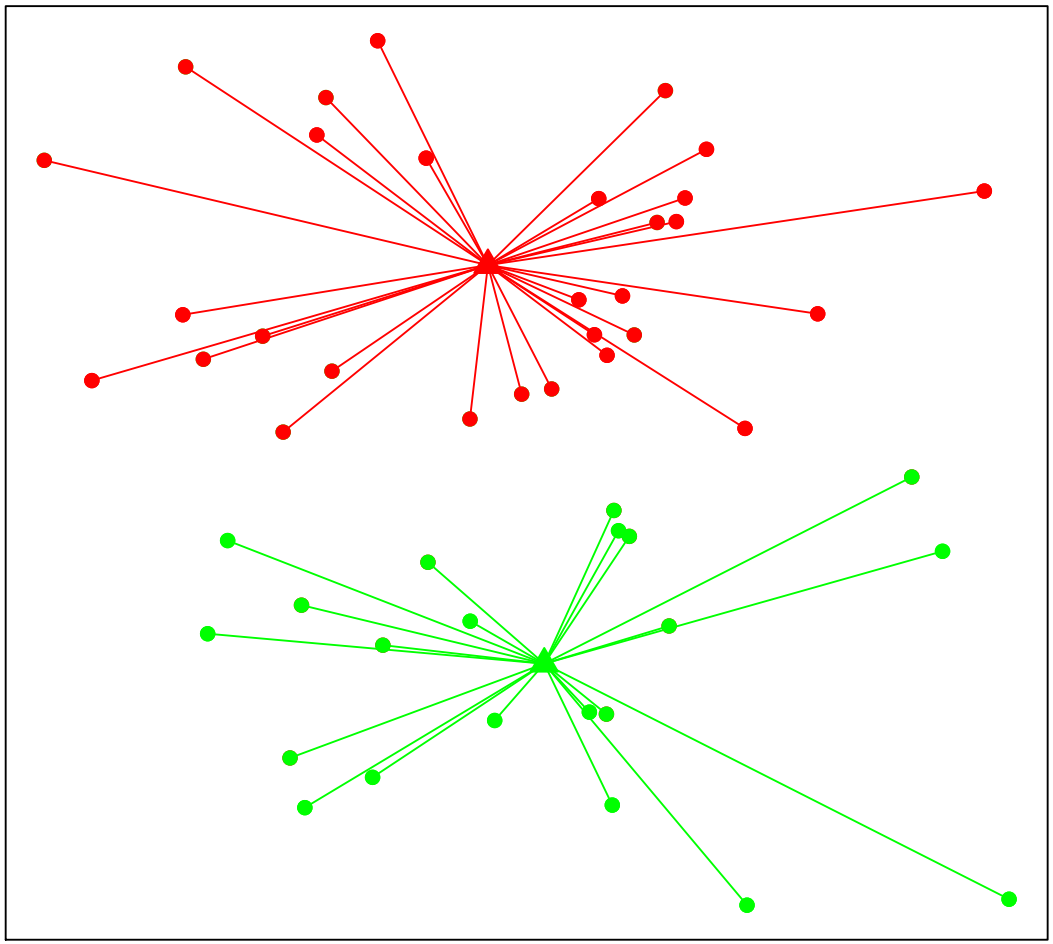
K means graphical illustration



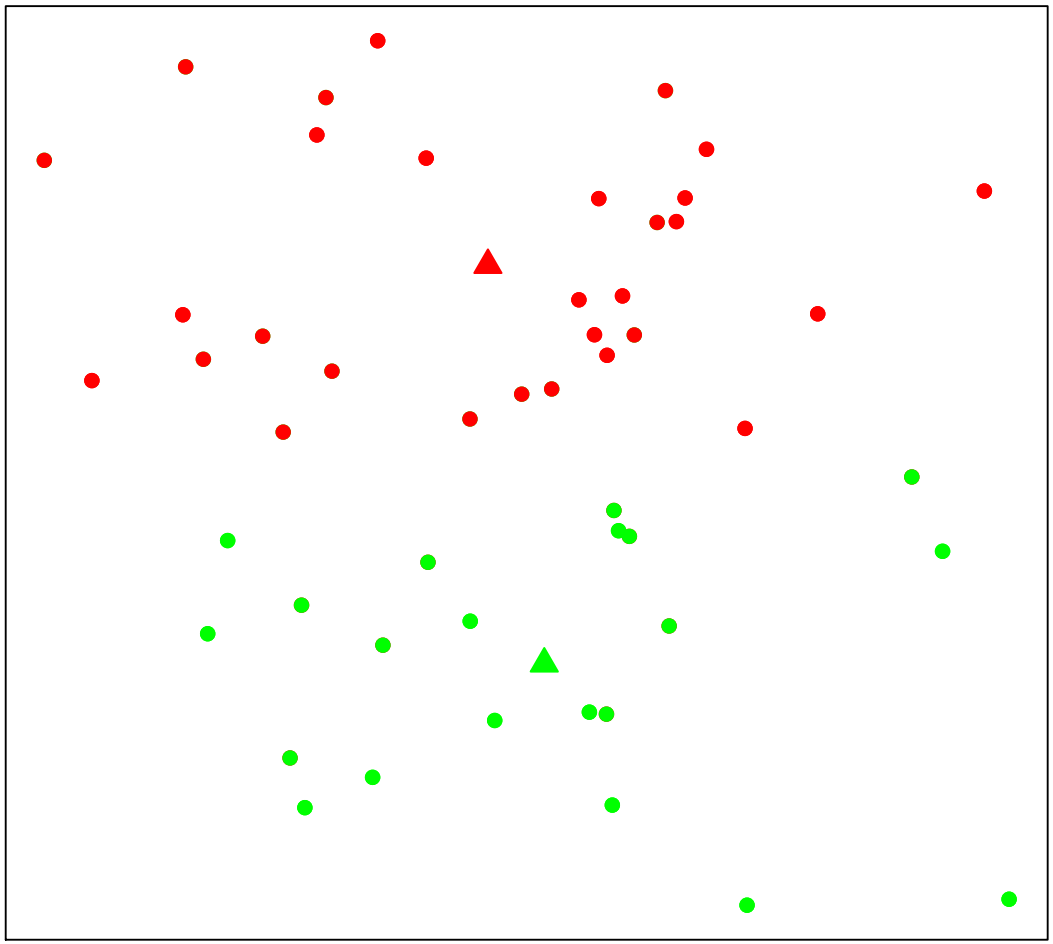
K means graphical illustration



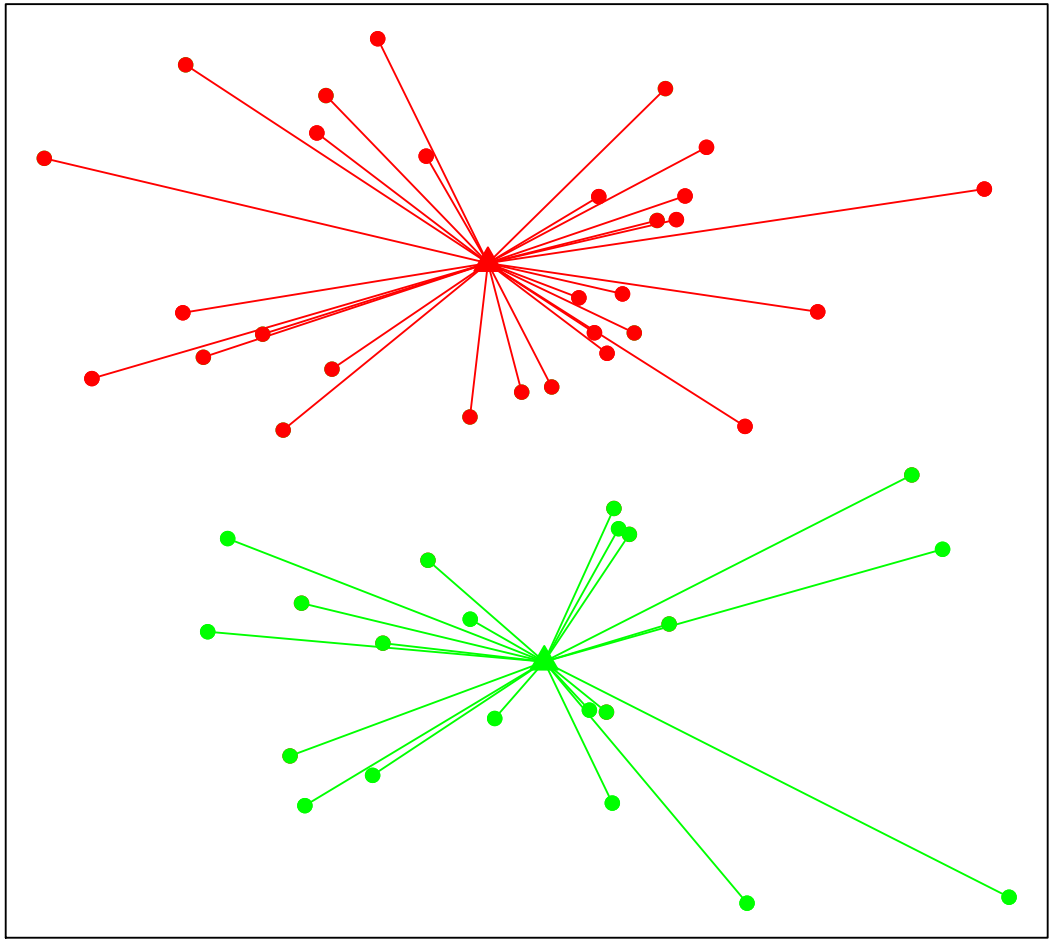
K means graphical illustration



K means graphical illustration



K means graphical illustration

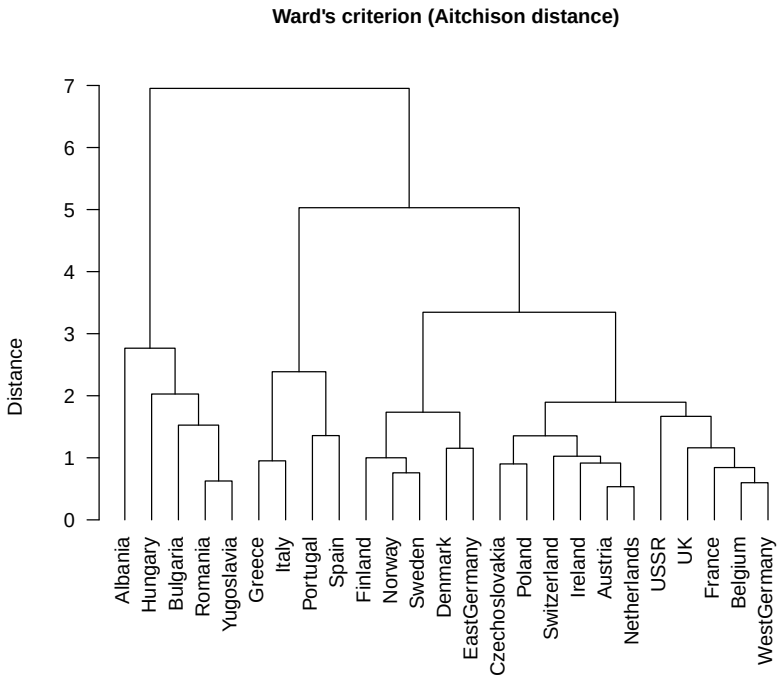


K-means

After convergence, it is recommended to

- Try different initial clusters, and compare the final clusters obtained.
- Try different numbers of clusters K.
- Compare cluster means and within-cluster variances.

Example: K-means of protein consumption data



We tentatively try $k = 4$

	Redm	Whim	Eggs	Milk	Fish	Cereals	Starchy	Nuts	FruVeg
1	0.46	0.21	-1.12	0.84	-2.03	2.47	-0.77	0.22	-0.29
2	0.46	0.21	-0.62	1.31	0.21	1.34	-0.21	-1.75	-0.96
3	0.55	0.38	-0.64	1.08	-0.79	1.43	-0.32	-1.21	-0.48
4	0.20	-0.56	-1.04	0.43	0.03	1.63	-0.61	-0.16	0.08

	1	2	3	4
Size	5.00	5.00	11.00	4.00
SS	7.24	2.96	6.81	4.22
Var	1.81	0.74	0.68	1.41

	SS	%
WSS	21.2	33.4
BSS	42.4	66.6
TSS	63.6	100.0

	Dendrogram			
K-means	1	2	3	4
1	5	0	0	0
2	0	0	5	0
3	0	11	0	0
4	0	0	0	4

K-means 1,000 Genomes individuals (CEU, JPT and YRI).

- Subset of 1,000 Genomes individuals (CEU, JPT and YRI).
- 500,000 SNPs from CHR 20; MAF > 0.05
- mind missing values

	Cluster		
	1	2	3
Size	107	104	99
SS	2259563	1582909	1659549
Var	21317	15368	16934

	SS	%
WSS	5502021	83.2
BSS	1109336	16.8
TSS	6611357	100.0

	1	2	3
CEU	0	0	99
JPT	0	104	0
YRI	107	0	0

	$k = 2$	%	$k = 3$	%	$k = 4$	%
WSS	6085198	92.0	5502021	83.2	5469181	82.7
BSS	526159	8.0	1109336	16.8	1142176	17.3
TSS	6611357	100.0	6611357	100.0	6611357	100.0

K-means issues

- The result depends on the initial random assignment
- The result is not necessarily optimal
- Perform multiple runs, and choose the best run

Model-based clustering

- Previous approaches do not make any distributional assumptions
- Probabilistic models can be used in clustering and this is called model-based clustering
- Finite mixture model

$$g(x|\boldsymbol{\pi}, \boldsymbol{\theta}) = \pi_1 f_1(x|\boldsymbol{\theta}_1) + \pi_2 f_2(x|\boldsymbol{\theta}_2) + \cdots \pi_k f_k(x|\boldsymbol{\theta}_k)$$

- With $\pi_i > 0$ and $\sum_{i=1}^{k} \pi_i = 1$
- Each f_i is a probability distribution for the ith cluster.
- Usually $f_i \sim N(\mu, \boldsymbol{\Sigma})$, but not necessarily so.
- The posterior probabilities that observation x_j pertains to the ith cluster can be calculated

$$\frac{\pi_i f_i(x_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{k} \pi_i f_i(x_j|\boldsymbol{\theta}_i)}$$

Procedure

- A value or estimate of the number of clusters k is needed
- The finite mixture model is estimated by maximum likelihood
- For each observation, the posterior probabilities of pertaining to j cluster are calculated
- Each observation is assigned to the cluster for which it has the largest posterior probability

NIST autosomal STRs

	CSF1PO		D10S1248		D12S391		D13S317		D16S539		D18S51		D19S433		D1S1656	
1	11	12	14	14	17	21	11	12	11	11	17	18	14	14	12	...
2	9	11	12	13	16	16	11	13	11	12	16	16	11	15	14	...
3	10	12	14	15	16	20	12	12	11	12	12	17	13	13	16	...
4	11	12	11	14	16	19	12	13	9	12	15	17	14	16	10	...
5	8	12	14	14	15	18	11	12	9	11	17	19	14	14	14	...
6	12	12	14	16	18	19	11	13	11	12	15	16	14	15	14	...
.	.	.	.	.	.	.	.	.	.	.	.	.				
.	.	.	.	.	.	.	.	.	.	.	.	.				
.	.	.	.	.	.	.	.	.	.	.	.	.				

Hill, C. et al. (2013) U.S. population data for 29 autosomal STR loci. *Forensic science international: Genetics* 497 7(3):e82–3.

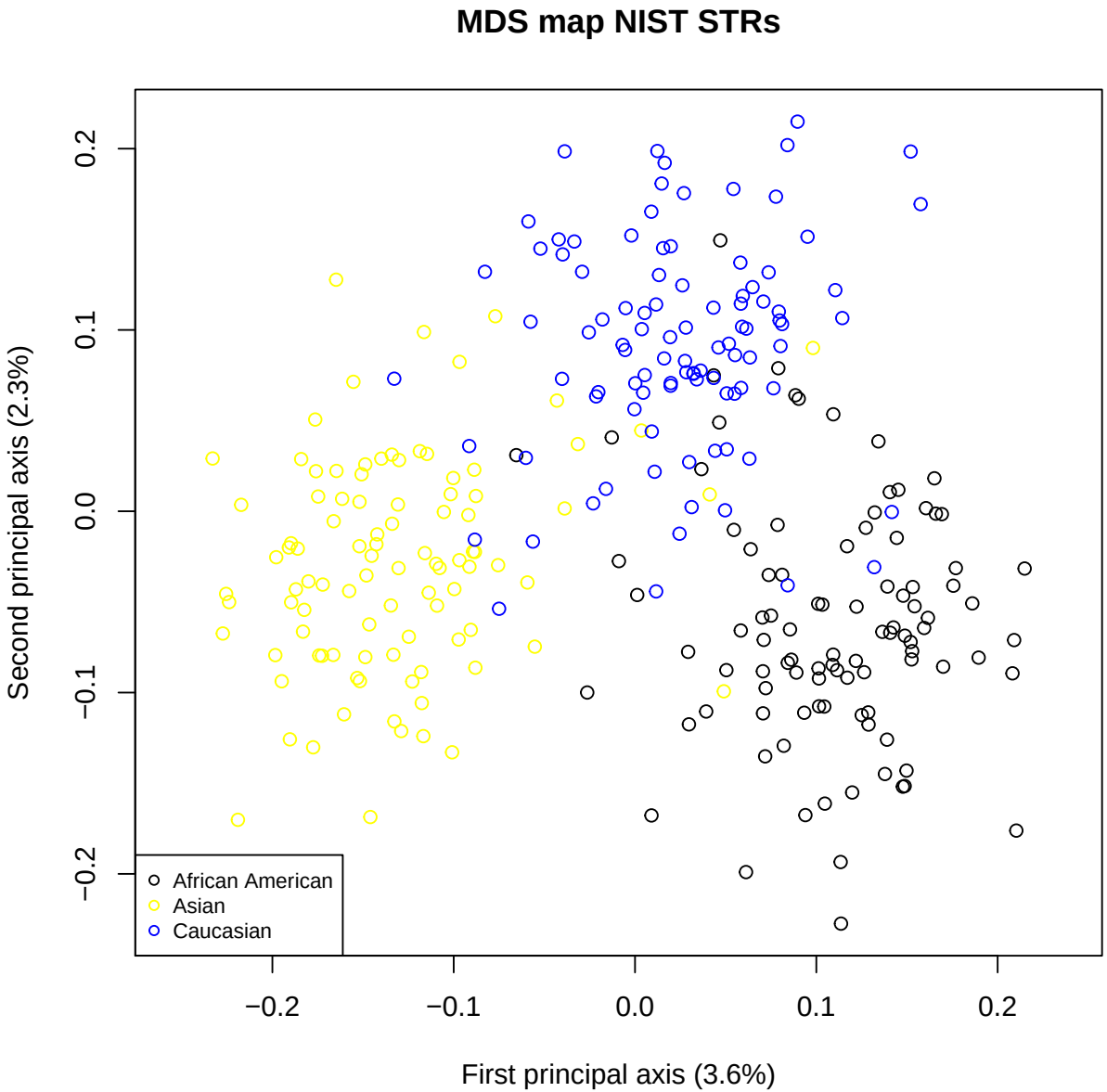
The data:

- 29 autosomal STRs
- Consider individuals with African-American, Asian and Caucasian ancestry
- Sample sizes balanced by subsampling

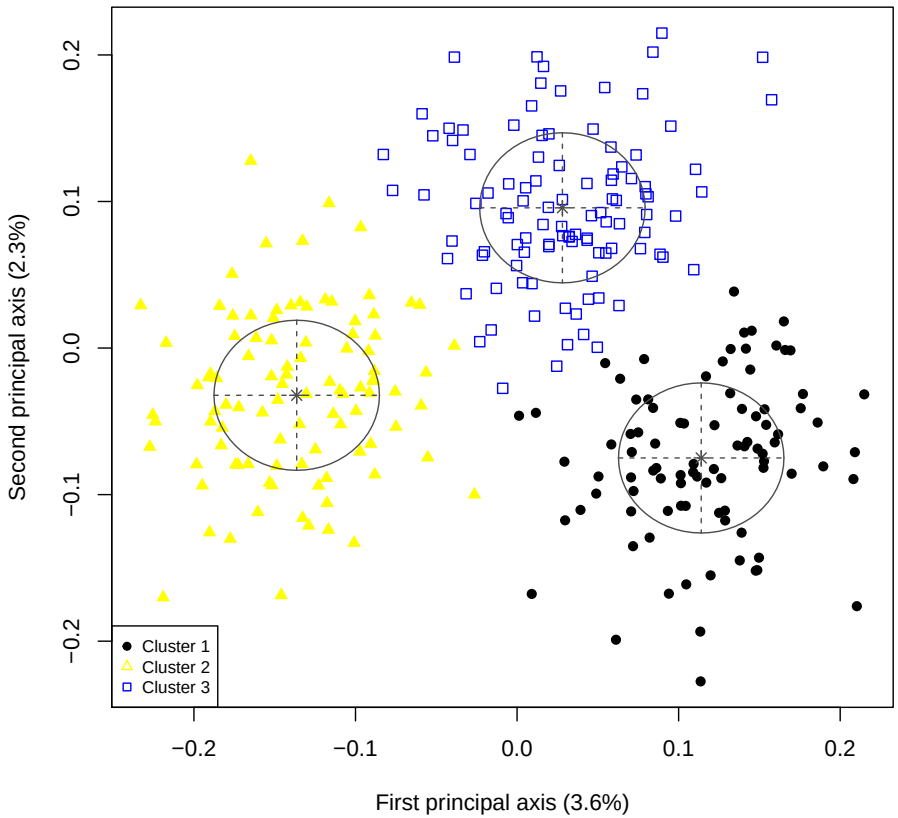
Prior to model-based clustering:

- STRs coded as binary variables
- Quantification of the data by MDS based on Jaccard metric

MDS map of NIST STRs



Model-based clustering with NIST STRs



	cluster		
	1	2	3
Prob	0.31	0.36	0.33
x_1	0.11	0.03	-0.14
x_2	-0.07	0.10	-0.03

Ancestry	Cluster		
	1	2	3
Afr. Am.	85	10	2
Asian	1	6	90
Caucasian	4	87	6

Cluster validity indices

- Choose the optimal number of clusters according to some (numerical) criterion
- Several criteria have been developed
- Some popular criteria:
  - Total within-cluster sum-of-squares (WSS)
  - Pseudo F-statistics (Calinski-Harabasz, 1974)
  - Silhouette coefficient (Rousseeuw, 1987)
  - ....
- Multiple indices can be calculated and used together to decide upon a number of clusters

Pseudo F statistics

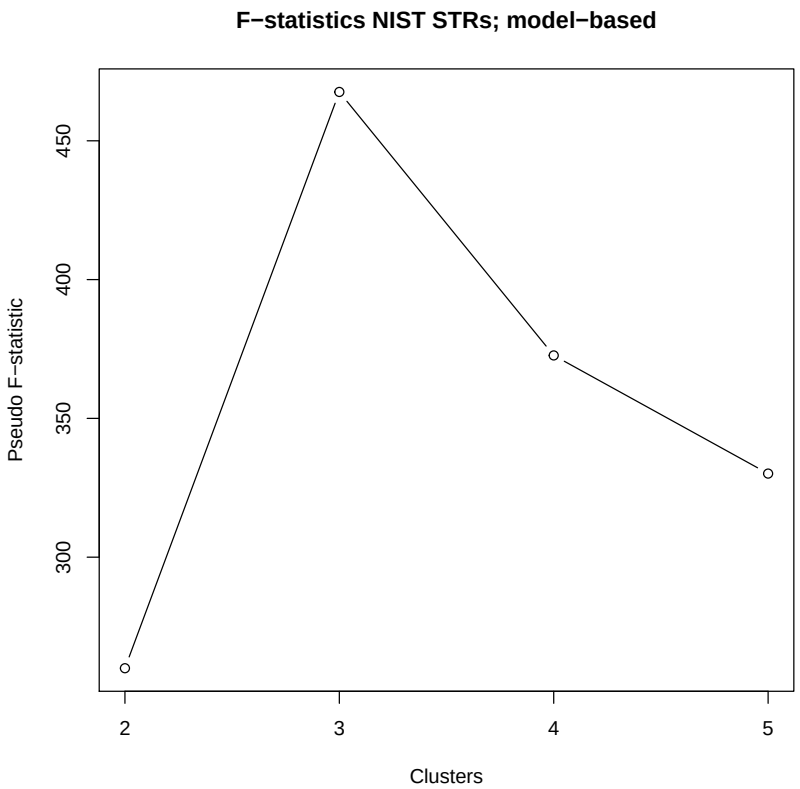
$$F = \frac{GSS/(K-1)}{WSS/(N-1)}$$

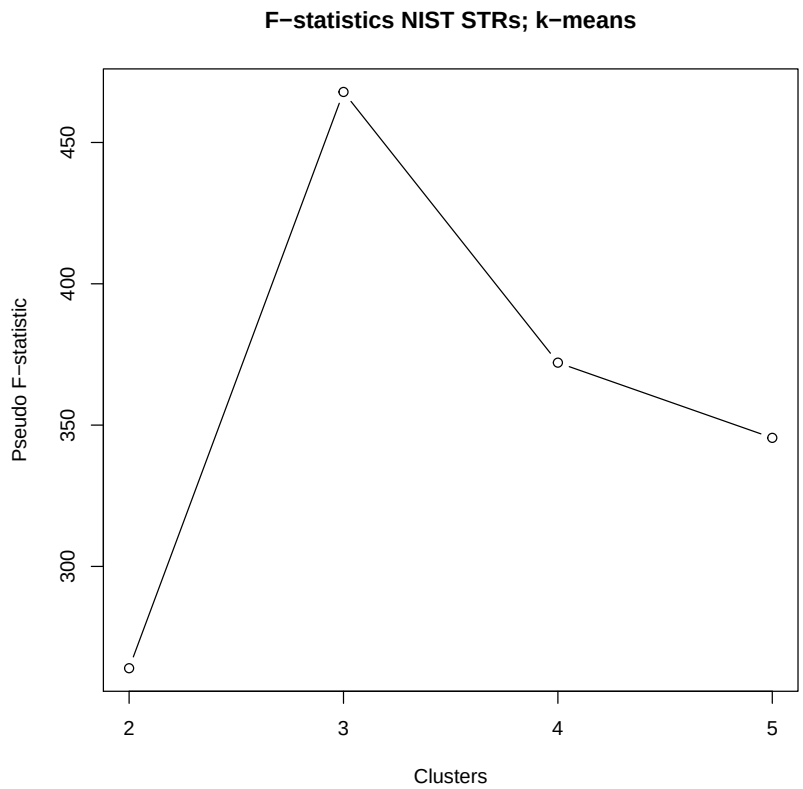
with:

- K = number of groups
- N = sample size
- GSS = between-group sum-of-squares
- WSS = within-group sum-of-squares

Choose the number of clusters that maximizes F

Example F statistics





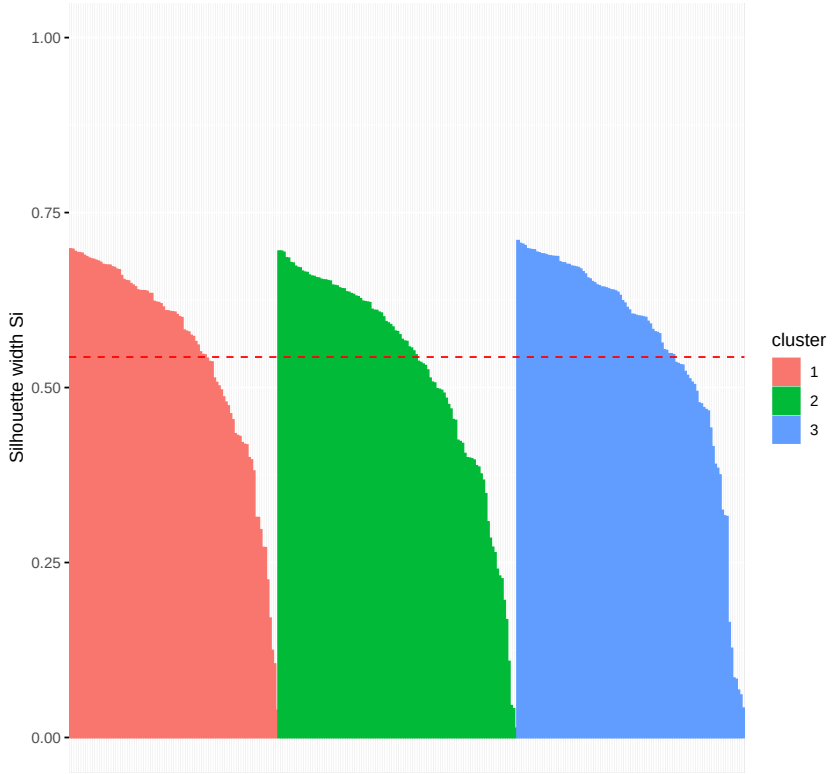
Silhouette scores and sihouette coefficient

- Let a_i be the average distance between observation i and all other points in the same cluster.
- Let b_i be the minimal average distance between observation i and all other points in another cluster.
- The silhouette score is defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \text{ and satisfies } -1 \leq s_i \leq 1$$

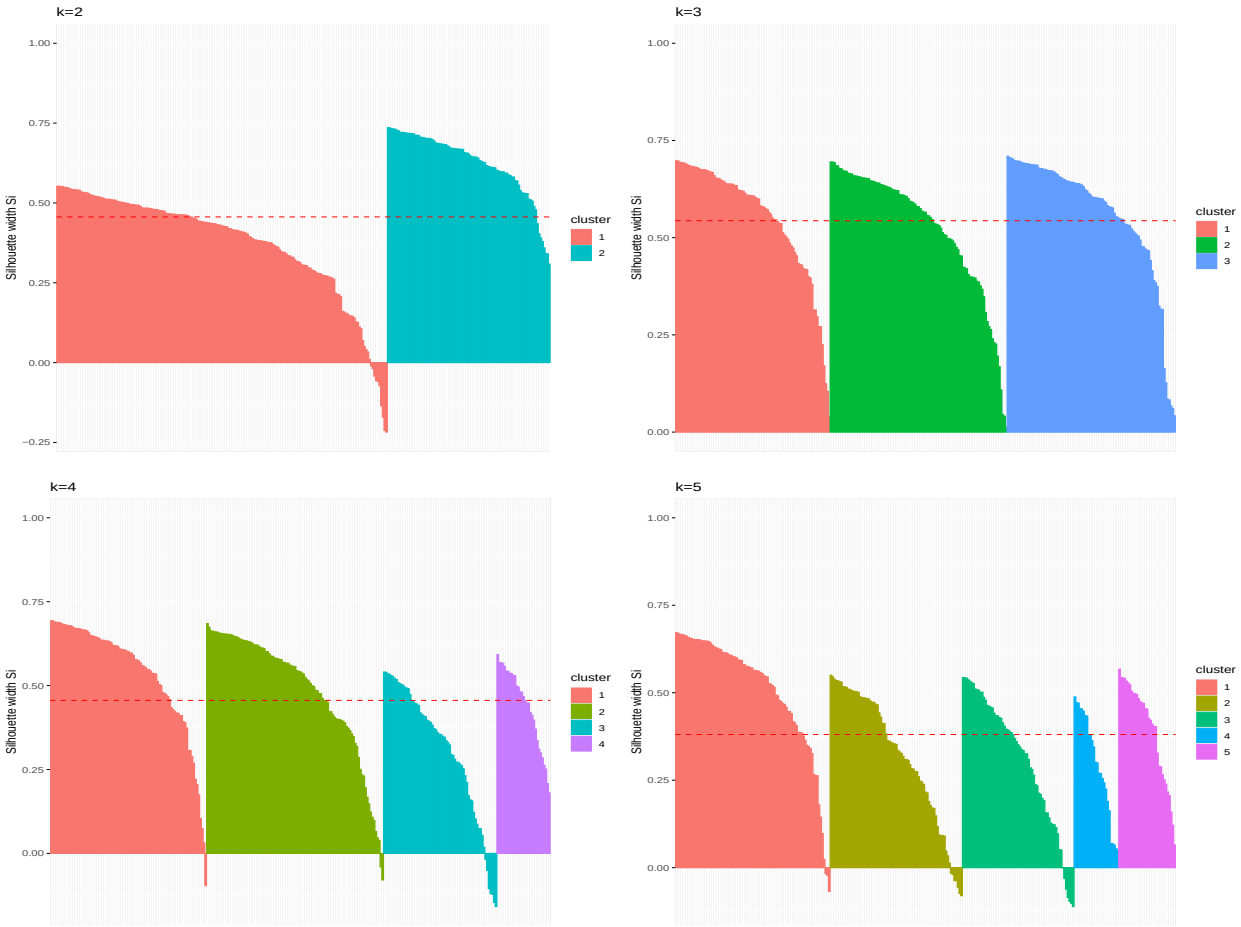
- s_i measures how well a case matches its own cluster.
- Note $a_i > 0$ and $b_i > 0$, and $s_i \approx 0$ means the case is on the boundary of two clusters.
- $s_i < 0$ means the case matches another cluster better.
- s_i can be averaged over all observations to give the average silhouette score.
- Choose the number of clusters that maximizes this average.

Silhouette scores NIST data ($k = 3$ with model-based clustering)

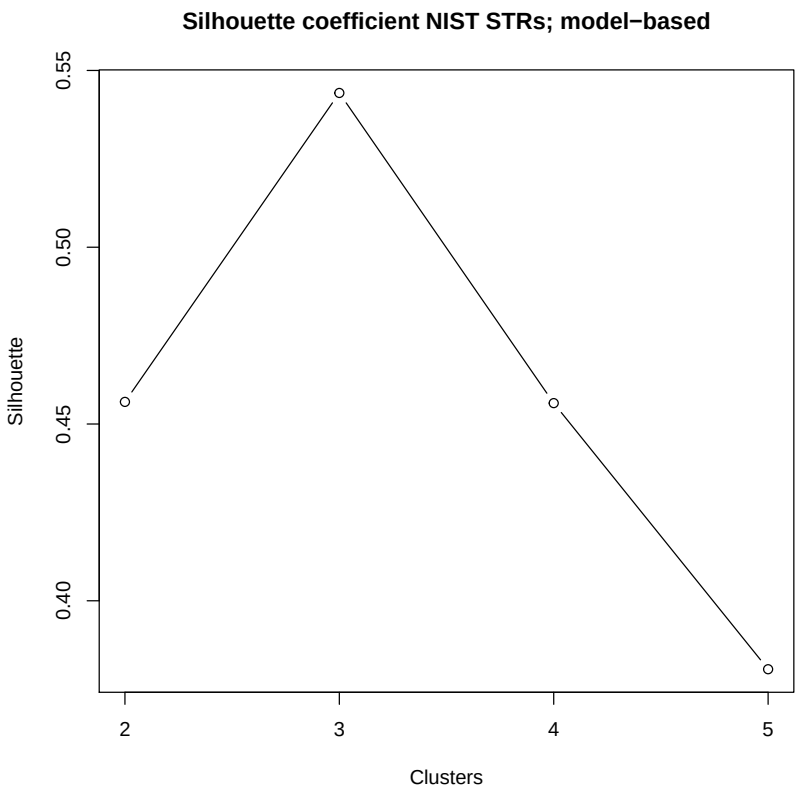


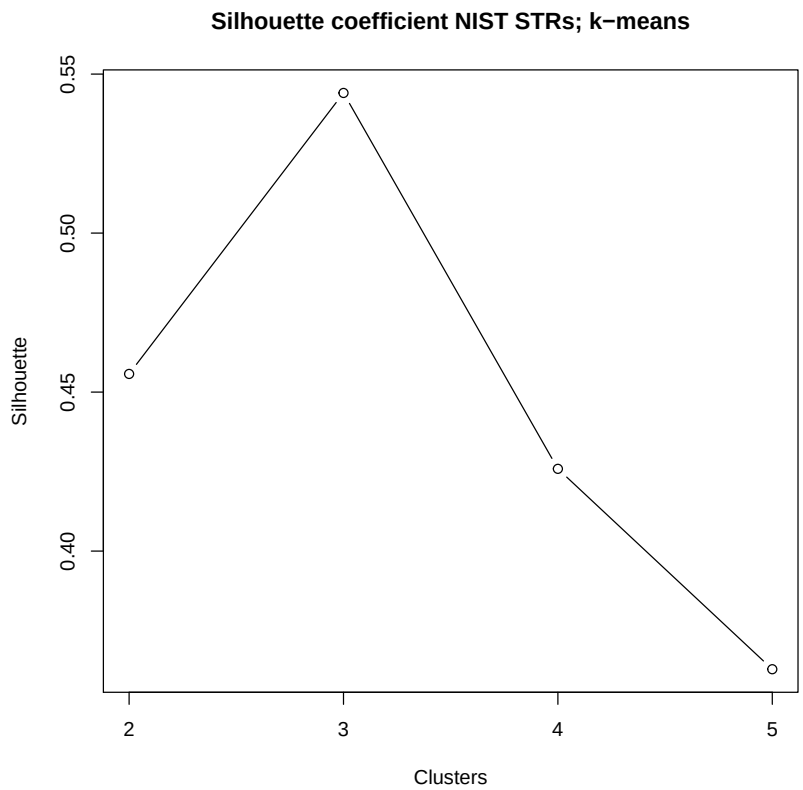
	All	1	2	3
s	0.54	0.55	0.52	0.56

For varying k



Average silhouette scores for varying k





After a cluster analysis

After obtaining the clusters:

- Are the clusters really different? (manova/anova)
- How homogeneous is each cluster?
- Which variables discriminate the clusters? (descriptive statistics per group, LDA/QDA)

Final remarks

In cluster analysis, the user has to make several choices:

1. the variables to include
2. possible transformations
3. the algorithm to use (hierarchical, divisive, model-based, ...)
4. the distance measure to use (Euclidean, City-Block, Mahalanobis, ...)
5. a measure of distance between clusters
6. metrics to assess the obtained clustering
7. ....

References

- Manly, B.F.J. (1989) Multivariate statistical methods: a primer. 3rd edition. Chapman and Hall, London.
- Johnson & Wichern (2002) Applied Multivariate Statistical Analysis. 5th edition. Prentice Hall, Chapter 12.
- Everitt, B.S., Landau, S., Lees, M. & Stahl, D. (2011) Cluster Analysis. 5th edition. Wiley.