Module 18 Multivariate Analysis for Genetic data
Session 10 Cluster Analysis II

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

26th Summer Institute in Statistical Genetics (SISG 2021)

Contents

1. Model-based clustering
2. Example STR data set
3. Cluster validation

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.

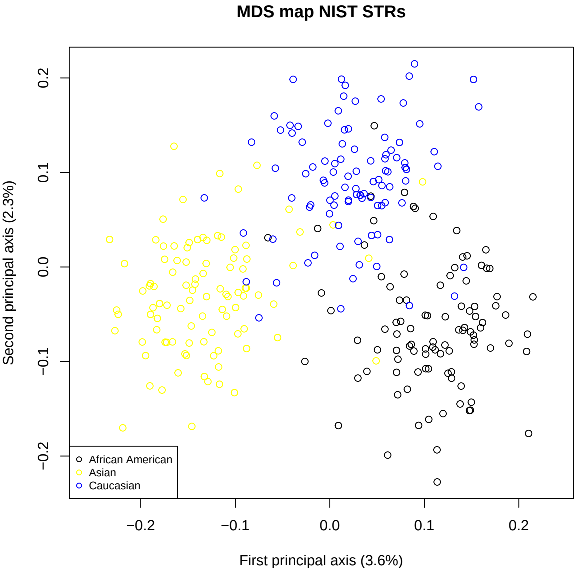
The data:

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

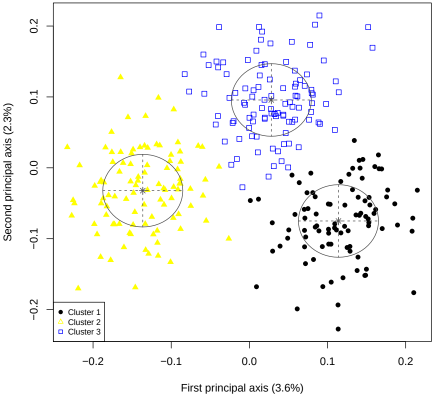
Prior to model-based clustering:

- 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

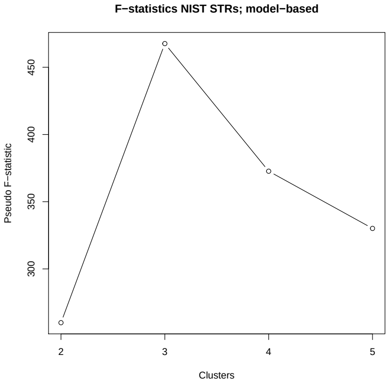
$$F = \frac{GSS/(K-1)}{WSS/(N-1)}$$

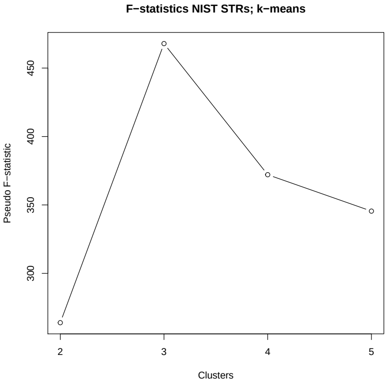
with:

- 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


F−statistics NIST STRs; model−based


F−statistics NIST STRs; k−means

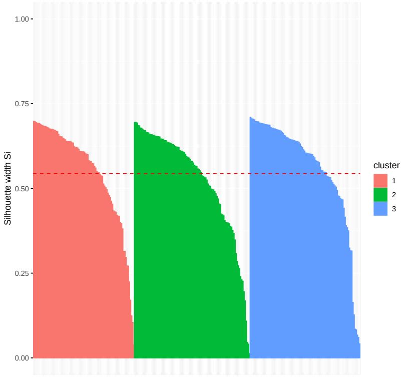
Silhouette scores and sihouette coefficient

- 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

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

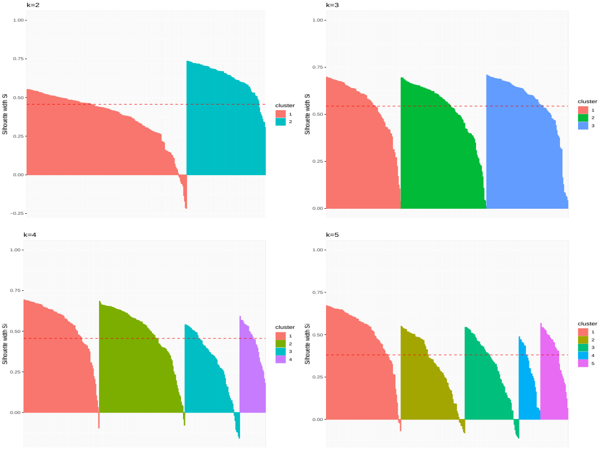
- s_i measures how well a case matches its own cluster.
- s_i can be averaged over all observations to give the average silhouette score.
- Choose the number of clusters that maximizes this average.

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

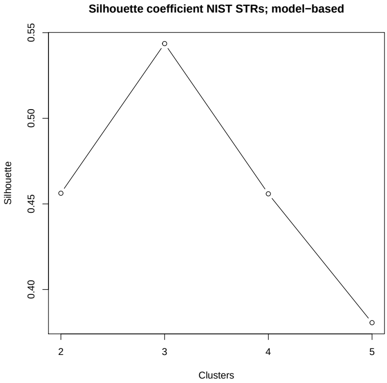


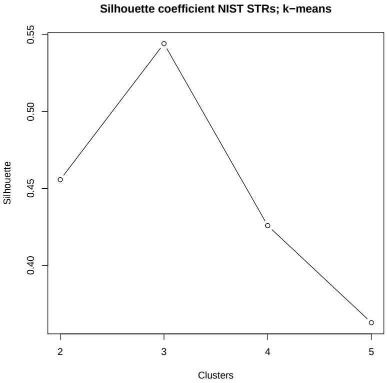
	All	1	2	3
s	0.54	0.55	0.52	0.56

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.