

This document is an historical remnant. It belongs to the collection Skeptron Web Archive (included in Donald Broady's archive) that mirrors parts of the public Skeptron web site as it appeared on 31 December 2019, containing material from the research group Sociology of Education and Culture (SEC) and the research programme Digital Literature (DL). The contents and file names are unchanged while character and layout encoding of older pages has been updated for technical reasons. Most links are dead. A number of documents of negligible historical interest as well as the collaborators' personal pages are omitted.

The site's internet address was since Summer 1993 www.nada.kth.se/~broady/ and since 2006 www.skeptron.uu.se/broady/sec/.

IV — What is Cluster Analysis?

Reference:

B. Le Roux, *L'analyse géométrique des données multidimensionnelles*, Dunod 2014, Chapters 10 & 11.

IV.1. The Aim of Cluster Analysis

Construct homogeneous clusters of objects (in GDA subclouds of points) so that:

- *compactness* criterion: objects within a same cluster are as much similar as possible;
- *separability* criterion: objects belonging to different clusters are as little similar as possible.

The greater the similarity (or homogeneity) within a cluster and the greater the difference between clusters the better the clustering.

heterogeneity between clusters — homogeneity within clusters

Types of Clustering

- ① algorithms leading to **partitions**.

Partitional clustering decomposes a data set into a set of disjoint clusters.

two following requirements:

- 1) each group contains at least one point,
- 2) each point belongs to exactly one group.

clustering around moving centers or *K-means cluster analysis*.

- ② algorithms leading to **hierarchical clustering** (the paradigm of natural sciences): system of nested clusters represented by a hierarchical tree or *dendrogram*.

- ▶ **ascending** algorithms (AHC)
- ▶ **descending** algorithms (segmentation methods):
problems of discrimination and regression by gradual segmentation of the set of objects → binary decision tree (methods AID, CART, etc.).

The methods of type 1 are *geometric* methods.

The method of type AHC is *geometric* if the distance is Euclidean and the aggregation index is the variance index.

The methods of type "segmentation" are not geometric.

The number of partitions into k clusters of n objects

n		k		
5 objects	into	2 clusters	=	15
10 objects	into	2 clusters	=	511
10 objects	into	5 clusters	=	42 525

etc.

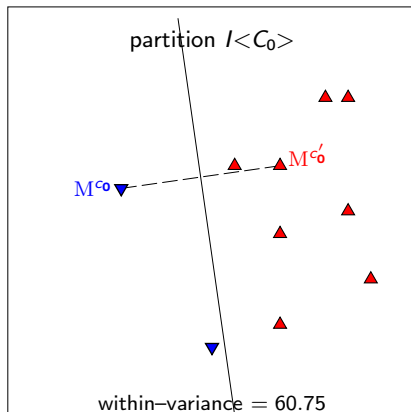
Except for small n , it is impossible to enumerate all the partitions of n individuals into k clusters

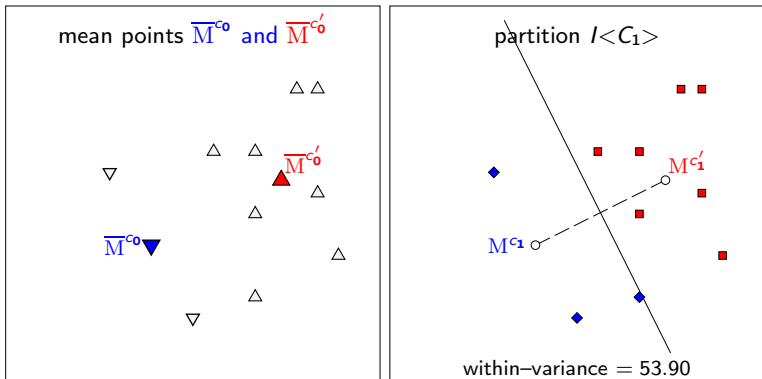
IV.2. K -means Clustering

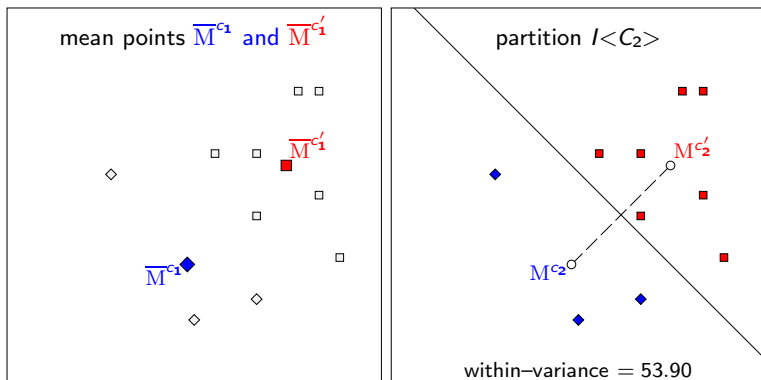
or aggregation around *moving centers*

- 1 Fix the number of clusters, say C ;
- 2 Choose (randomly or not) C initial class centers;
- 3 Assign each object to the closest center \rightarrow new clusters;
- 4 Determine the centers of the new clusters;
- 5 Repeat the assignment;
- 6 Stop the algorithm when 2 successive iterations provide the same clusters.

Choose 2 initial centers: M^{c_0} and $M^{c'_0}$







- Advantage: method is fast
- Disadvantage: The solution depends on the choice of initial centers.

IV.3. Ascending Hierarchical Clustering (AHC)

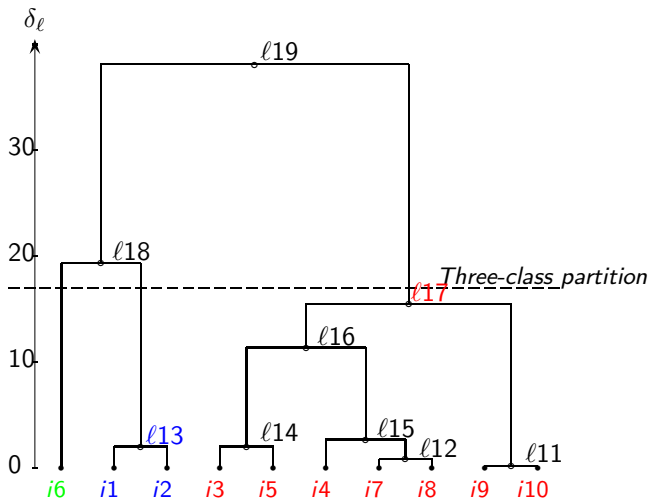
Clusters =

either the objects to be clustered (one–element class),
or the clusters of objects generated by the algorithm.

At each step, one groups the two elements which are the closest, hence the representation by a *hierarchical tree* or dendrogram.

We have to define the notion of “close”, that is, the *aggregation index*.

Target example: hierarchical tree



Ascending/agglomerative Hierarchical Clustering

Start with the basic objects (one–element clusters)
proceed to successive aggregations
until all objects are grouped in a single class.

AHC works “bottom–up”.

IV.4. Euclidean Clustering

- 1 Objects = *points of Euclidean cloud*.
- 2 *Aggregation index* (variance index) is the contribution of the two centers of the classes to be grouped (Ward's index).

Grouping property

If 2 clusters are *grouped*, the variance *decreases* from an amount equal to the contribution of the two centers of the clusters that are grouped.

Basic Algorithm

- **Step 1.** Calculate the contributions of the $9 \times 10/2 = 45$ pairs of points:

δ	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
i_2	2								
i_3	11.6	4							
i_4	6.8	3.2	4						
i_5	14.4	6.8	2	2					
i_6	13	17	27.4	10.6	20.2				
i_7	13	10.6	12.2	2.6	5.8	5.2			
i_8	14.6	9.8	8.2	1.8	2.6	10	0.8		
i_9	29.2	20.8	13.6	8	5.2	19.4	5	2.6	
i_{10}	31.4	21.8	13	9	5	23.2	6.8	3.6	0.2

Minimum index **0.2** for the pair of points $\{i_9, i_{10}\}$ which are aggregated (fig. 1), hence the mean point ℓ_{11} and a derived *cloud of 9 points* (fig. 2).

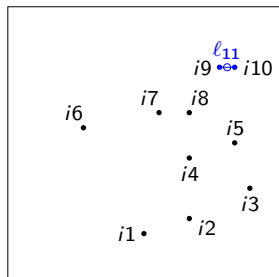
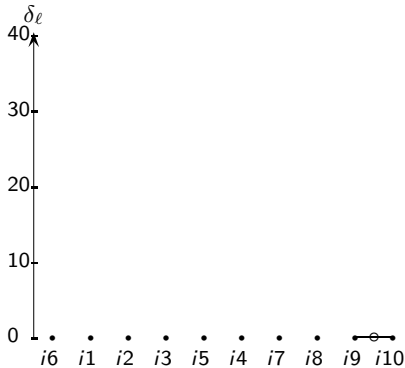


Figure 1



- **Step 2.** Calculate the aggregation index between the new point ℓ_{11} and the 8 other points.

New minimum 0.8 for $\{i7, i8\}$ which aggregated (fig. 2), hence the new point ℓ_{12} and a derived *cloud of 8 points* (fig. 3).

	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$i7$	$i8$
ℓ_{11}	40.33	28.33	17.67	11.27	6.73	28.33	7.8	4.07

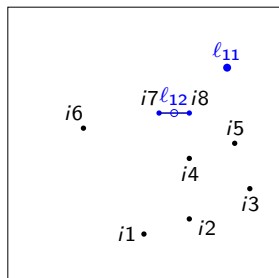
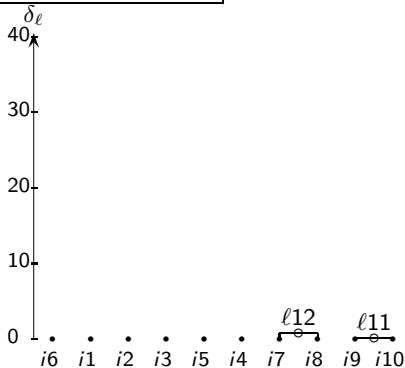


Figure 2



- **Step 3.** Iterate the procedure

Aggregation index between ℓ_{12} and the 7 other points

	i_1	i_2	i_3	i_4	i_5	i_6	ℓ_{11}
ℓ_{12}	18.13	13.33	13.33	2.67	5.33	9.87	8.2

Minimum = 2 for $\{i_1, i_2\}$, $\{i_3, i_5\}$ and $\{i_4, i_5\}$, aggregation of i_1 and i_2 (fig. 3), hence the point ℓ_{13} and a *cloud of 7 points* (fig. 4).

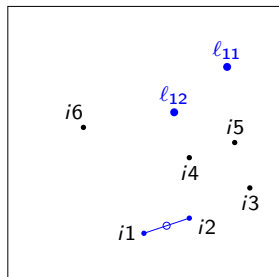
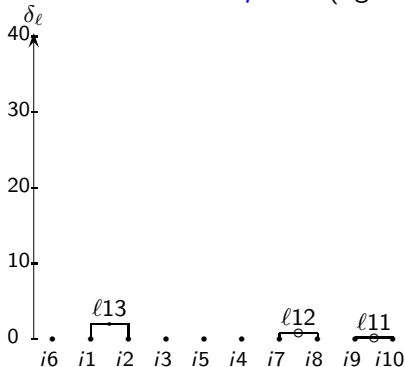


Figure 3

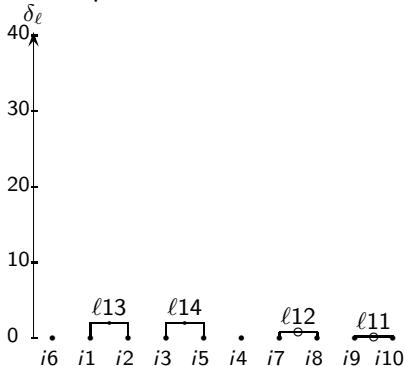
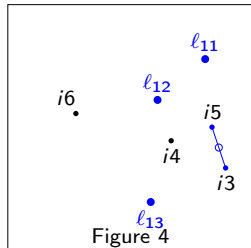


- **Step 4.** Iterate the procedure

Aggregation index between l_{13} and the 6 other points

	i_3	i_4	i_5	i_6	l_{11}	l_{12}
l_{13}	9.73	6.00	13.47	19.33	50.5	22.6

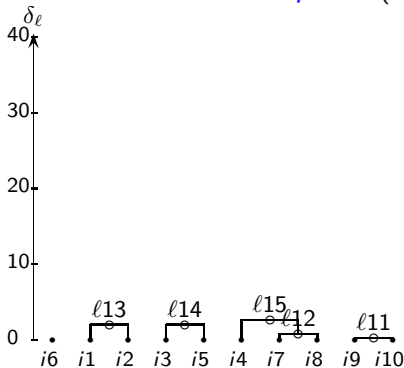
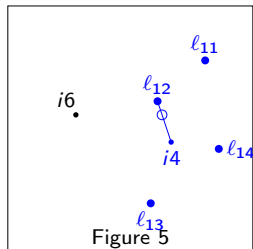
Minimum of index = 2 for the two pairs $\{i_3, i_5\}$ and $\{i_4, i_5\}$. Aggregation of i_3 and i_5 (fig. 4), hence the point l_{14} and the *cloud of 6 points* (fig. 5).



- **Step 5.** Aggregation index between l_{14} and the 5 other points

	i_4	i_6	l_{11}	l_{12}	l_{13}
l_{14}	3.33	31.07	17.33	13.00	16.4

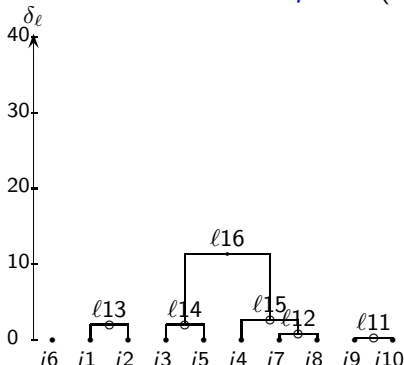
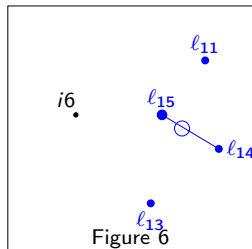
→ aggregation of l_{12} and i_4 at level 2.67 (fig. 5), hence the point l_{15} and the *cloud of 5 points* (fig. 6).



- **Step 6.** Aggregation index between l_{15} and the 4 other points

	i_6	l_{11}	l_{13}	l_{14}
l_{15}	12.03	12.49	20.61	11.33

→ aggregation of l_{15} and l_{14} at level 11.33 (fig. 6), hence the point l_{16} and the *cloud of 4 points* (fig. 7).



- **Step 7.** Aggregation index between l_{16} and the 3 other points

	i_6	l_{11}	l_{13}
l_{16}	21.67	15.57	20.86

 \rightarrow aggregation of l_{16} and l_{11} at level 15.57 (fig. 7),
 hence the point l_{17} and the *cloud of 3 points* (fig. 8).

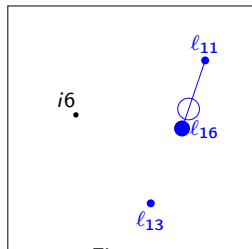
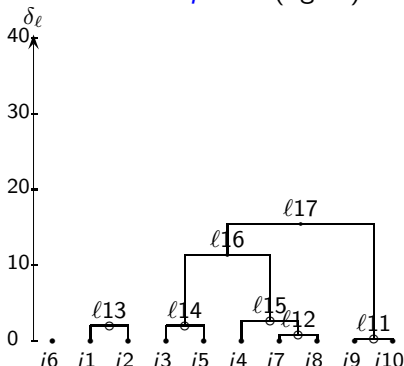


Figure 7



- Step 8.

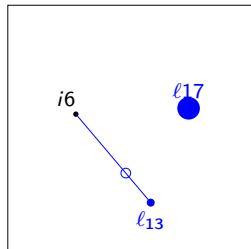
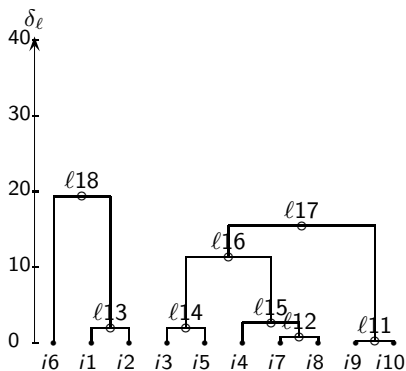


Figure 8



The three-class partition \mathcal{A} ($l14$), \mathcal{B} ($i6$), \mathcal{C} ($l17$) (already studied) with mean points A ($l13$), B ($i6$), C ($l17$) (fig. 8).

• Step 9.

Table of the contributions of the 3 pairs of points

	l_{13}	i_6	l_{17}
l_{13}	—		
i_6	19.33	—	
l_{17}	35.21	24.93	—

\Rightarrow grouping of i_6 and l_{13} at level 19.33 (fig. 9).

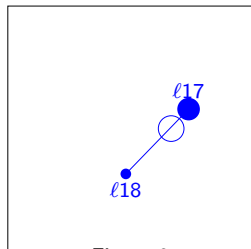
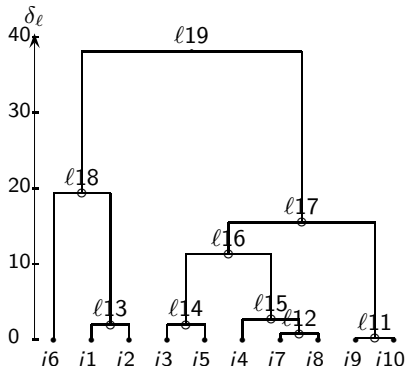
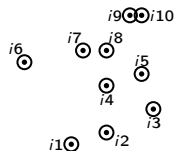


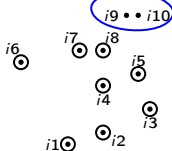
Figure 9



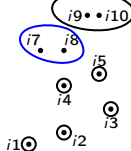
Step 0



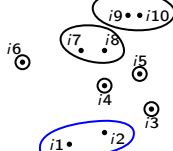
Step 1



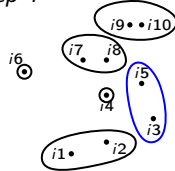
Step 2



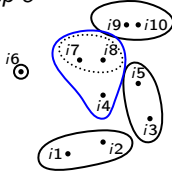
Step 3



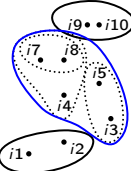
Step 4



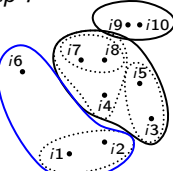
Step 5



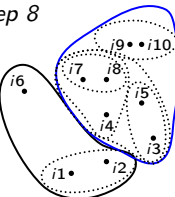
Step 6



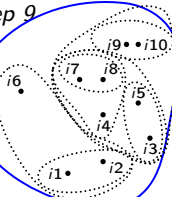
Step 7



Step 8



Step 9



Successive Steps of the AHC

ℓ	δ_ℓ	clusters	n	class description	<i>Between Var</i>	η_ℓ^2
ℓ_{19}	38.095	$\ell_{18} \ell_{17}$	10	$i_9 i_{10} i_3 i_5 i_4 i_7 i_8 i_6 i_1 i_2$	ℓ_{19}	
ℓ_{18}	19.333	$\ell_{13} \ell_6$	3	$i_6 i_1 i_2$	ℓ_{18}	38.10 .414
ℓ_{17}	15.571	$\ell_{16} \ell_{11}$	7	$i_9 i_{10} i_3 i_5 i_4 i_7 i_8$	ℓ_{17}	57.43 .624
ℓ_{16}	11.333	$\ell_{15} \ell_{14}$	5	$i_3 i_5 i_4 i_7 i_8$	ℓ_{16}	73.00 .793
ℓ_{15}	2.667	$\ell_{12} \ell_4$	3	$i_4 i_7 i_8$	ℓ_{15}	84.33 .917
ℓ_{14}	2.	$\ell_5 \ell_3$	2	$i_3 i_5$	ℓ_{14}	87.00 .957
ℓ_{13}	2.	$\ell_2 \ell_1$	2	$i_1 i_2$	ℓ_{13}	89.00 .967
ℓ_{12}	0.8	$\ell_8 \ell_7$	2	$i_7 i_8$	ℓ_{12}	91.00 .989
ℓ_{11}	0.2	$\ell_{10} \ell_9$	2	$i_9 i_{10}$	ℓ_{11}	91.80 .998
						92.00 1

Sum of the 9 level indices = 92 (variance of the cloud).

Between-variance of the 2-class partition = 38.095.

Between-variance of the 3-class partition = 38.095 + 19.333 = 57.43, etc.

IV.5. Interpretation of clusters

Interpretation is based on active variables then supplementary variables

Categorical variables

1. **descriptive criterion:**

Categories over-represented:

The relative frequency of the category in the cluster (f_c)
is 5% higher than the frequency in the whole set (f)
or is twice the one in the whole set.

$$f_c - f > 0.05 \quad f_c/f > 2$$

Categories under-represented: $f_c - f < -0.05 \quad f_c/f < 2$

2. **inductive criterion:**

The hypergeometric test of comparison of the frequency in the cluster to the reference frequency is significant.

Numerical variables

Variables retained for the interpretation:

1. **descriptive criterion:**

$$\frac{\text{mean for the cluster} - \text{mean for the overall set}}{\text{standard deviation for the whole set}} \geq 0.5$$

or

$$\frac{\text{mean for the cluster} - \text{mean for the overall set}}{\text{standard deviation for the whole set}} \leq -0.5$$

2. **inductive criterion:**

The combinatorial test of comparison of the mean in the cluster to the overall mean is significant.

Interpretation of the Hierarchy

The hierarchical tree is constructed in ascending direction.

- Reading a clustering (ascending and descending reading)
- Choice of the number of clusters:
from the diagram of level indexes.
- study of the successive dichotomies

Assignment of Supplementary Objects to Clusters

Mixed Clustering

IV.6. Other Aggregation Indices

- **Minimal jump**. the smallest distance between the elements of the 2 clusters = *single linkage clustering*.
- **Maximal jump**. The largest distance between elements of the two clusters = *diameter index*, or *complete linkage clustering*.
- **Mean distance** . Weighted mean of distances between the points of 2 clusters = *average linkage clustering*.

IV.7. Divisive Hierarchical Clustering

Start with one cluster and, at each step, split a cluster until only one–element clusters remain.

In this case, we need to decide which cluster will be split at each step and how to do the splitting.

Methods: CHAID and CART