Course on GDA

Geometric Data Analysis

Brigitte Le Roux

November 20–22, 2023

UPPSALA
UNIVERSITET

Stuctured Data Analysis

I.1 Introduction

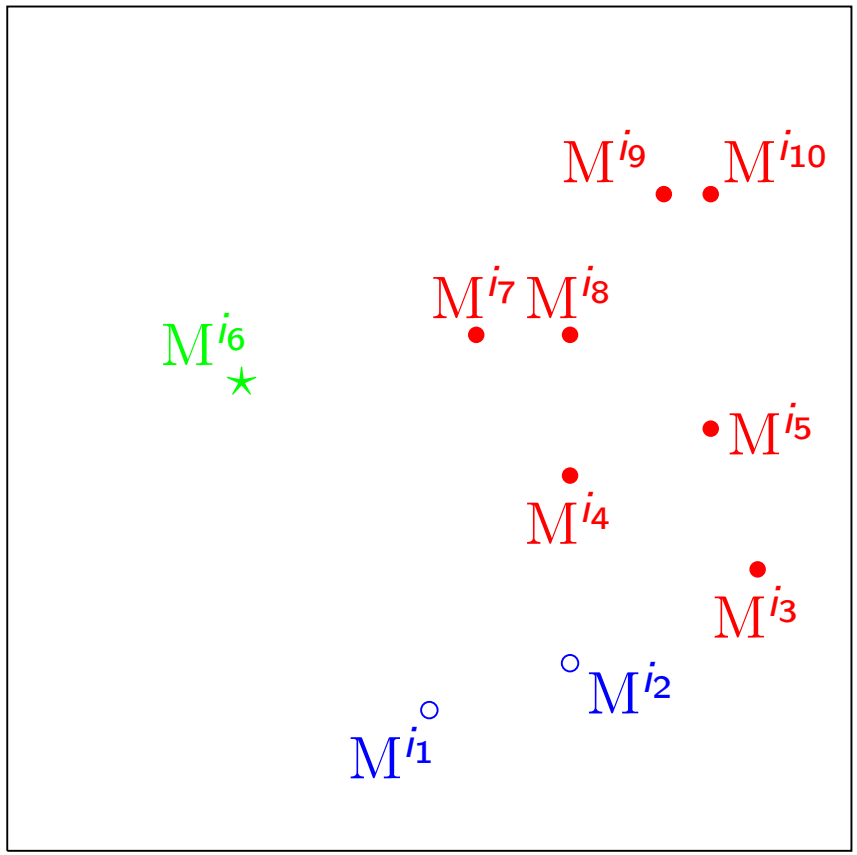
- *Structuring factors*:
  relevant variables describing the two basic sets that do not serve to construct the Euclidean clouds.
- *Structured data*:
  data tables whose basic sets are equipped with structuring factors;
- *Structured Data Analysis*, we mean the embedding of structuring factors into data analysis, in the line of Analysis of Variance—including its multivariate extension (MANOVA)—while preserving the principles of the construction of clouds.

I.2 Nesting

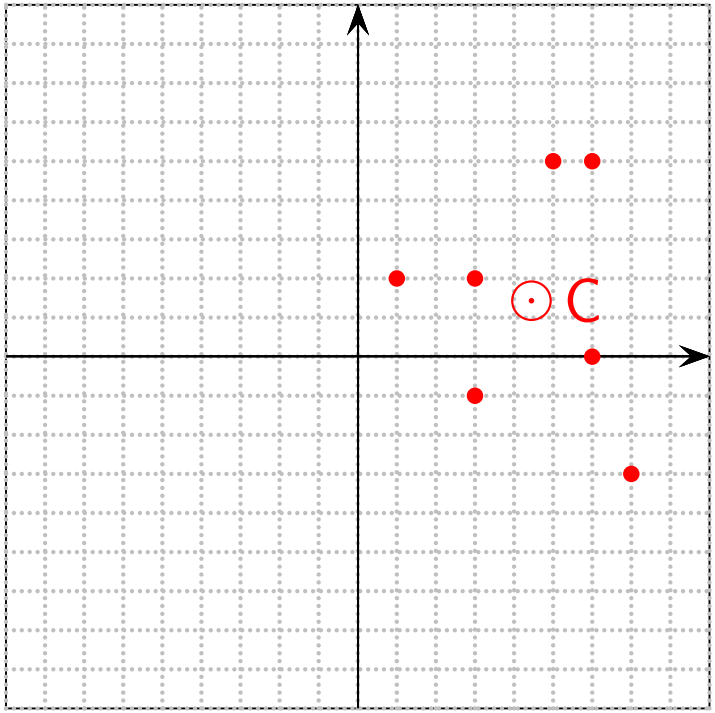
Consider a partition of the cloud into 3 classes.

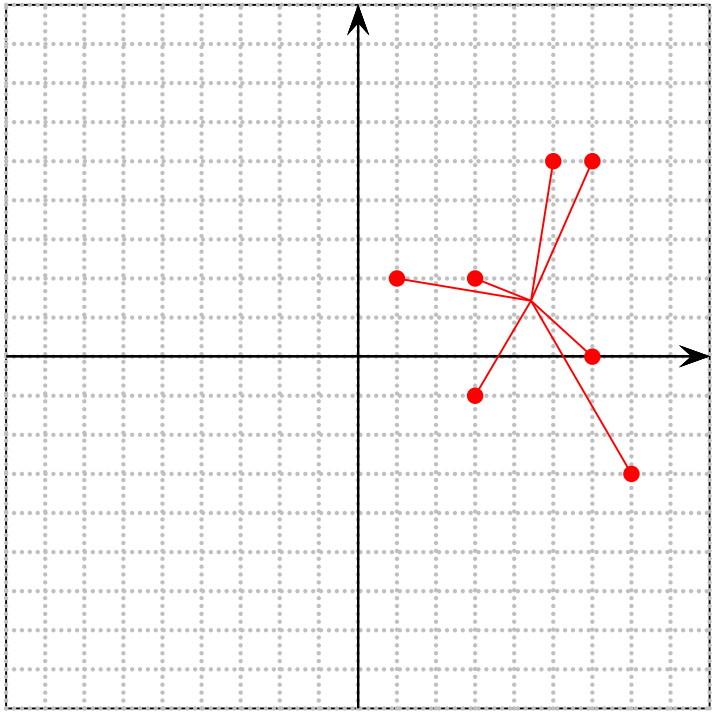
A part of a cloud defines a *subcloud*.

$\mathcal{A}$: subcloud of 2 points
$\{M^{i_1}, M^{i_2}\}$ (dipole)
$\mathcal{B}$: subcloud of 1 point
$\{M^{i_6}\}$
$\mathcal{C}$: subcloud of 7 points
$\{M^{i_3}, M^{i_4}, M^{i_5}, M^{i_7}, M^{i_8}, M^{i_9}, M^{i_{10}}\}$



Subcloud





Coordinates of point C: (4.43,1.43)
Weight of the cloud: 7
Variance of the cloud: $\mathrm{Var}_{\mathcal{C}} = 11.6425$

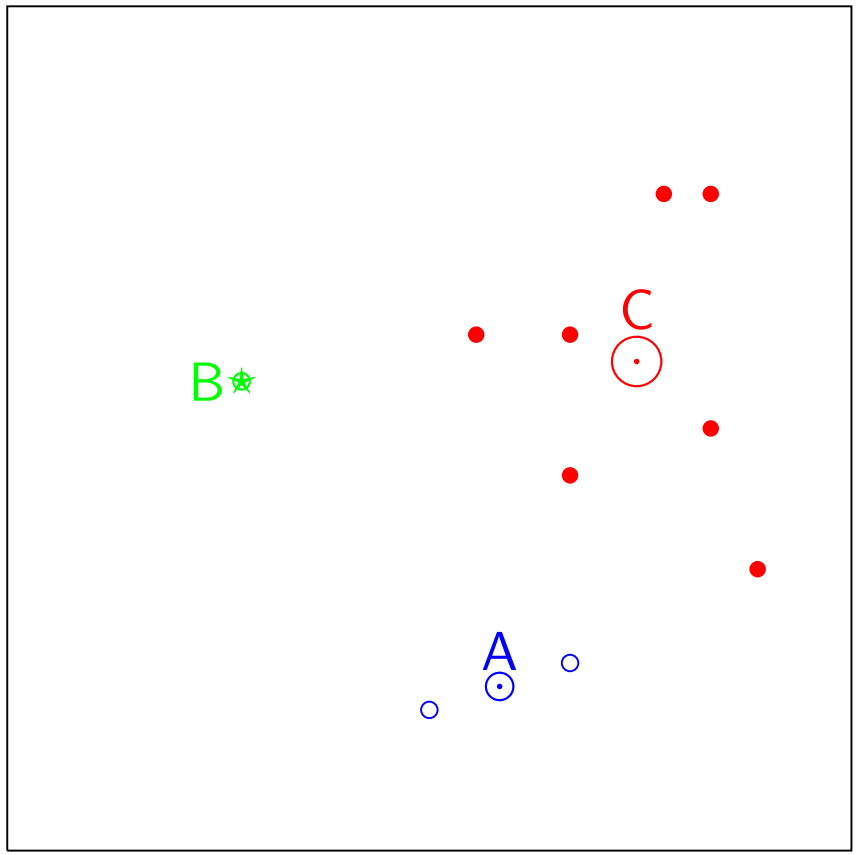
Mean points of subclouds

A, B, C are the mean points moyens of subcloudss $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$;
their weights are 2, 1, 7.

In a Euclidean points, by grouping:

- points are averaged,
- weights are added.

	Coordinates			Variances
	x_1	x_2	weight	
A	1.5	-5.5	$n_A = 2$	2.5
B	-4	1	$n_B = 1$	0
C	4.429	1.429	$n_C = 7$	11.6425
	$\overline{x}_1 = 3$	$\overline{x}_2 = 0$	$n = 10$	



Between cloud

The 3 mean points (A,2), (B,1) et (C,7) define the *between-cloud* associated with the partition.

The between cloud is weighted. Its weight: n ($n = 10$); its mean point: G.

- between-variance = variance of the between cloud :

$$\frac{n_A}{n}(\mathrm{GA})^2 + \frac{n_B}{n}(\mathrm{GB})^2 + \frac{n_C}{n}(\mathrm{GC})^2$$
$$= \frac{2}{10} \times 32.5 + \frac{1}{10} \times 50 + \frac{7}{10} \times 4.08163 = 6.5 + 5 + 2.857 = 14.357$$

- within-variance = weighted average of the variances of the subclouds

$$= \frac{2}{10} \times 2.5 + 0 + \frac{7}{10} \times 11.633 = 8.65$$

total variance = between-variance + within-variance

$$\eta^2 = \frac{\text{between-variance}}{\text{total variance}} \text{ (eta-deux)} \qquad \eta^2 = \frac{14.35}{23} = 0.62$$

I.3 Structured data analysis of the Taste Example

Study of Gender

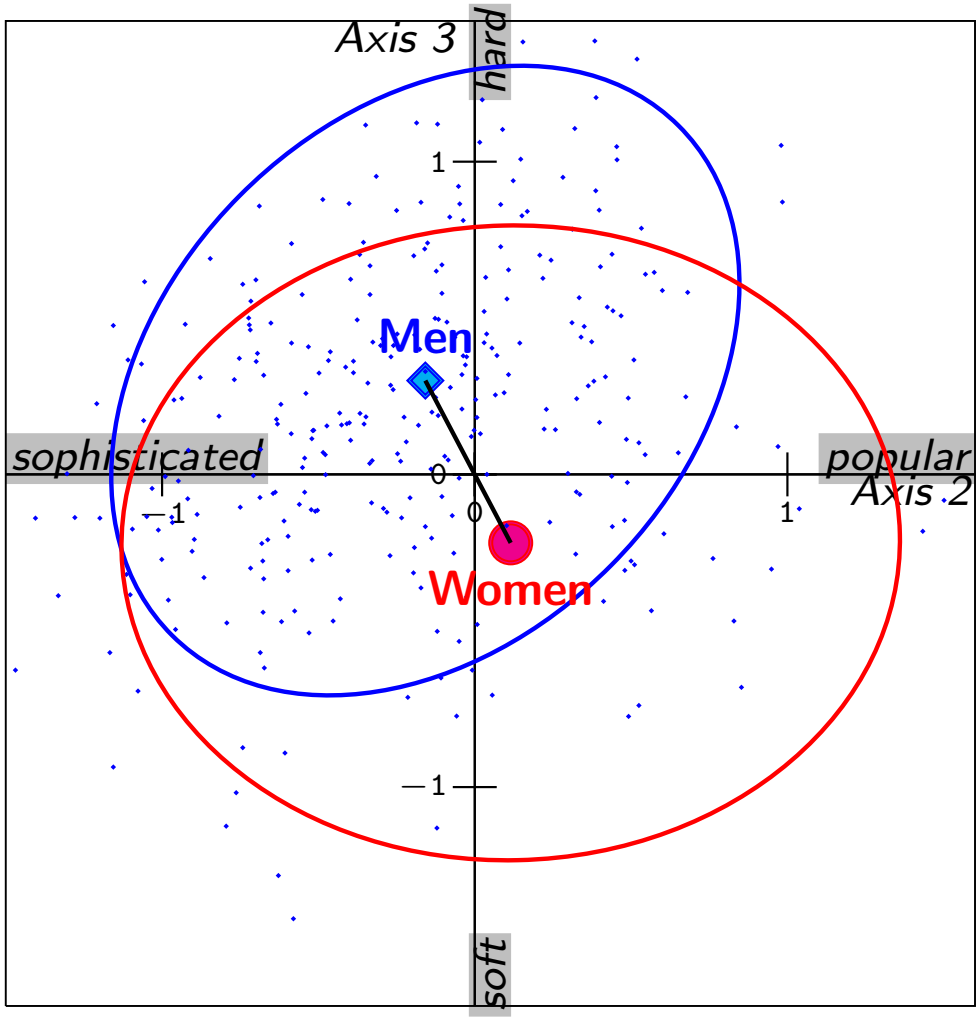
Gender		Mean point coordinates			Variances		
	weight	Axis 1	Axis 2	Axis 3	Axis 1	Axis 2	Axis 3
men	513	−0.112	−0.158	+0.300	.2915	.2528	.2567
women	702	+0.082	+0.115	−0.219	.4639	.3916	.2613
				within–Gender	.3911	.3330	.2593
				between–Gender	.0092	.0182	.0657
				total (λ)	.4004	.3512	.3250

scaled deviations (deviation divided by $\sqrt{\lambda}$)

axis 1: $\dfrac{-0.112 - 0.082}{\sqrt{0.4004}} = -0.308;$

axis 2: $\dfrac{-0.273}{\sqrt{0.3512}} = -0.461;$

axis 3: $\dfrac{+0.519}{\sqrt{0.3512}} = -0.910 > 0.4$: the deviation is large

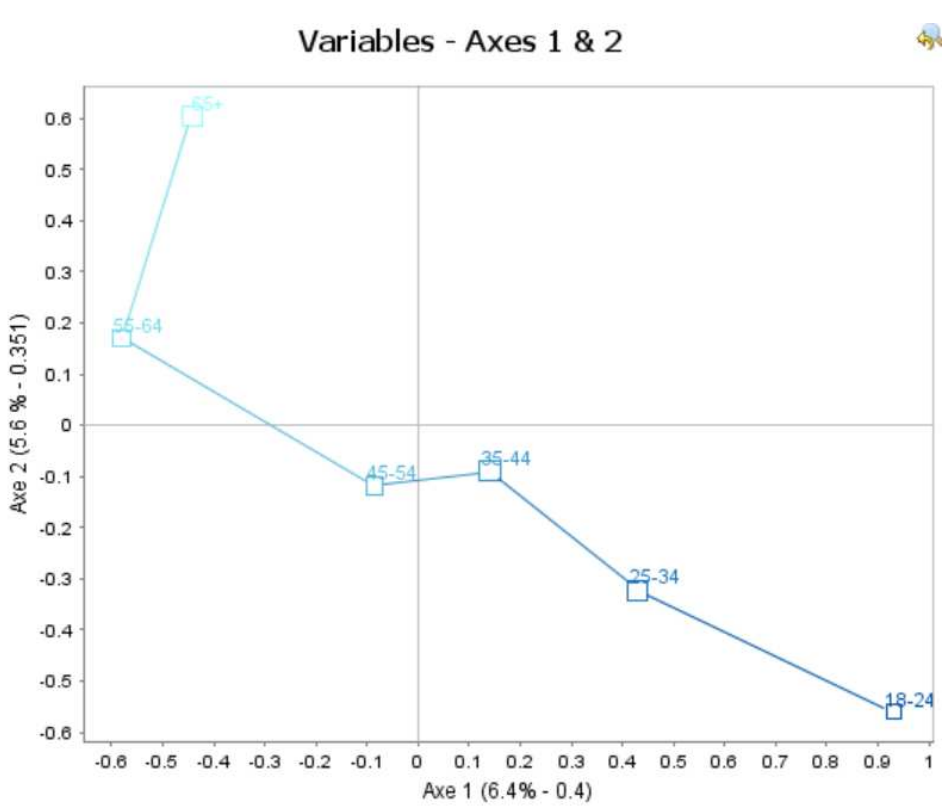


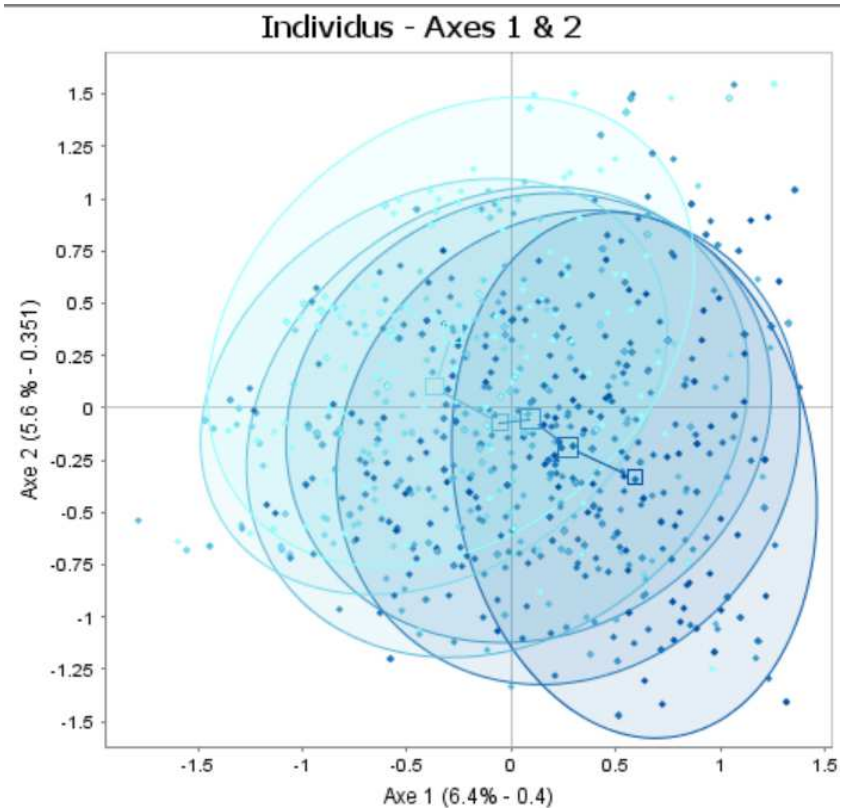
Exemple Taste. Subcloud of Men and of Women in principal plane 2-3

Study of Age

6 age groups $\longrightarrow$ 6 subclouds

Age		Coordinates			Variances		
	weight	Axis 1	Axis 2	Axis 3	Axis 1	Axis 2	Axis 3
18-24	93	+0.589	−0.332	+0.014	.1916	.3946	.2581
25-34	248	+0.272	−0.191	−0.014	.3083	.3225	.2934
35-44	258	+0.089	−0.053	+0.052	.3371	.2880	.3406
45-54	191	−0.054	−0.070	−0.047	.3604	.3176	.3120
55-64	183	−0.367	+0.101	−0.013	.3121	.2459	.4095
$\geq$ 65	242	−0.281	+0.359	0.000	.3401	.3143	.3078
				within	.3206	.3068	.3240
				between	.199	.126	.003
				total (λ)	.4004	.3512	.3250
				η^2	.0798	.0444	.0010





In plane 1-2

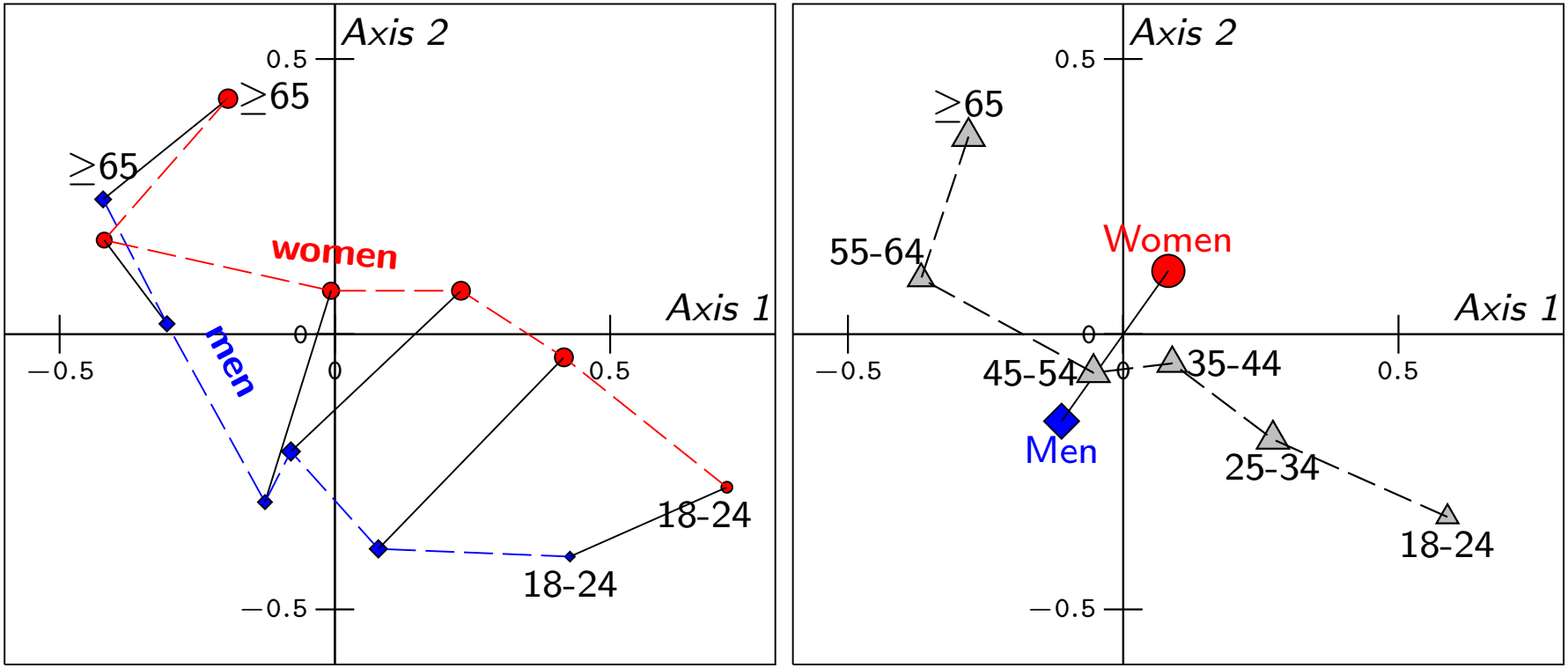
between-variance = 0.0798+0.0444= 0.1242

$$\eta^2 = \frac{0.0798 + 0.0444}{0.4004 + 0.3512} = 16.5\%$$

I.4 Crossing of Age and Gender

In the cloud of individuals, the crossing of *Age* and *Gender* induces a cloud of $6 \times 2 = 12$ category mean points.

Gender×	men			women			*Age*		
Age	*n*	Axis 1	Axis 2	*n*	Axis 1	Axis 2	*n*	Axis 1	Axis 2
18-24	40	+0.4267	−0.4043	53	+0.7121	−0.2781	93	+0.589	−0.332
25-34	106	+0.0792	−0.3904	142	+0.4163	−0.0417	248	+0.272	−0.191
35-44	117	−0.0799	−0.2130	141	+0.2292	+0.0792	258	+0.089	−0.053
45-54	74	−0.1273	−0.3049	117	−0.0073	+0.0791	191	−0.054	−0.070
55-64	84	−0.3055	+0.0185	99	−0.4188	+0.1711	183	−0.367	+0.101
≥65	92	−0.4209	+0.2452	150	−0.1945	+0.4282	242	−0.281	+0.359
Gender	513	−0.112	−0.158	702	+0.082	+0.115			

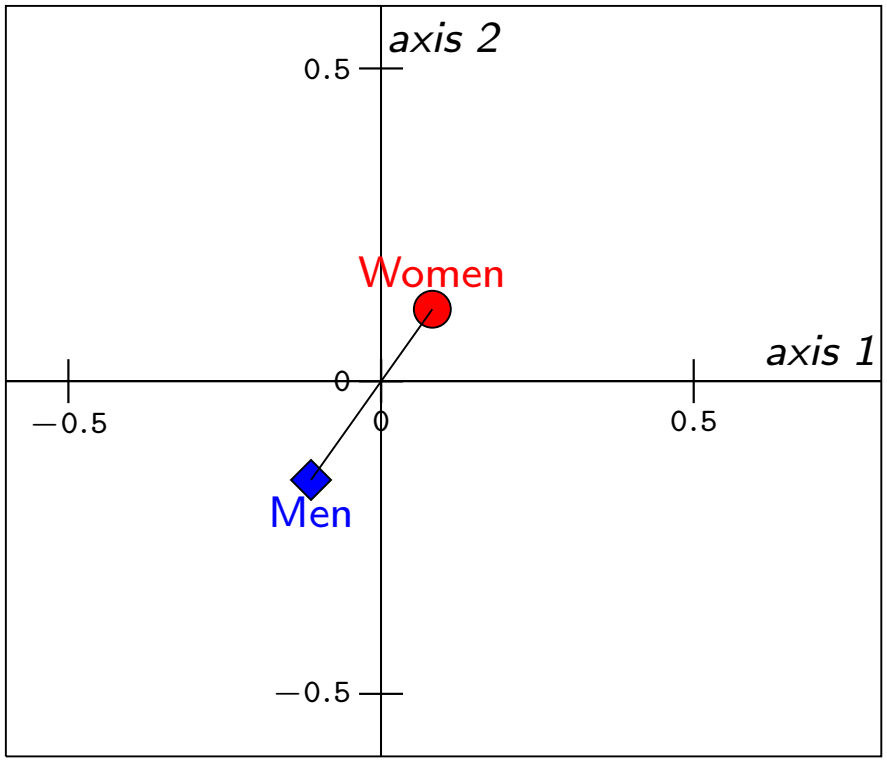


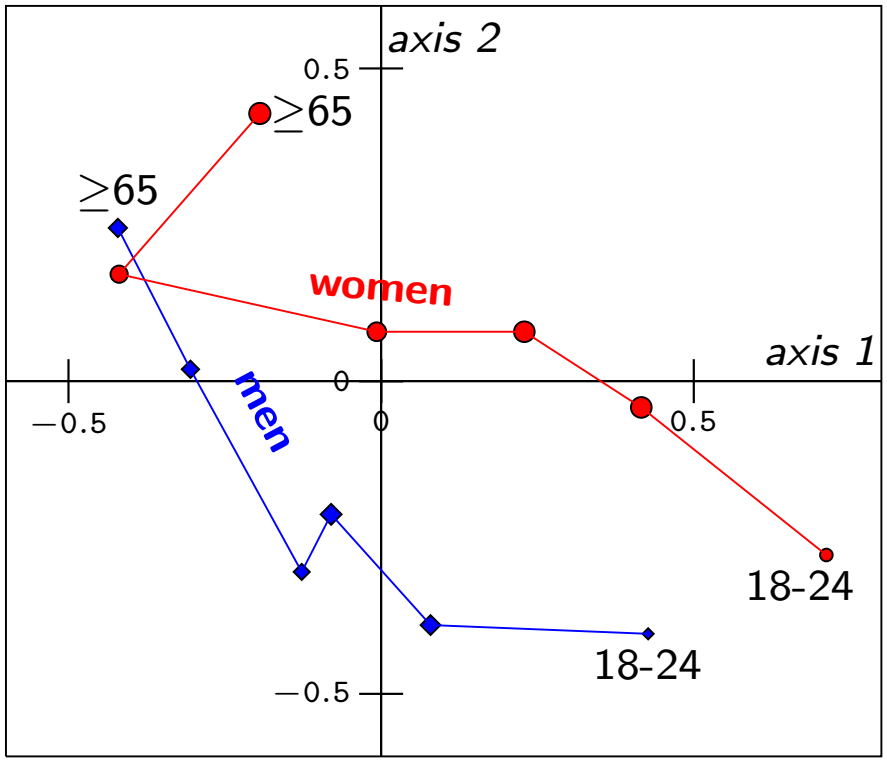
Age×*Gender* cloud — *Gender* and *Age* clouds, in plane 1-2

variance of *Age*×*Gender* cloud: 0.1586
variance of *Gender* cloud: 0.0274
variance of *Age* cloud: 0.1242
$0.1586 \neq 0.0274 + 0.1242$

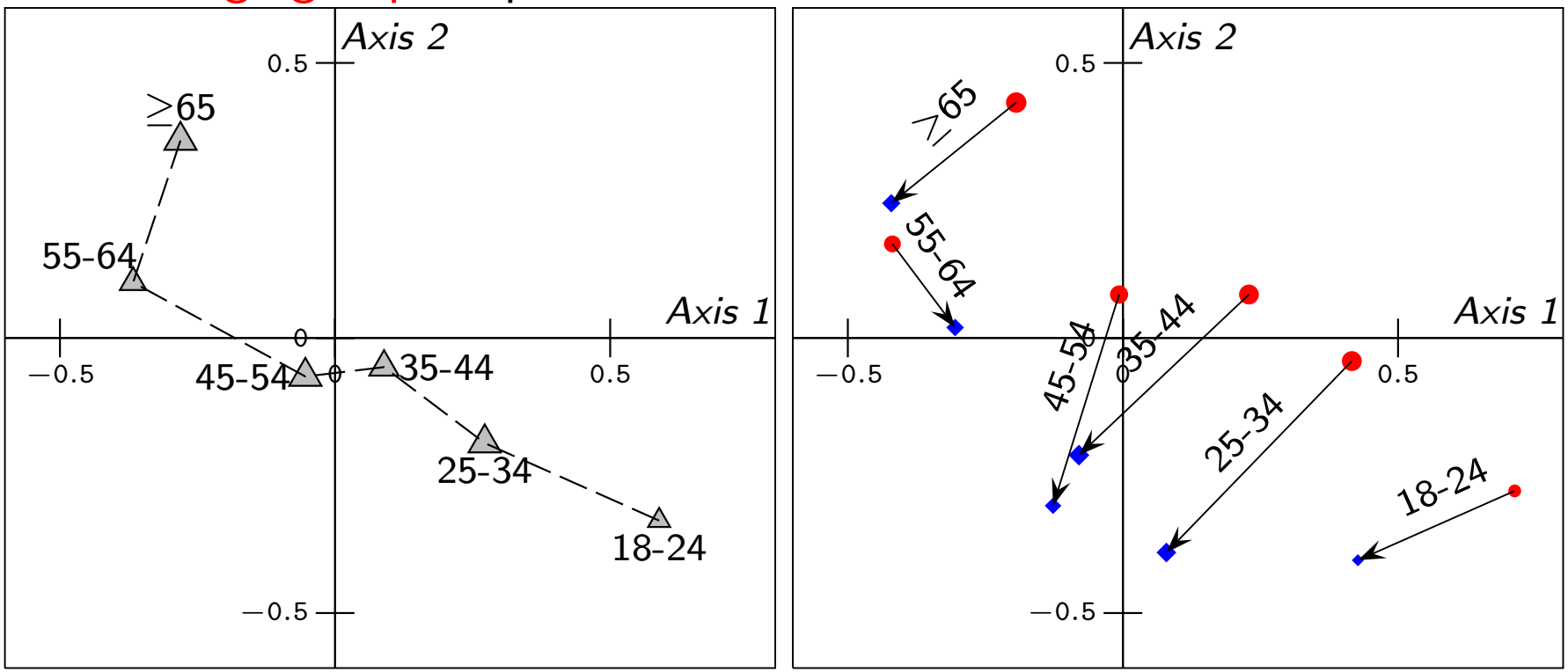
Additive breakdown of variances of the *Age*×*Gender* cloud and variances of the *Age*×*Gender* cloud, for axis 1, axis 2 and plane 1-2.

	between genders	*Age* within–*Gender*	*Age*× *Gender*
Axis 1	0.0092	0.0862	0.0954
Axis 2	0.0182	0.0450	0.0632
Plane 1-2	0.0274	0.1312	0.1586





Main effect (vector) of *Gender* and six within–effects (vectors) of *Gender* for each age group, in plane 1-2.



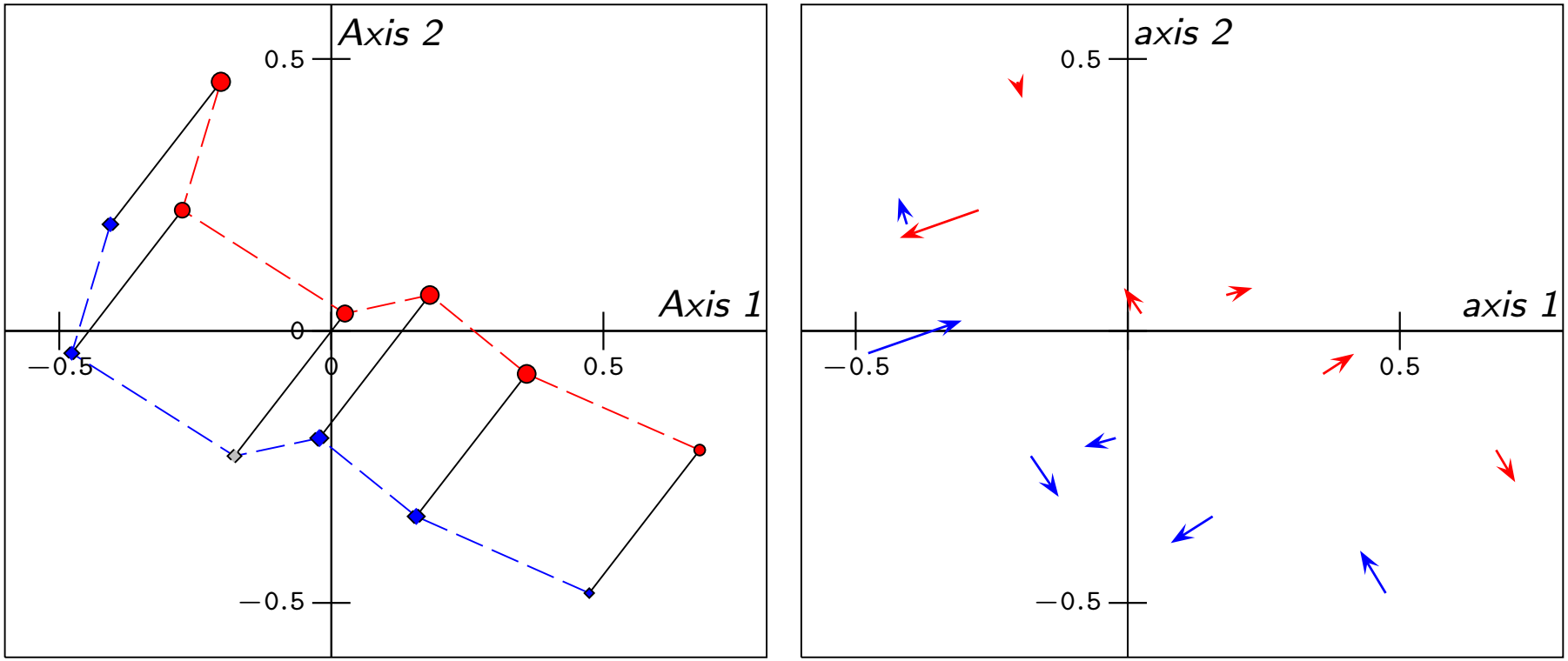
Age	18-24	25-34	35-44	45-54	55-64	≥65
Axis 1	0.0200	0.0278	0.0237	0.0034	0.0032	0.0121
Axis 2	0.0039	0.0298	0.0212	0.0350	0.0058	0.0079
Plane 1-2	0.0239	0.0576	0.0448	0.0384	0.0090	0.0200
weights	93	248	258	191	183	242

Variances of the 6 dipoles

Additive breakdown of variance of the *Age*×*Gender* cloud and variances of the *Age*×*Gender* cloud, for axis 1, axis 2 and plane 1-2.

	between ages	*Gender* within–*Age*	*Age*× *Gender*
Axis 1	0.0798	0.0157	0.0954
Axis 2	0.0444	0.0188	0.0632
Plane 1-2	0.1242	0.0345	0.1586

Additive and Interaction clouds



Additive cloud fitted to the *Gender*×*Age* cloud (left) and the 12 deviations (vectors) from the *Gender*×*Age* cloud (right).

Interaction

	additive	inter-action	*Age*×*Gender*
Axis 1	0.0898	0.0056	0.0954
Axis 2	0.0611	0.0021	0.0632
Plane 1-2	0.1509	0.0077	0.1586

Interaction:
however weak, is revealed by the fact that the variance of the additive cloud (0.1509) is smaller than the variance of *Gender*×*Age* cloud (0.1586).

Low *correlation between factors* is revealed by the fact that
V_G (0.0274) is not far from $V_{G_{\text{within}A}}$ (0.034),
V_A (0.1242) is not far from $V_{A_{\text{within}G}}$ (0.131)
the variance of the additive cloud (0.1509) is close to the sum of the between variances ($V_G + V_A = 0.1516$).

References

— Le Roux B. (2014). *Analyse géométrique des données multidimensionnelles*, chap.9 (pp. 303–320), Paris: Dunod
— Le Roux B. (2014). Structured Data Analysis, Chap 12 (pp185-203) in *Visualization and verbalization of data*, Blasius, J., & Greenacre, M. (Eds.).. CRC Press.
— Le Roux, B., & Rouanet, H. (2010). Multiple correspondence analysis (Vol. 163). Sage.
— Le Roux B., Rouanet H., Savage M. and Warde A., Class and Cultural Division in the UK, *Sociology* 2008; 42; 1049-1071. (http://soc.sagepub.com/cgi/reprint/42/6/1049).
— Silva E., Le Roux B., Cultural capital of couples: Tensions of elective affinities, *Poetics*, 39, 547-565.

What is Cluster Analysis?

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

V.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 divides a data set into a set of disjoint clusters (partition).
   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 $\rightarrow$ 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.

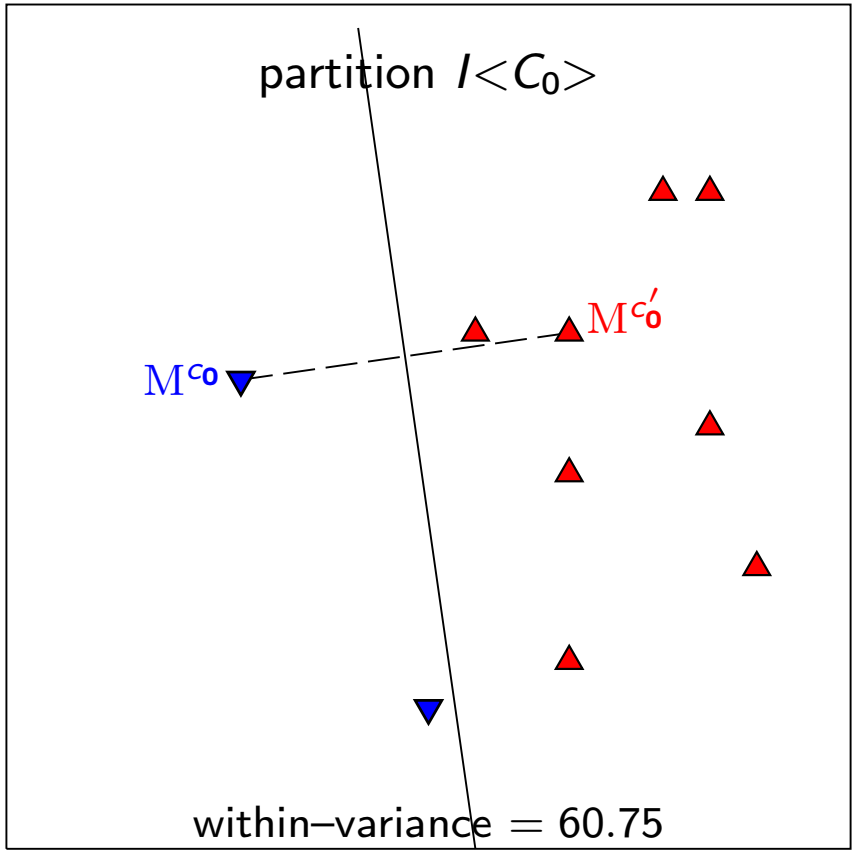
Except for small n, it is impossible to enumerate all the partitions of n individuals into k clusters

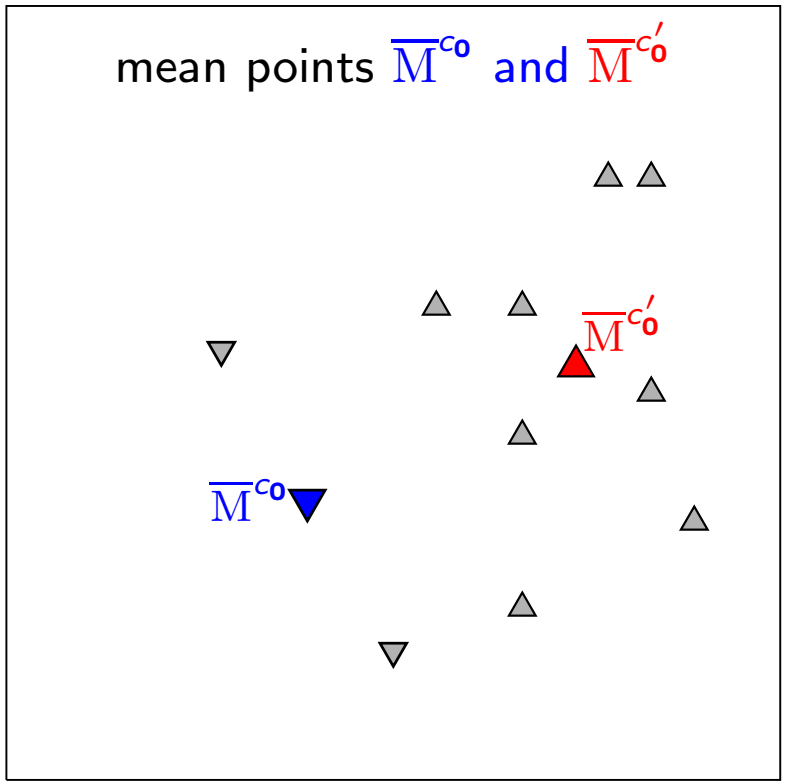
V.2. K–means Clustering

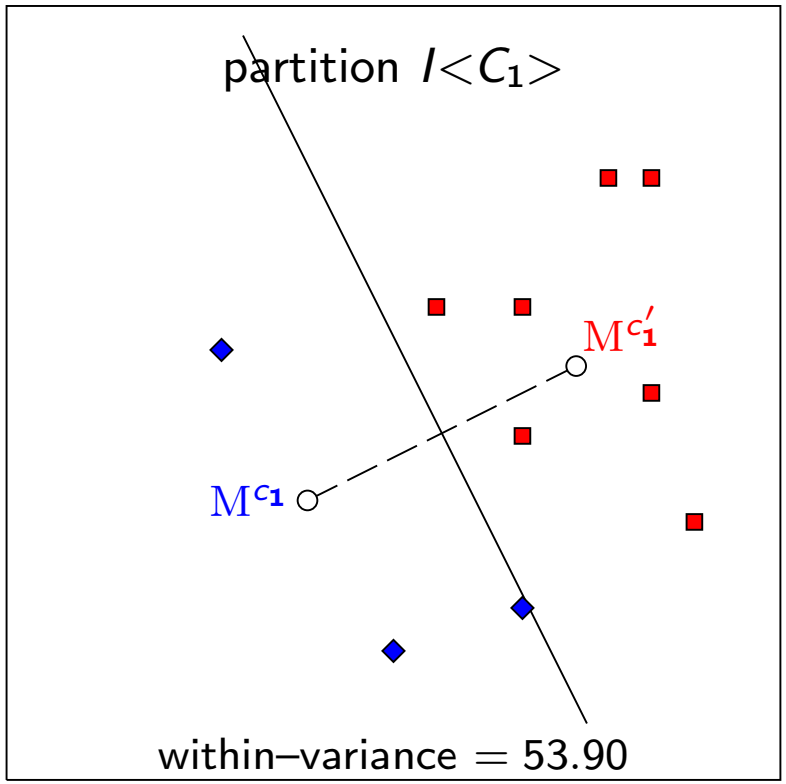
or aggregation around *moving centers*

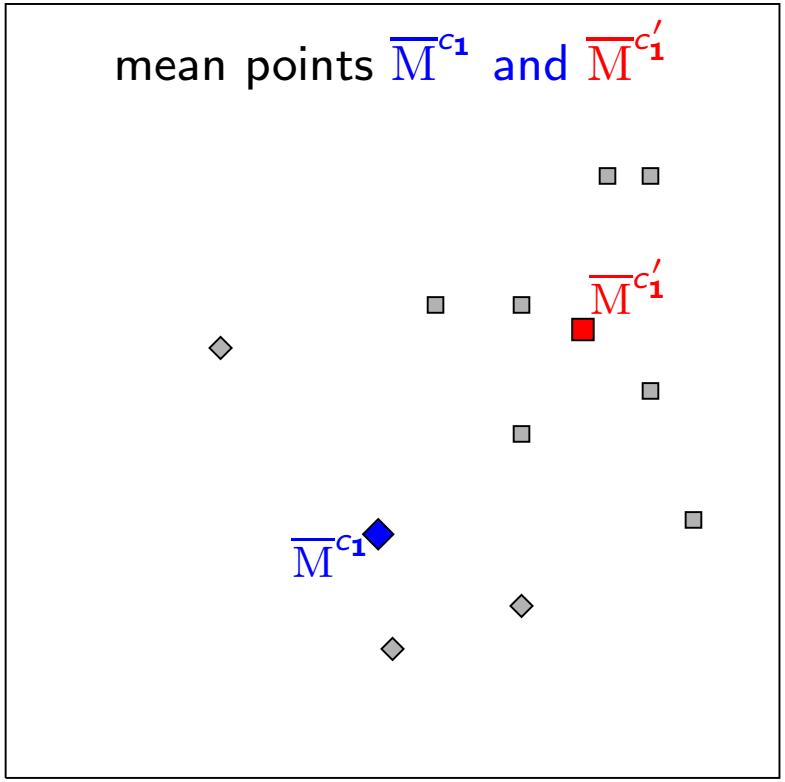
1. Fix the number of clusters, say C;
2. Choose (randomly or not) C initial class centers;
3. Assign each object to the closest center $\rightarrow$ new clusters;
4. Determine the centers of the new clusters;
5. Repeat the assignment;
6. Stop the algorithm when 2 successive iterations provide the same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

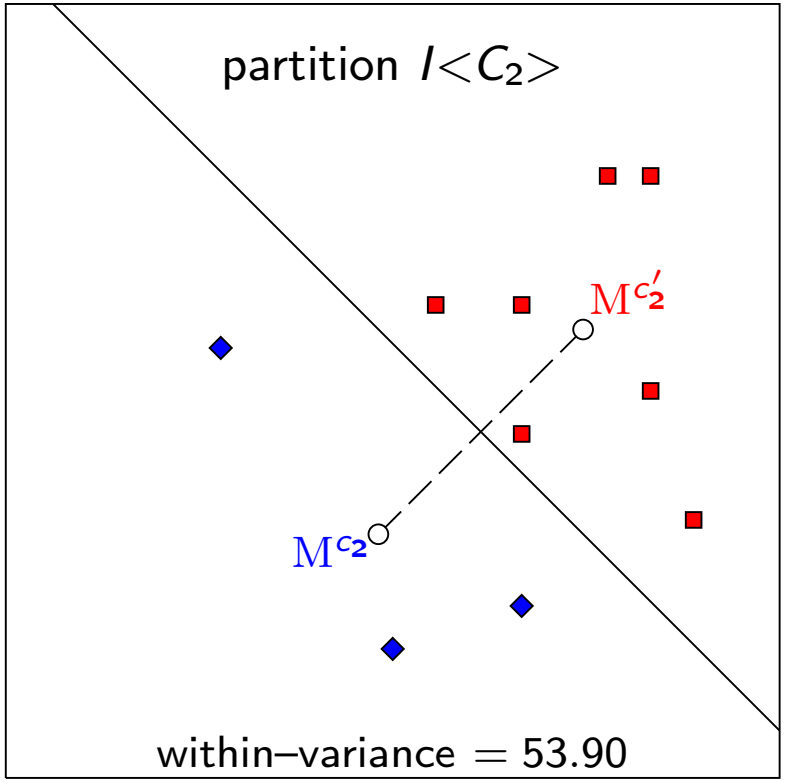
Choose 2 initial centers: M^{c_0} and $M^{c'_0}$




mean points $\overline{M}^{C_0}$ and $\overline{M}^{C_0'}$
$\overline{M}^{C_0'}$
$\overline{M}^{C_0}$


partition $I<C_1>$
$M^{C_1'}$
M^{C_1}
within–variance = 53.90





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

V.3. Ascending Hierarchical Clustering (AHC)

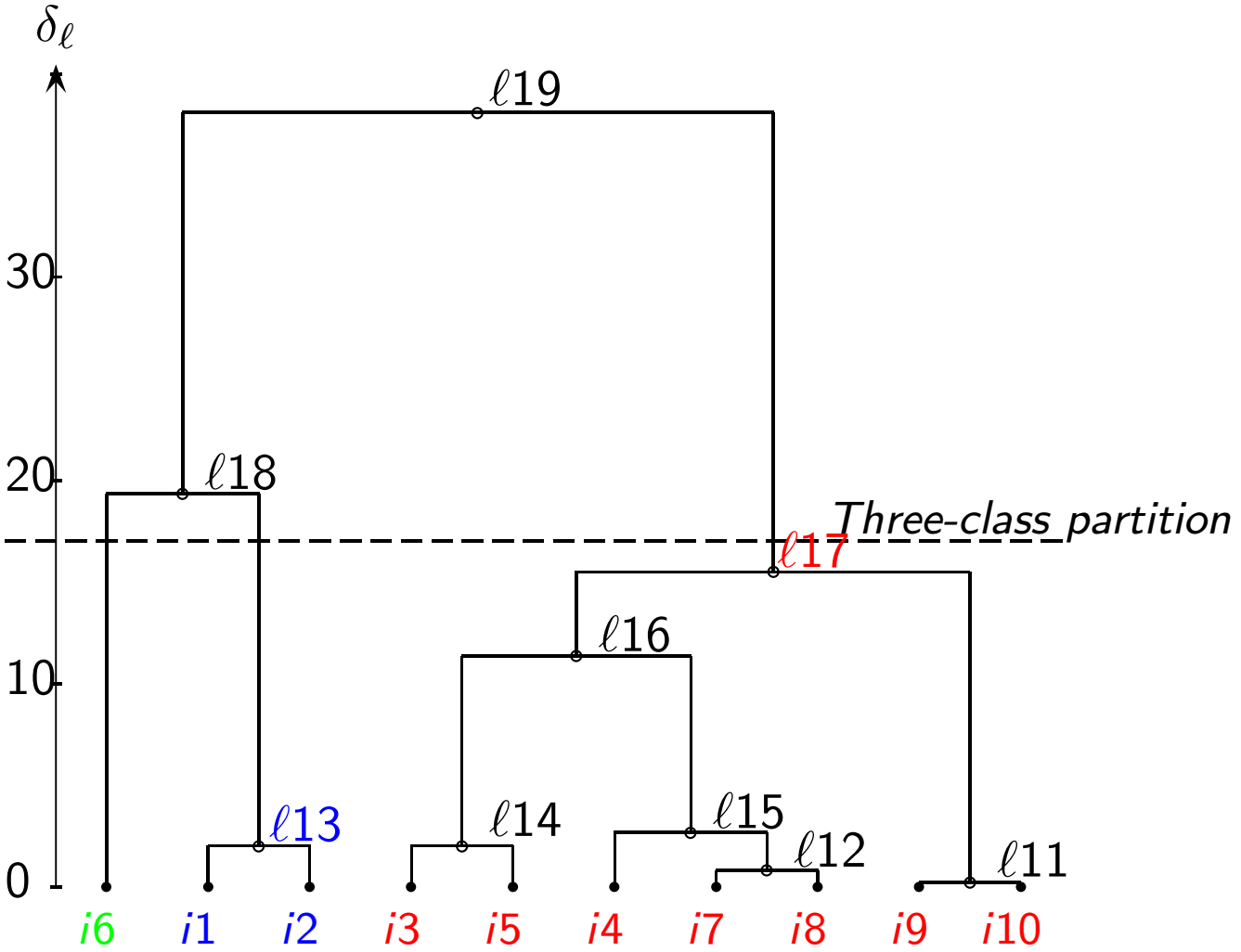
Clusters =

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".

V.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	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$i7$	$i8$	$i9$
$i2$	0.5								
$i3$	2.9	1.0							
$i4$	1.7	0.8	1.0						
$i5$	3.6	1.7	0.5	0.5					
$i6$	3.25	4.25	6.85	2.65	5.05				
$i7$	3.25	2.65	3.05	0.65	1.45	1.3			
$i8$	3.65	2.45	2.05	0.45	0.65	2.5	0.2		
$i9$	7.3	5.2	3.4	2	1.3	4.85	1.25	0.65	
$i10$	7.85	5.45	3.25	2.25	1.25	5.8	1.7	0.9	0.05

Minimum index 0.05 for the pair of points $\{i9, i10\}$ which are aggregated (fig. 1), hence the mean point ℓ_{11} and a new *cloud of 9 points* (fig. 2).

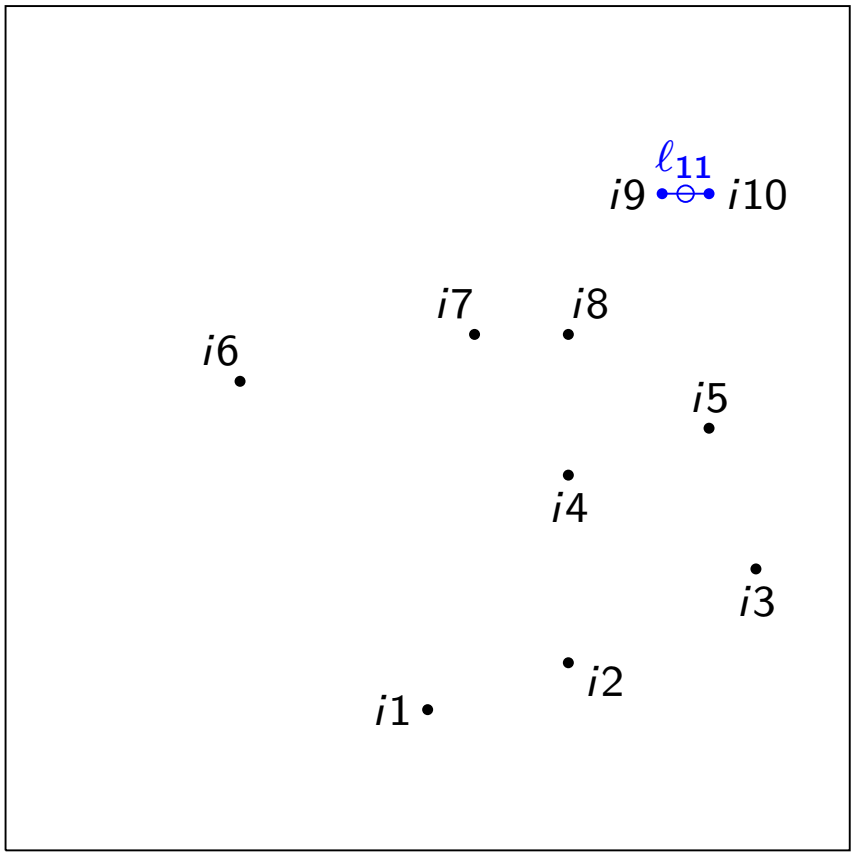
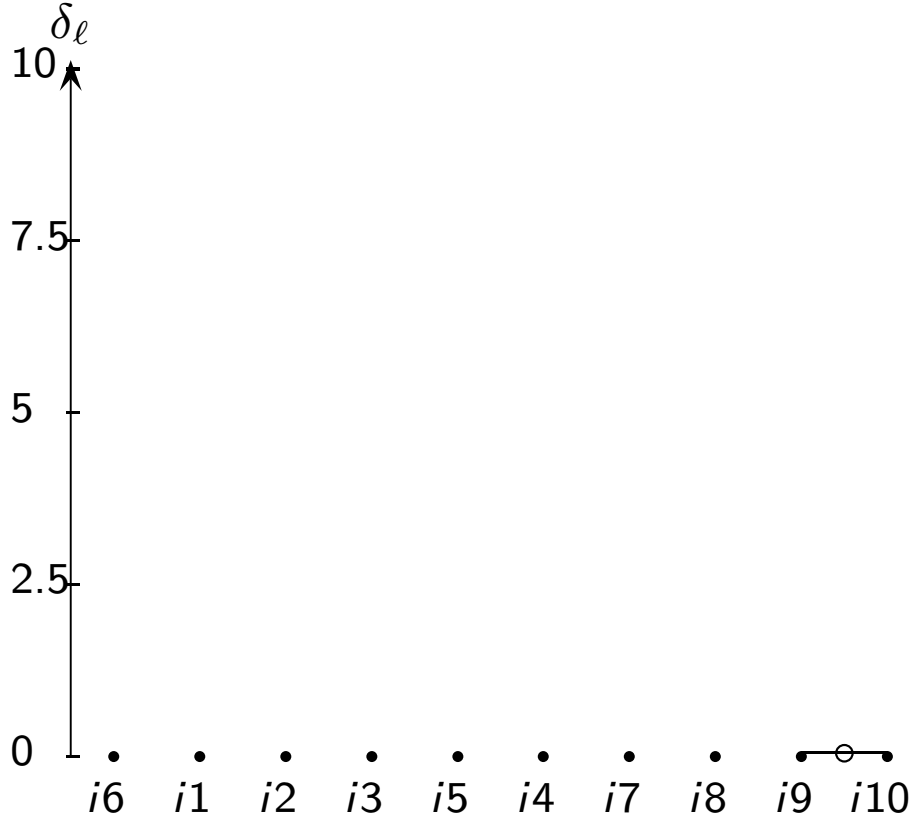


Figure 1



• **Step 2.** Calculate the aggregation index between the new point ℓ_{11} and the 8 other points. New minimum 0.2 for $\{i7, i8\}$ (fig. 2), hence the new point ℓ_{12} and a new *cloud of 8 points* (fig. 3).

	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$i7$	$i8$
ℓ_{11}	10.0825	7.0825	4.4175	2.8175	1.6825	7.0825	1.95	1.0175

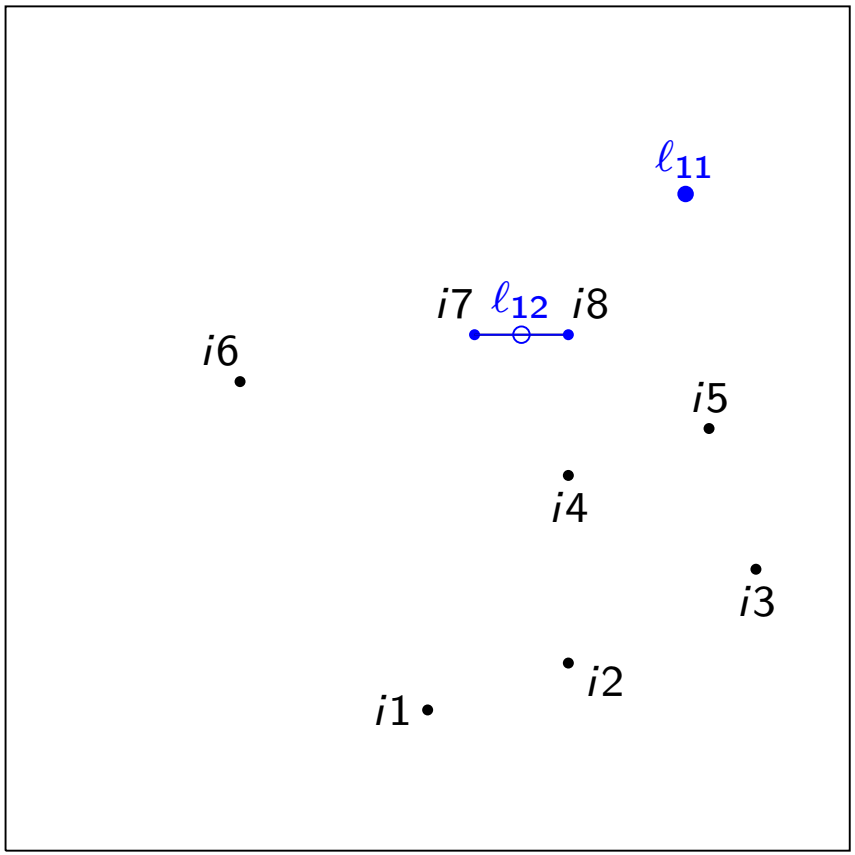
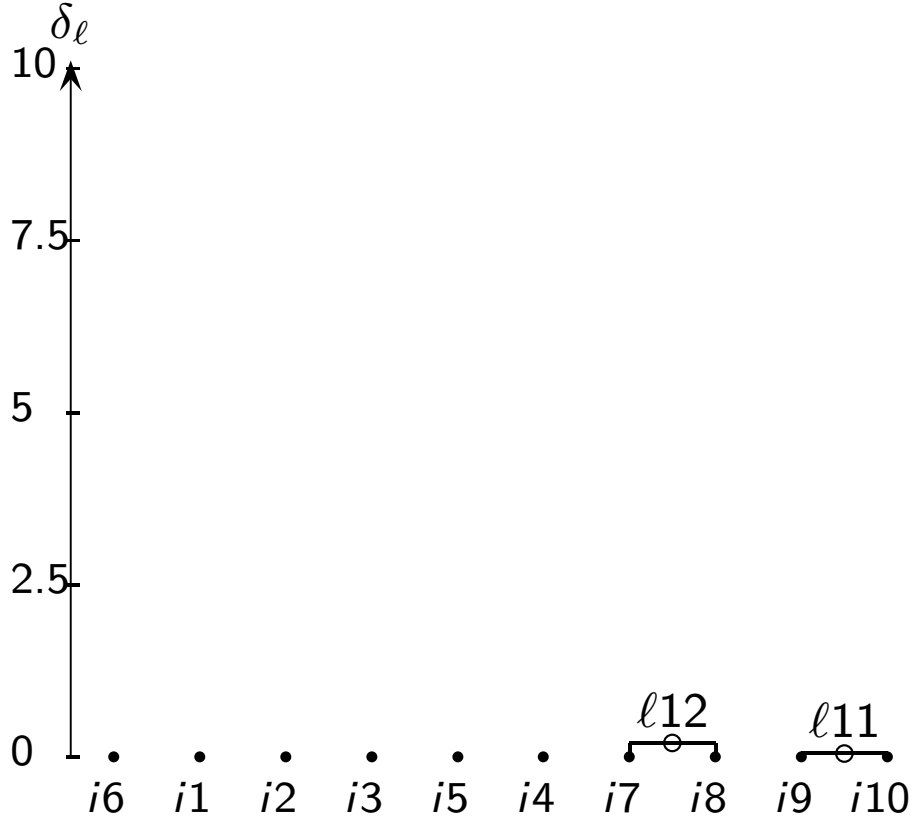


Figure 2



• **Step 3.** Iterate the procedure: aggregation index between $\ell 12$ and the 7 other points

	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$\ell 11$
ℓ_{12}	4.5325	3.3325	3.3325	0.6675	1.3325	2.4675	2.05

Minimum = 0.5 for $\{i1, i2\}$, $\{i3, i5\}$ and $\{i4, i5\}$, aggregation of $i1$ and $i2$ (fig. 3), hence the point $\ell 13$ and a *new cloud of 7 points* (fig. 4).

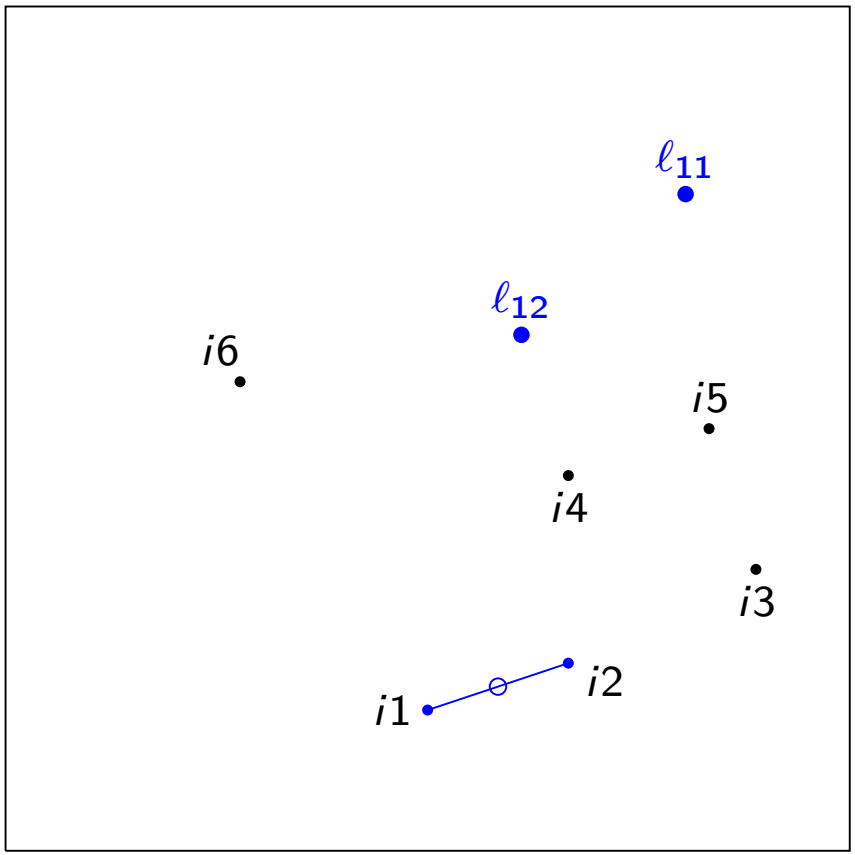
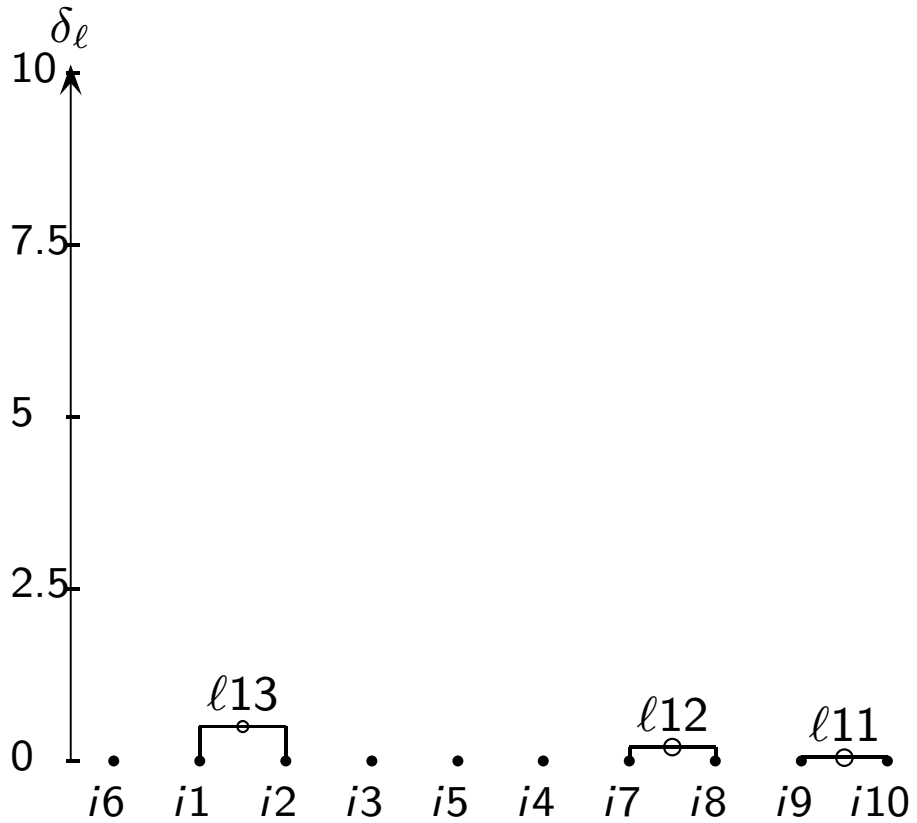


Figure 3



• **Step 4.** Iterate the procedure: aggregation index between $\ell 13$ and the 6 other points

	$i3$	$i4$	$i5$	$i6$	$\ell 11$	$\ell 12$
$\ell 13$	2.4325	1.5	3.3675	4.8325	12.625	5.65

Minimum of index = 0.5 for the two pairs $\{i3, i5\}$ and $\{i4, i5\}$.

Aggregation of $i3$ and $i5$ (fig. 4), hence the point ℓ_{14} and the *new cloud of 6 points* (fig. 5).

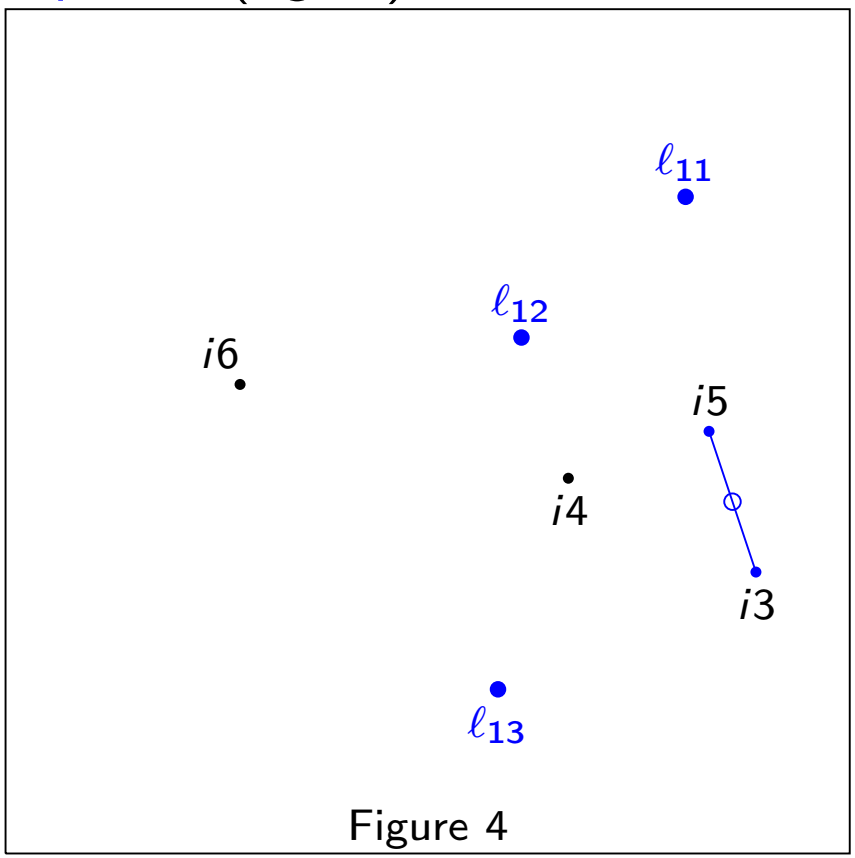
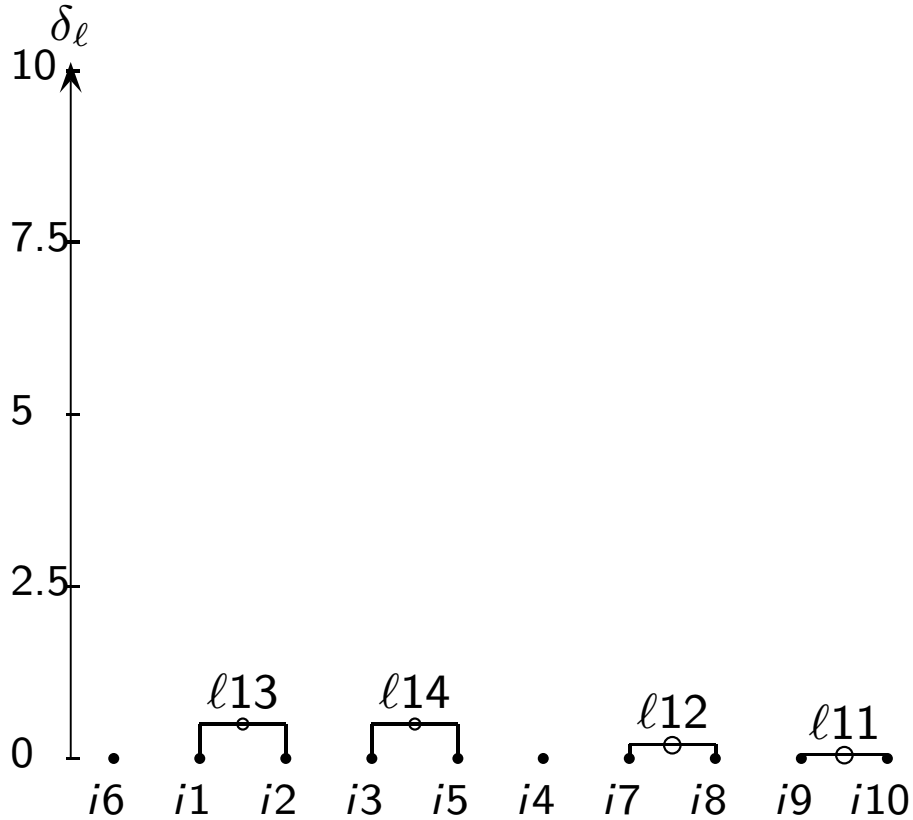


Figure 4



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

	$i4$	$i6$	ℓ_{11}	ℓ_{12}	ℓ_{13}
ℓ_{14}	0.8325	7.7675	4.3325	3.25	4.1

$\rightarrow$ aggregation of ℓ_{12} and $i4$ at level 0.6675 (fig. 5), hence the point ℓ_{15} and the *cloud of 5 points* (fig. 6).

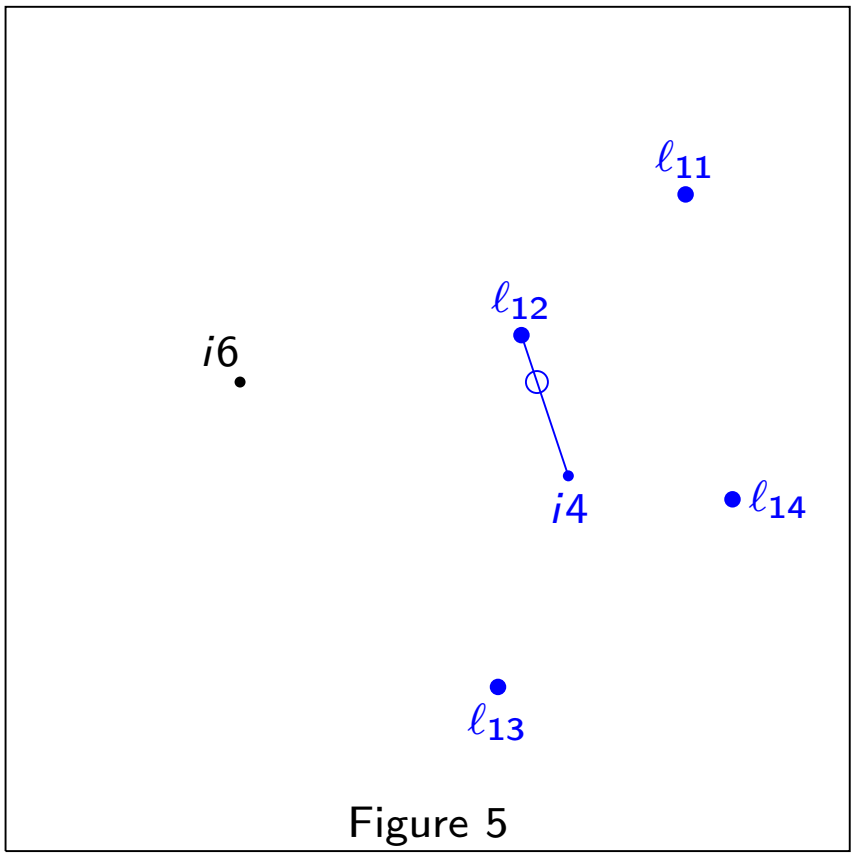
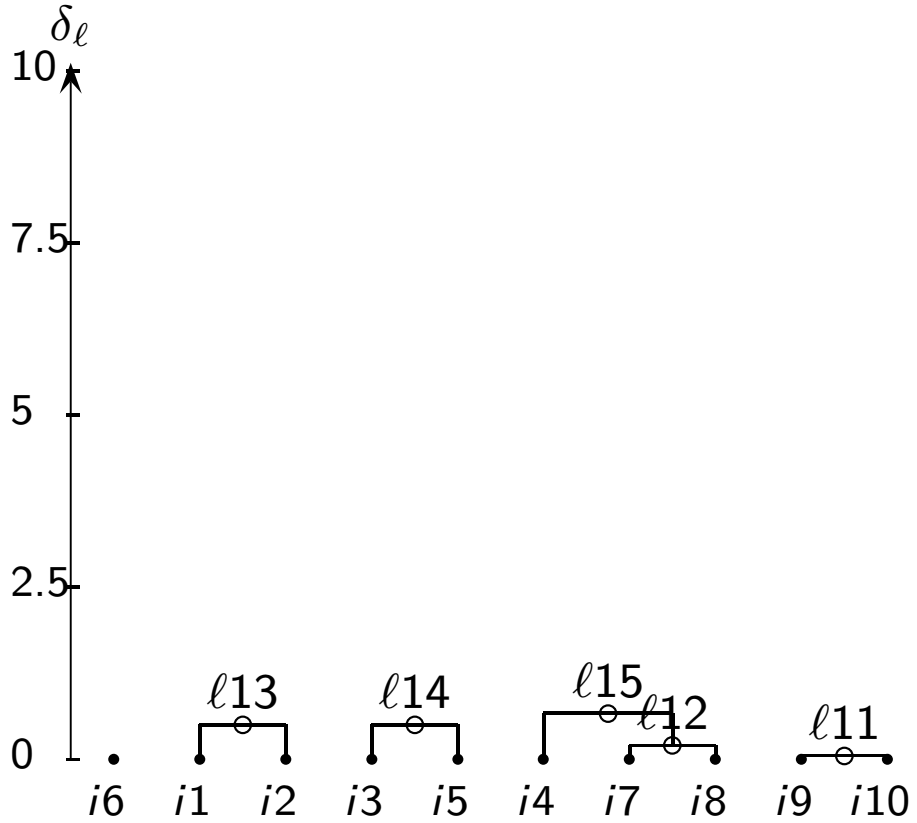


Figure 5



- **Step 6.** Aggregation index between $\ell 15$ and the 4 other points

	$i6$	ℓ_{11}	$\ell_{[13]}$	$\ell_{[14]}$
ℓ_{15}	3.0075	3.1225	5.1525	2.8325

$\rightarrow$ aggregation of ℓ_{15} and ℓ_{14} at level 2.8325 (fig. 6), hence the point ℓ_{16} and the *cloud of 4 points* (fig. 7).

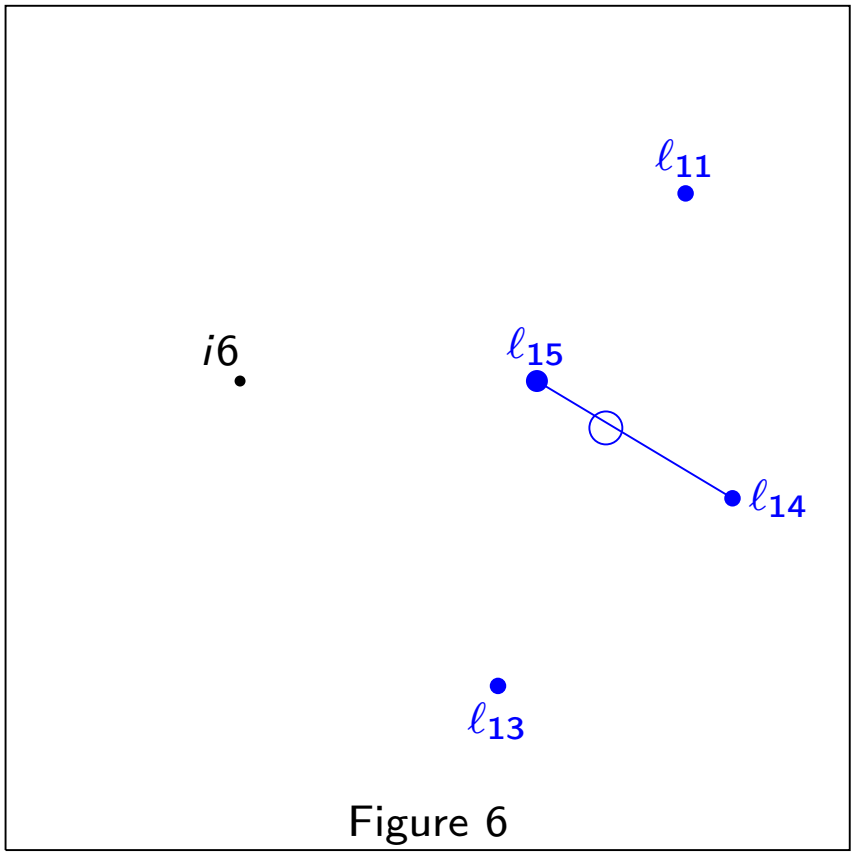
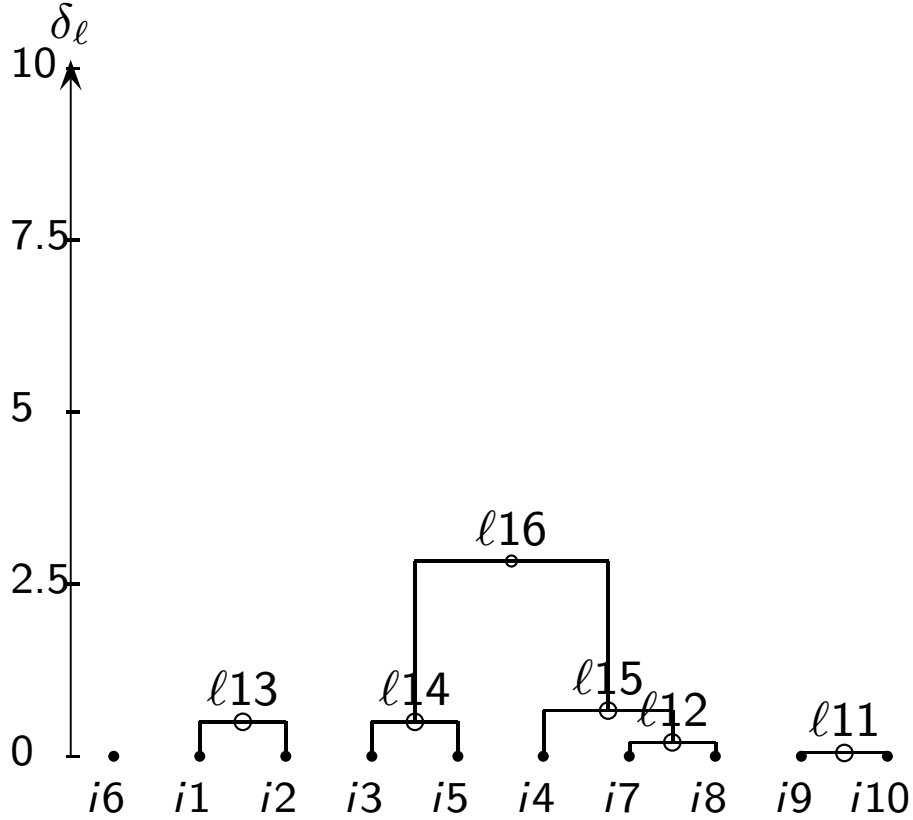


Figure 6



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

	$i6$	ℓ_{11}	ℓ_{13}
$\ell 16$	5.4175	3.8925	5.215

$\rightarrow$ aggregation of ℓ_{16} and ℓ_{11} at level 3.8925 (fig. 7), hence the point ℓ_{17} and the *cloud of 3 points* (fig. 8).

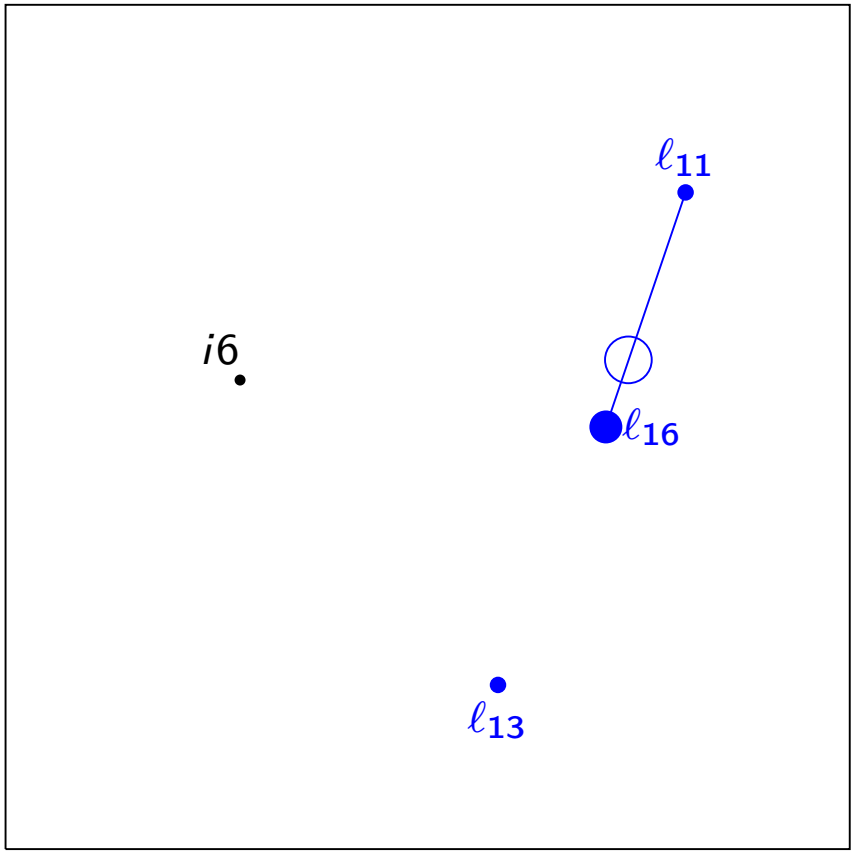
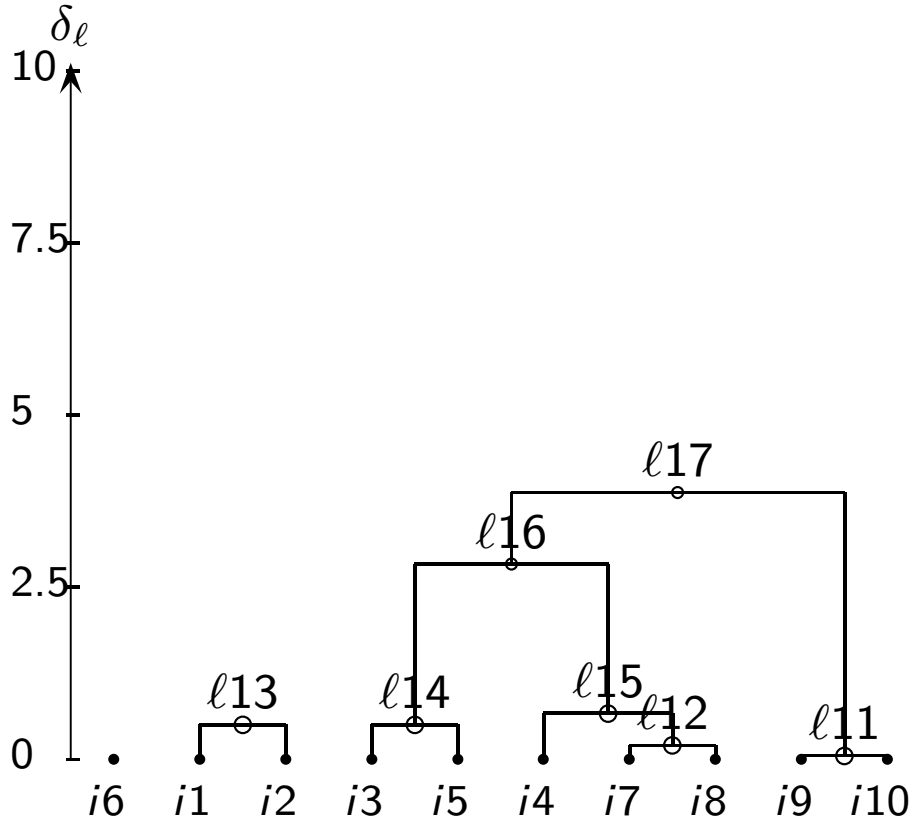


Figure 7



- Step 8.

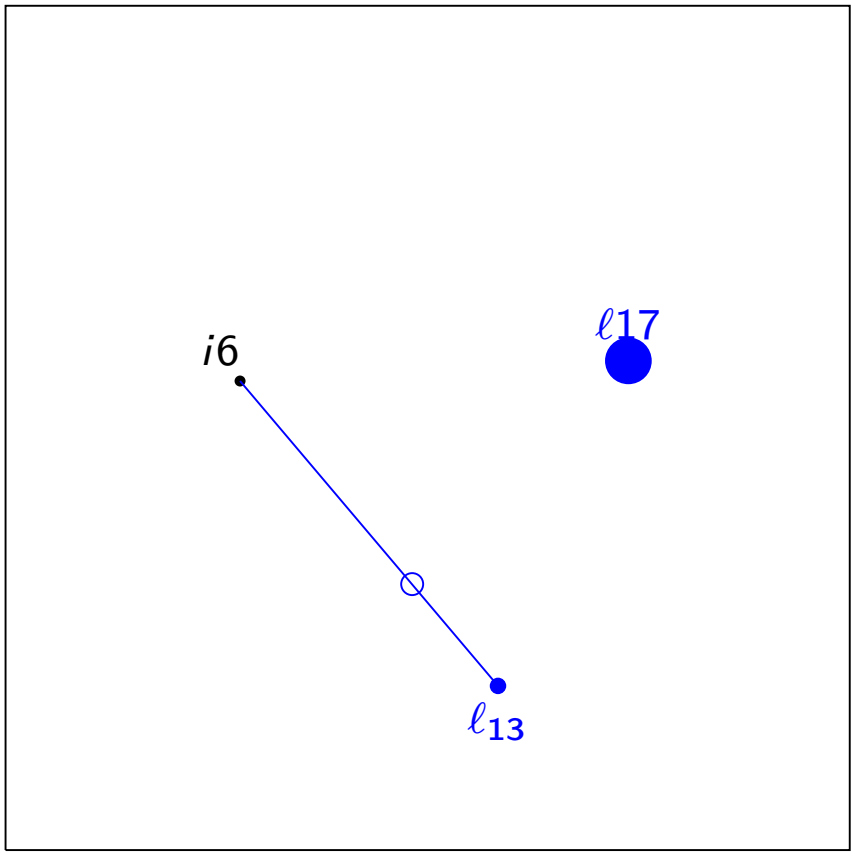
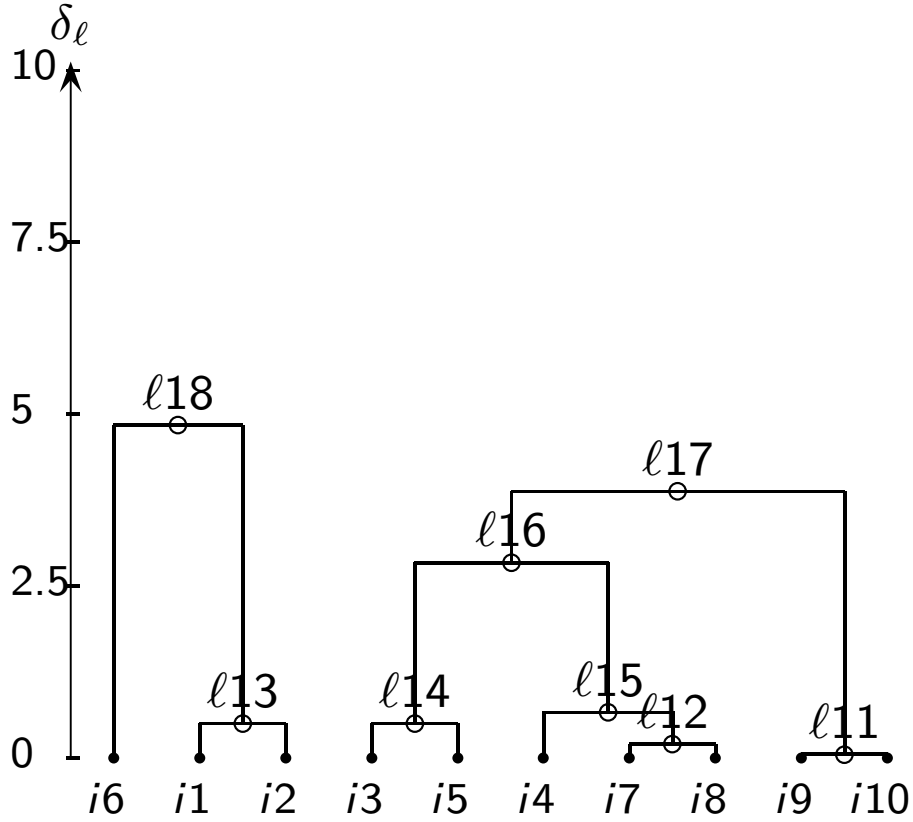


Figure 8



The three-class partition $\mathcal{A}$ ($\ell 14$), $\mathcal{B}$ ($i6$), $\mathcal{C}$ ($\ell 17$) (already studied) with mean points A ($\ell 13$), B ($i6$), C ($\ell 17$) (fig. 8).

- Step 9.

Contributions of the 3 pairs of points

	$\ell 13$	$i6$	$\ell 17$
$\ell 13$	—		
$i6$	4.8325	—	
$\ell 17$	8.8025	6.2325	—

$\Rightarrow$ grouping of i_6 and ℓ_{13} at level 4.8325 (fig. 9).

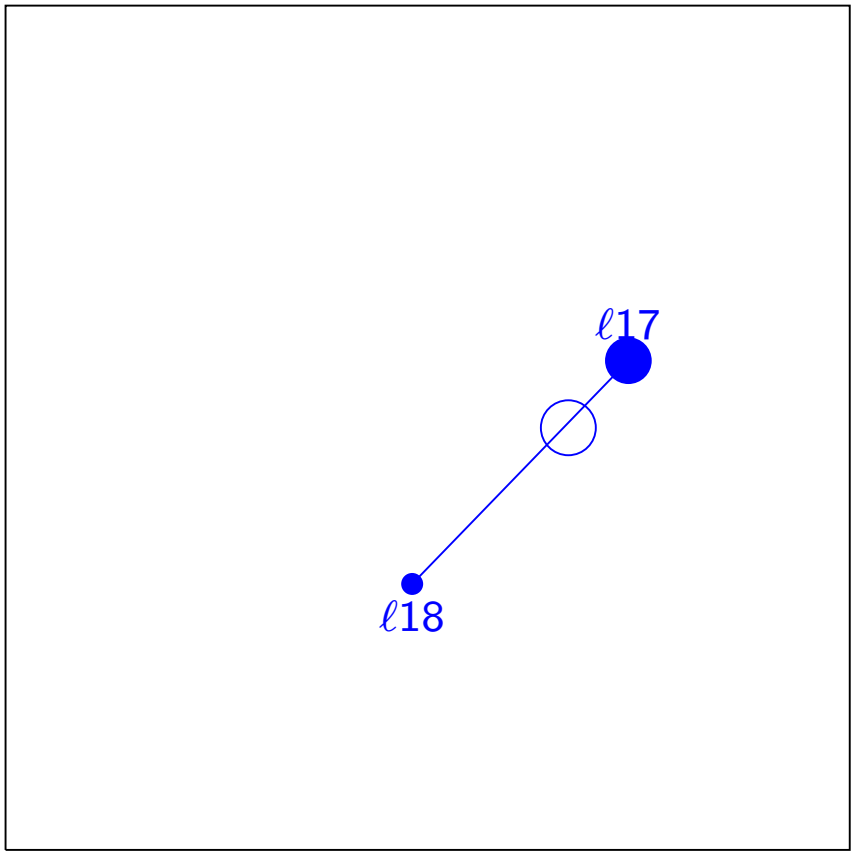
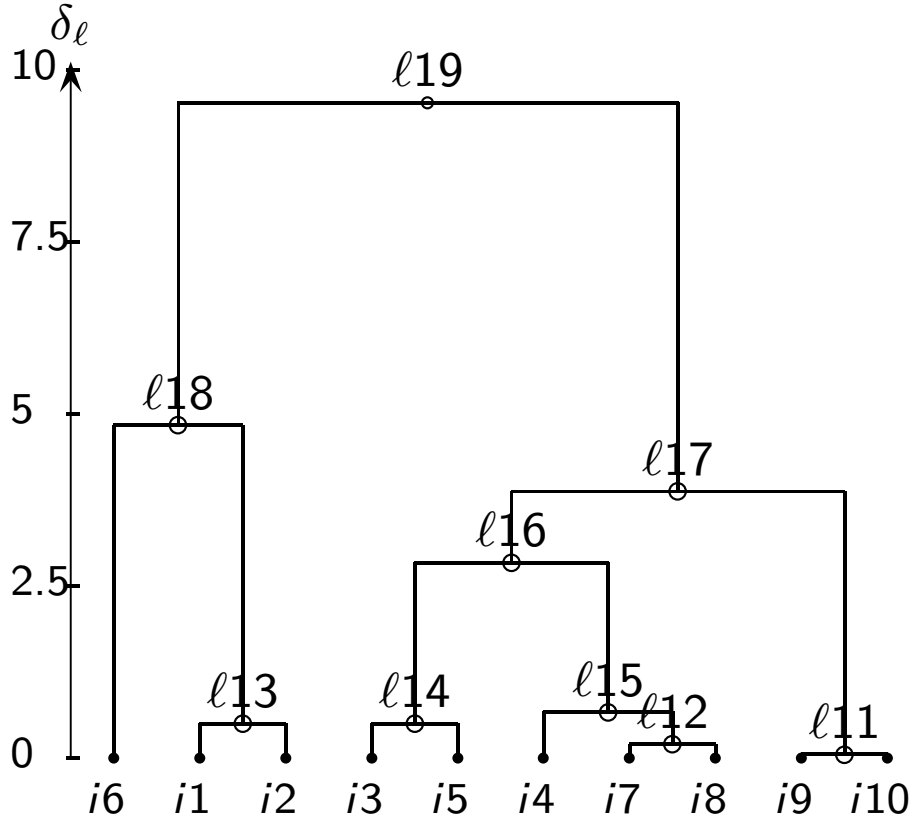
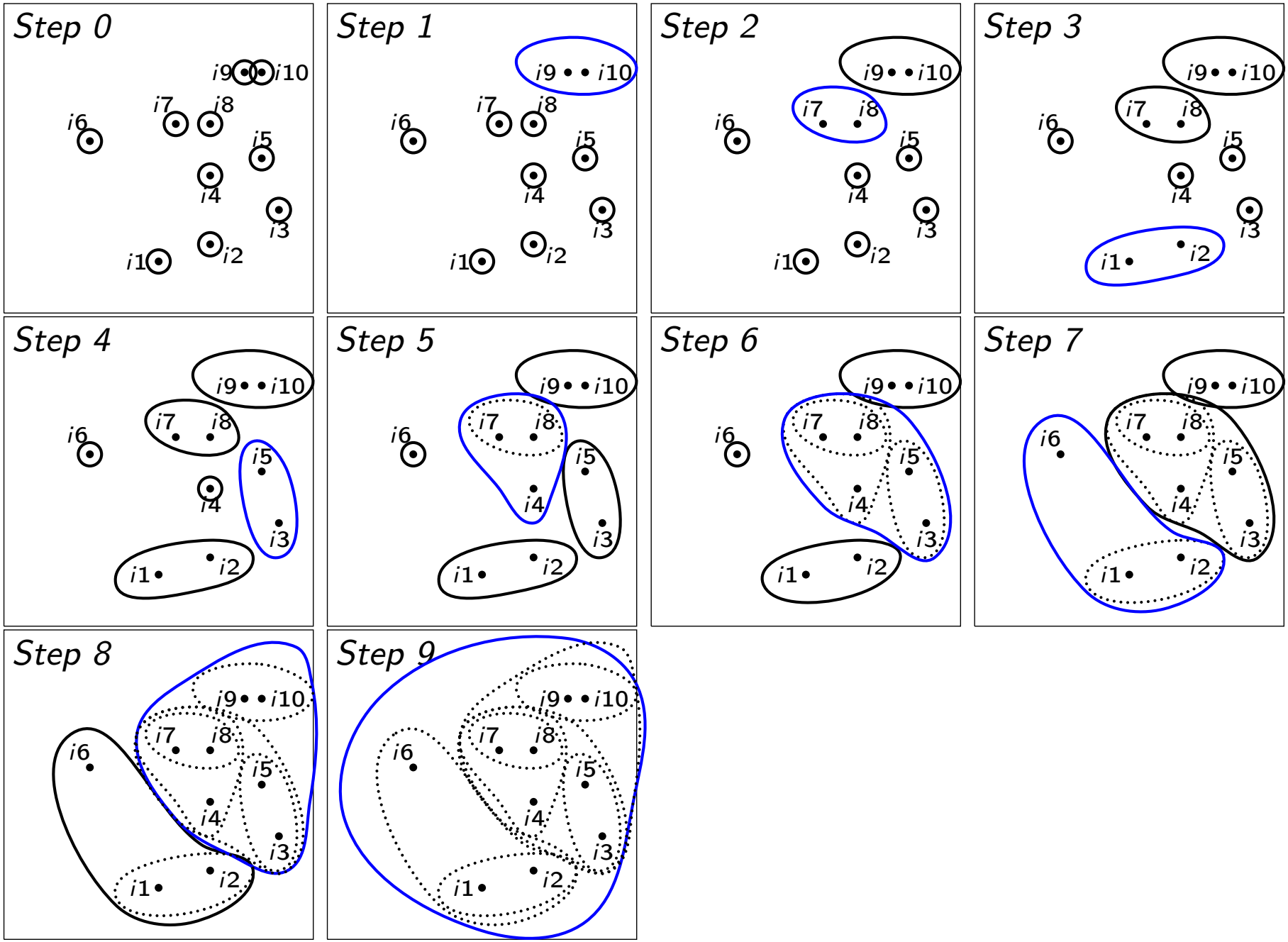


Figure 9




Step 0
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
Step 9
i1
i2
i3
i4
i5
i6
i7
i8
i9 • • i10

Successive Steps of the AHC

ℓ	δ_ℓ	clusters		n	class description
ℓ_{19}	9.52375	ℓ_{18}	ℓ_{17}	10	i_9 i_{10} i_3 i_5 i_4 i_7 i_8 i_6 i_1 i_2
ℓ_{18}	4.83325	ℓ_{13}	ℓ_6	3	i_6 i_1 i_2
ℓ_{17}	3.89275	ℓ_{16}	ℓ_{11}	7	$i9$ i_{10} i_3 i_5 i_4 i_7 i_8
ℓ_{16}	2.83325	ℓ_{15}	ℓ_{14}	5	i_3 i_5 i_4 i_7 i_8
ℓ_{15}	0.66675	ℓ_{12}	ℓ_4	3	i_4 i_7 i_8
ℓ_{14}	0.5	ℓ_5	ℓ_3	2	i_3 i_5
ℓ_{13}	0.5	ℓ_2	ℓ_1	2	i_1 i_2
ℓ_{12}	0.2	ℓ_8	ℓ_7	2	i_7 i_8
ℓ_{11}	0.05	ℓ_{10}	ℓ_9	2	i_9 i_10

	Between Var.	η^2_ℓ
ℓ_{19}	9.52375	.414
ℓ_{18}	14.3575	.624
ℓ_{17}	18.25	.793
ℓ_{16}	21.0825	.917
ℓ_{15}	21.75	.957
ℓ_{14}	22.25	.967
ℓ_{13}	22.75	.989
ℓ_{12}	22.95	.998
ℓ_{11}	23.00	1

Sum of the 9 level indexes = 23 (variance of the cloud).

Between-variance of the 2-class partition = 9.52375.

Between-variance of the 3-class partition = 9.52375 + 4.83325 = 14.3575, etc.

V.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.

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

Categories under-represented: $f_c - f < -0.05 \qquad f_c/f < 1/2$

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

the procedure
— is based on the successive dichotomies of the hierarchy,
— takes into account the geometric properties of the clusters, particularly their shape.

Reference:
F. Cassor & B Le Roux. Assigning Changes Over Time Using GDA Methods: Application to the French "Barometer of Political Trust", Chapter 19 (pp. 325–344), in *Empirical Investigation of Social Space*; J. Blasius, F. Lebaron, B. Le Roux, A. Schmitz (*Eds*): Springer Nature (2020).

Mixed Clustering

Consolidation of a partition

To improve a partition, we can use the method of aggregation around the moving centres (K-means) by taking as initial centre-points the mean points of the clusters of the partition obtained by AHC.

Mixed classification of large data sets

1. Draw a random sample from the set of n objects to be classified (say between 5 000 and 10 000 objects) and perform a AHC;
2. Perform a partition in a large number of clusters (say 500) of the random sample, and choose in each cluster the object close to the centre;
3. Carry out a mixed classification: proceed to a K-means clustering of the n objects by taking the 500 objects determined in the previous step as the initial centres of the clusters, then carry out the AHC of the 500 mean-points of the clusters obtained by the K-means method;
4. Decide from this hierarchy the (small) number of clusters of the final partition.

V.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*.

V.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

References

— Bourdieu, P. (1999) Une révolution conservatrice dans l'édition, *Actes de la recherche en Sciences Sociales*, 126–127 (A conservative revolution in publishing, *Translation studies*, 2008).

— Lebaron, F., Le Roux B. (eds), 2015, *La méthodologie de Pierre Bourdieu en action. Espace culturel, espace social et analyse des données*, Paris: Dunod.

— Le Roux B. (2014). *Analyse géométrique des données multidimensionnelles*, chap.10 (pp. 321–342), Paris: Dunod

— Le Roux, B., Cassor, F. (2015). Assigning Objects to Classes of a Euclidean Ascending Hierarchical Clustering in *Statistical learning and Data Sciences*; Eds Gammerman, A., Vovk, V., Papadopoulos, H., Springer.

— Le Roux, B., Rouanet, H.(2013). Geometric Data Analysis of Gifted Students' Individual Differences, Chapter 3, in *Individual Differences in Online Computer-based Learning: Gifted and Other Diverse Populations*, Suppes P. (editor), pp. 129–157, Stanford, CA: CSLI Publications.