IV — Introduction to Euclidean Classification

Brigitte.LeRoux@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/~lerb/
rouanet@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/~rouanet/

1 What is Euclidean Classification?

1.1 Introduction

a) Partition and hierarchy

Hierarchical classification: System of nested classes (the paradigm of natural sciences) represented by a hierarchical tree.

b) Qualities of classification: compactness and separability

c) Descending (or divisive) classification vs ascending (or agglomerative) classification

1.2 Ascending Hierarchical Classification (AHC) using Variance Criterion

Grouping property: If 2 classes are grouped together, the between–variance decreases from an amount equal to the contribution of the dipole defined by the centers of the 2 grouped classes.

Target Example (see II): Three–class partition $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ with between-variance 57.43 (variance of the cloud of 3 mean points (A,B,C) of classes). If $\mathcal{A}$ and $\mathcal{B}$ are grouped, the between–variance of the partition in 2 classes, that is, the variance of the cloud of 2 points (barycenter of A and B, C) is equal to 38.10.

Within-contribution of the pair (A,B): $\frac{\widetilde{n_{\mathrm{AB}}}}{n} \times (\mathrm{AB})^2 = 19.33$, with $\mathrm{AB}^2 = 290$ and $\widetilde{n_{\mathrm{AB}}} = \frac{1}{\frac{1}{2}+\frac{1}{1}} = 2/3$ (weight of dipole);

One has: $38.10 = 57.43 - 19.33$ (grouping property).

Ascending Hierarchical Classification: starting with the basic objects (one–element classes) proceed to successive aggregations, until all objects are grouped in a single class.

At each step, one groups 2 classes of the current partition.

Euclidean classification:

1. Objects = *points of Euclidean cloud*: distance between objets is Euclidean distance.

2. *Aggregation index* = variance index, that is, the contribution of the dipole associated with the 2 aggregated classes (Ward index).

At each step, the aggregated classes are those which lead to the minimal decrease of the between-variance.

1.3 Basic Algorithm

• **Step 1**. Calculate the contributions of the $9 \times 10/2 = 45$ dipoles

δ	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$i7$	$i8$	$i9$
$i2$	2								
$i3$	11.6	4							
$i4$	6.8	3.2	4						
$i5$	14.4	6.8	2	2					
$i6$	13	17	27.4	10.6	20.2				
$i7$	13	10.6	12.2	2.6	5.8	5.2			
$i8$	14.6	9.8	8.2	1.8	2.6	10	0.8		
$i9$	29.2	20.8	13.6	8	5.2	19.4	5	2.6	
$i10$	31.4	21.8	13	9	5	23.2	6.8	3.6	**0.2**

Example: For dipole $\{i1, i2\}$: $\widetilde{n_{12}} = 1/(\frac{1}{1} + \frac{1}{1}) = 0.5$, squared distance= $(0-6)^2 + (-12+10)^2 = 40$, hence the absolute contribution of dipole $\frac{0.5}{10} \times 40 = 2$.

Minimum index 0.2 for the pair of points $\{i9, i10\}$ which are aggregated (fig. 1), hence the mean point $\ell 11$ and a derived *cloud of 9 points* (fig. 2).

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

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

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

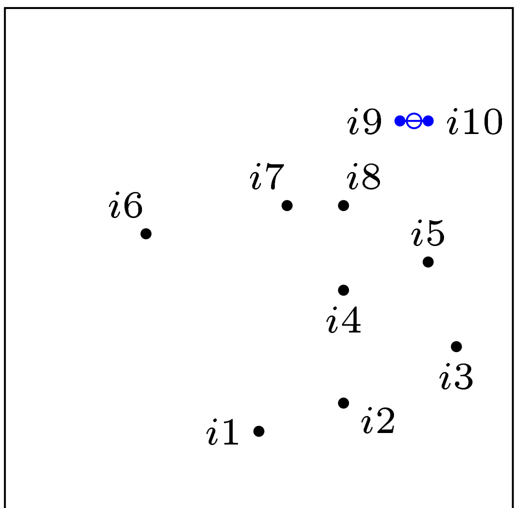


Figure 1

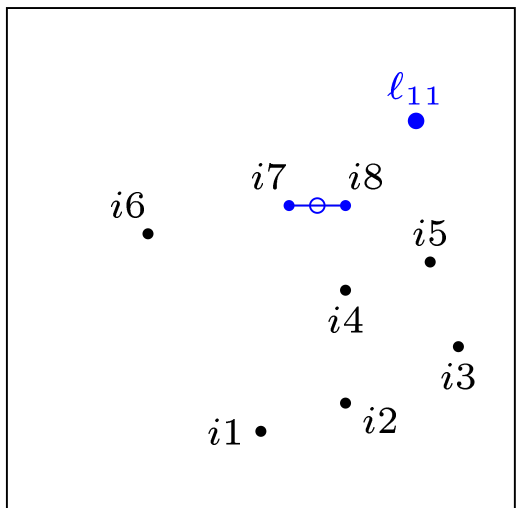


Figure 2

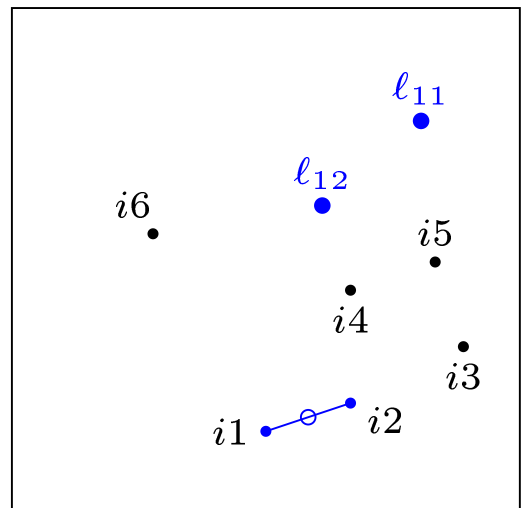


Figure 3

• **Step 3.** Iterate the procedure

Aggregation index between $\ell 12$ and the 7 other points

	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$\ell 11$
$\ell 12$	18.13	13.33	13.33	2.67	5.33	9.87	8.2

Minimum of index = 2 for the three pairs $\{i1, i2\}$, $\{i3, i5\}$ and $\{i4, i5\}$. We choose[a] to aggregate $i1$ and $i2$ (fig. 3), hence the point $\ell 13$ and a *cloud of 7 points* (fig. 4).

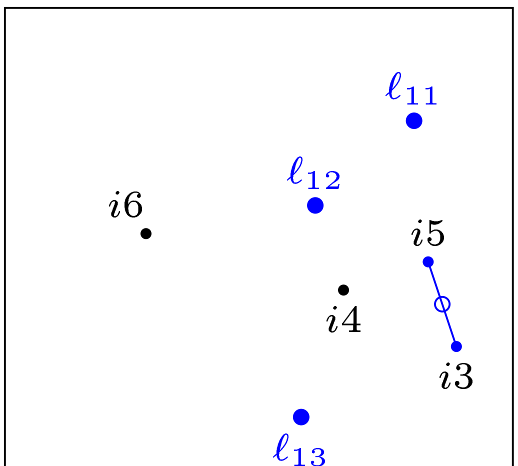


Figure 4

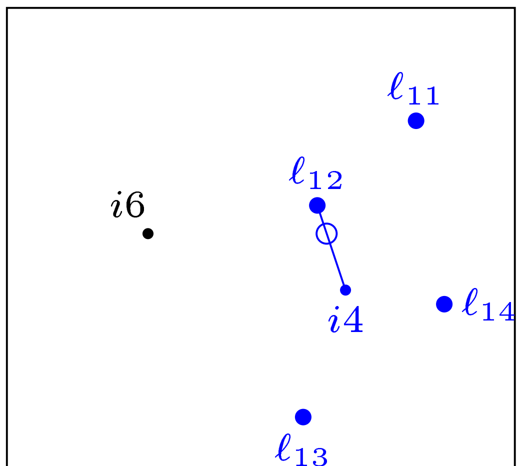


Figure 5

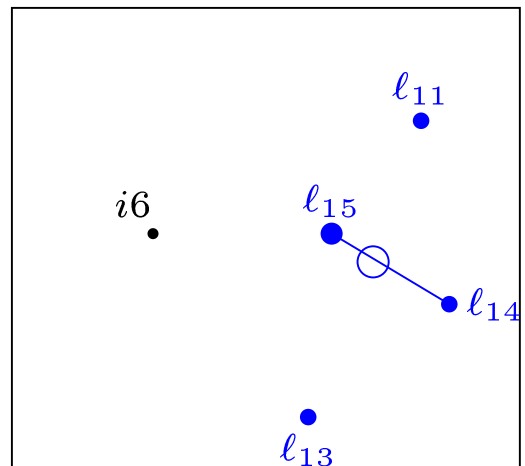


Figure 6

[a]In indeterminate cases different choices may yield different classifications.

Aggregation index between $\ell 13$ and the 6 other points

	$i3$	$i4$	$i5$	$i6$	$\ell 11$	$\ell 12$
$\ell 13$	9.73	6.00	13.47	19.33	50.5	22.6

Minimum of index = 2 for the two pairs $\{i3, i5\}$ and $\{i4, i5\}$. We choose to aggregate $i3$ and $i5$ (fig. 4), hence the point $\ell 14$ and the *cloud of 6 points* (fig. 5).

Aggregation index between $\ell 14$ and the 5 other points

	$i4$	$i6$	$\ell 11$	$\ell 12$	$\ell 13$
$\ell 14$	3.33	31.07	17.33	13.00	16.4

→ aggregation of $\ell 12$ and $i4$ at level 2.67 (fig. 5), hence the point $\ell 15$ and the *cloud of 5 points* (fig. 6).

Aggregation index between $\ell 15$ and the 4 other points

	$i6$	$\ell 11$	$\ell 13$	$\ell 14$
$\ell 15$	12.03	12.49	20.61	11.33

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

Aggregation index between $\ell 16$ and the 3 other points

	$i6$	$\ell 11$	$\ell 13$
$\ell 16$	21.67	15.57	20.86

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

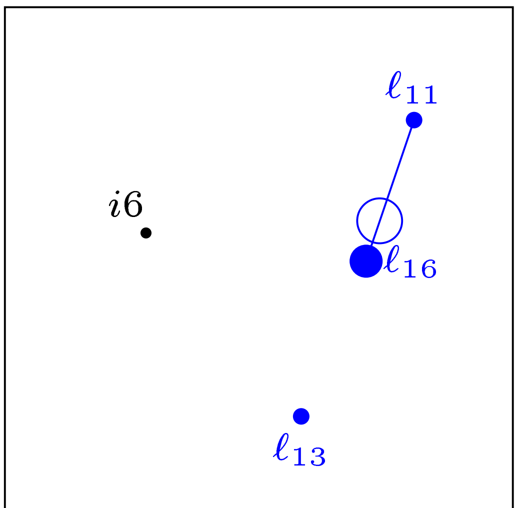


Figure 7

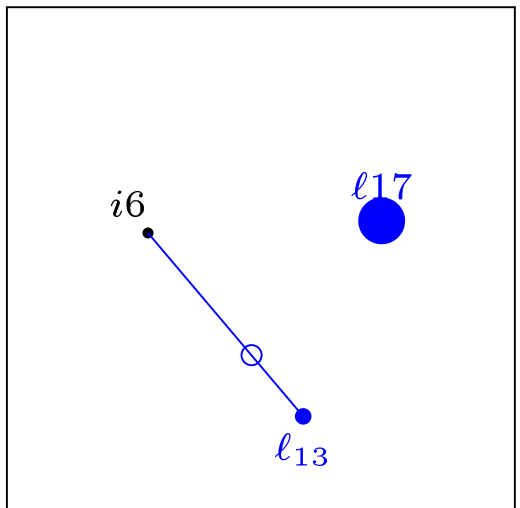


Figure 8

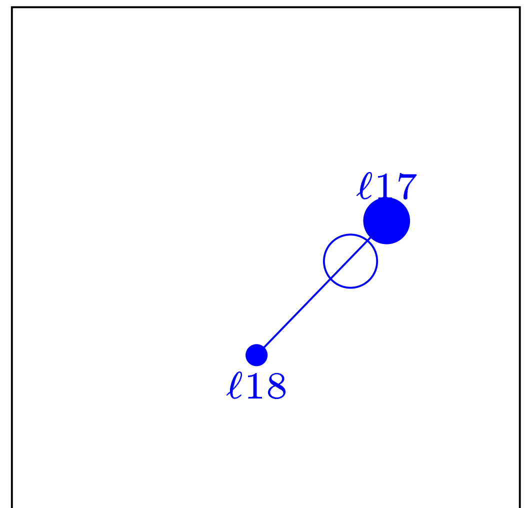


Figure 9

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

Table of the within-contributions of the 3 pairs of points

(distance)^2	weight	Contribution
$\text{AB}^2 = 290$	$\widetilde{n_{\text{AB}}} = \frac{1}{\frac{1}{2}+\frac{1}{1}} = 2/3$	$\text{Cta}_{(\text{A},\text{B})} = \frac{2/3}{10} \times 290 = 19.33$
$\text{AC}^2 = 226.33$	$\widetilde{n_{\text{AC}}} = \frac{1}{\frac{1}{2}+\frac{1}{7}} = 14/9$	$\text{Cta}_{(\text{A},\text{C})} = \frac{14/9}{10} \times 226.33 = 35.21$
$\text{BC}^2 = 284.90$	$\widetilde{n_{\text{BC}}} = \frac{1}{\frac{1}{1}+\frac{1}{7}} = 7/8$	$\text{Cta}_{(\text{B},\text{C})} = \frac{7/8}{10} \times 284.90 = 24.93$

At this step, we group A and B at level 19.33 (fig. 9).

Successive steps of the AHC

ℓ	δ_ℓ	classes		n	class description
$\ell 19$	38.095	$\ell 18$	$\ell 17$	10	$i9\, i10\, i3\, i5\, i4\, i7\, i8\, i6\, i1\, i2$
$\ell 18$	19.333	$\ell 13$	$\ell 6$	3	$i6\, i1\, i2$
$\ell 17$	15.571	$\ell 16$	$\ell 11$	7	$i9\, i10\, i3\, i5\, i4\, i7\, i8$
$\ell 16$	11.333	$\ell 15$	$\ell 14$	5	$i3\, i5\, i4\, i7\, i8$
$\ell 15$	2.667	$\ell 12$	$\ell 4$	3	$i4\, i7\, i8$
$\ell 14$	2.	$\ell 5$	$\ell 3$	2	$i3\, i5$
$\ell 13$	2.	$\ell 2$	$\ell 1$	2	$i1\, i2$
$\ell 12$	0.8	$\ell 8$	$\ell 7$	2	$i7\, i8$
$\ell 11$	0.2	$\ell 10$	$\ell 9$	2	$i9\, i10$

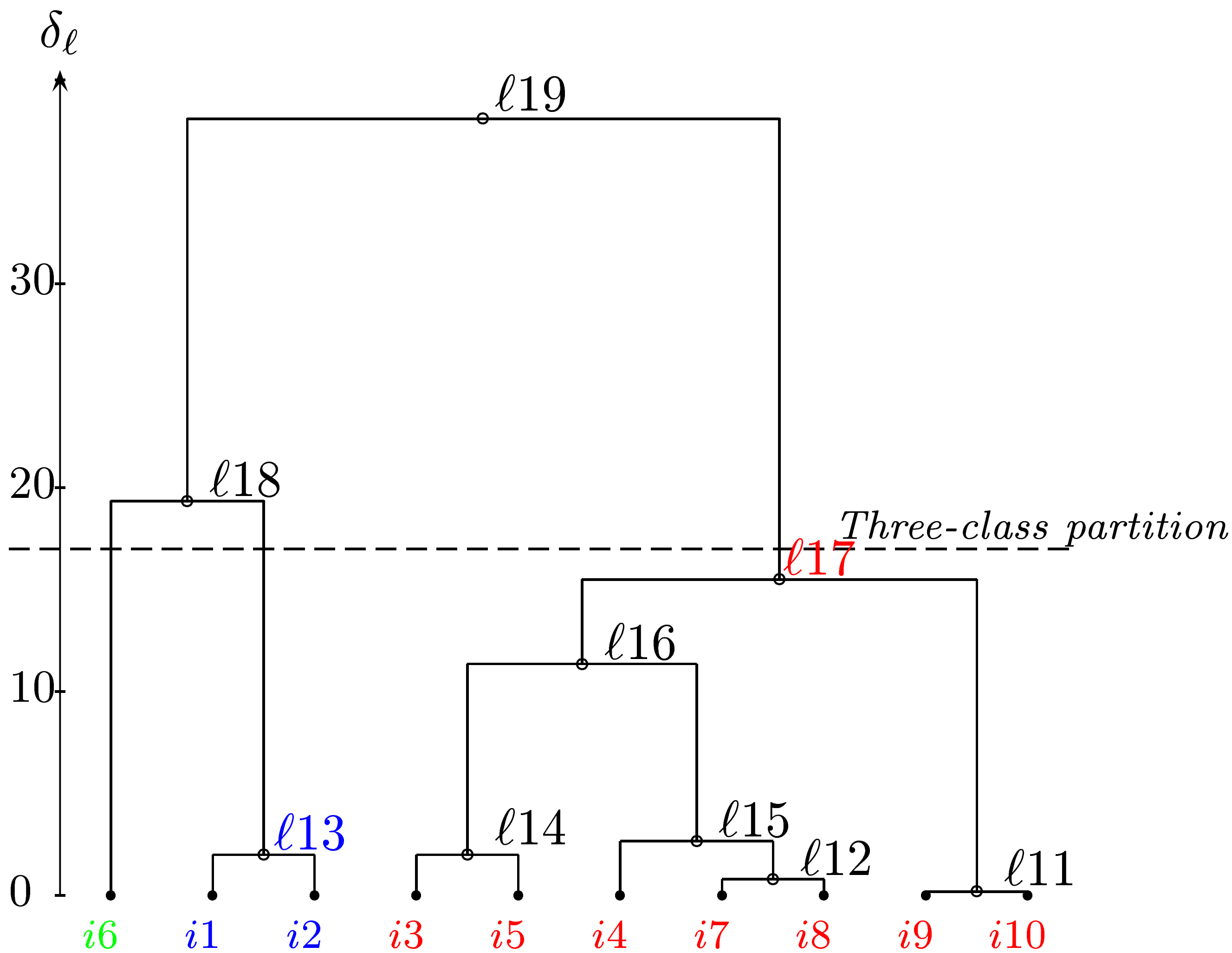
Between Var		η^2_ℓ
$\ell 19$	38.10	.414
$\ell 18$	57.43	.624
$\ell 17$	73.00	.793
$\ell 16$	84.33	.917
$\ell 15$	87.00	.957
$\ell 14$	89.00	.967
$\ell 13$	91.90	.989
$\ell 12$	91.80	.998
$\ell 11$	92.00	1

The sum of the nine level indices δ_ℓ is 92 (total 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.

Target example: hierarchical tree



References

— Benzécri J-P. (1992) *Correspondence Analysis Handbook*, (Part V), New York: Dekker (p. 561-635).

— Bourdieu P. (1999). Une révolution conservatrice dans l'édition [A conservative revolution in publishing], *Actes de la Recherche en Sciences Sociales*, Vol. 126-127, 3-28.

— Le Roux B. & Rouanet H. (2003). Geometric Analysis of Individual Differences in Mathematical Performance for epgy Students in the Third Grade. `www-epgy.stanford.edu/research/`.

— Le Roux B. & Rouanet H. (2004), *Geometric Data Analysis: from Correspondence Analysis to Structured Data Analysis* (chapter 3, p.106-116), Dordrecht: Kluwer.