Geometric Data Analysis

Brigitte Le Roux

Brigitte.LeRoux@mi.parisdescartes.fr
www.mi.parisdescartes.fr/∼lerb/

[1]MAP5/CNRS, Université Paris Descartes

[2]CEVIPOF/CNRS, SciencesPo Paris

GDA course — Sept. 12-16, 2016 — Uppsala

Table of Contents I

Table of Contents II

Table of Contents III

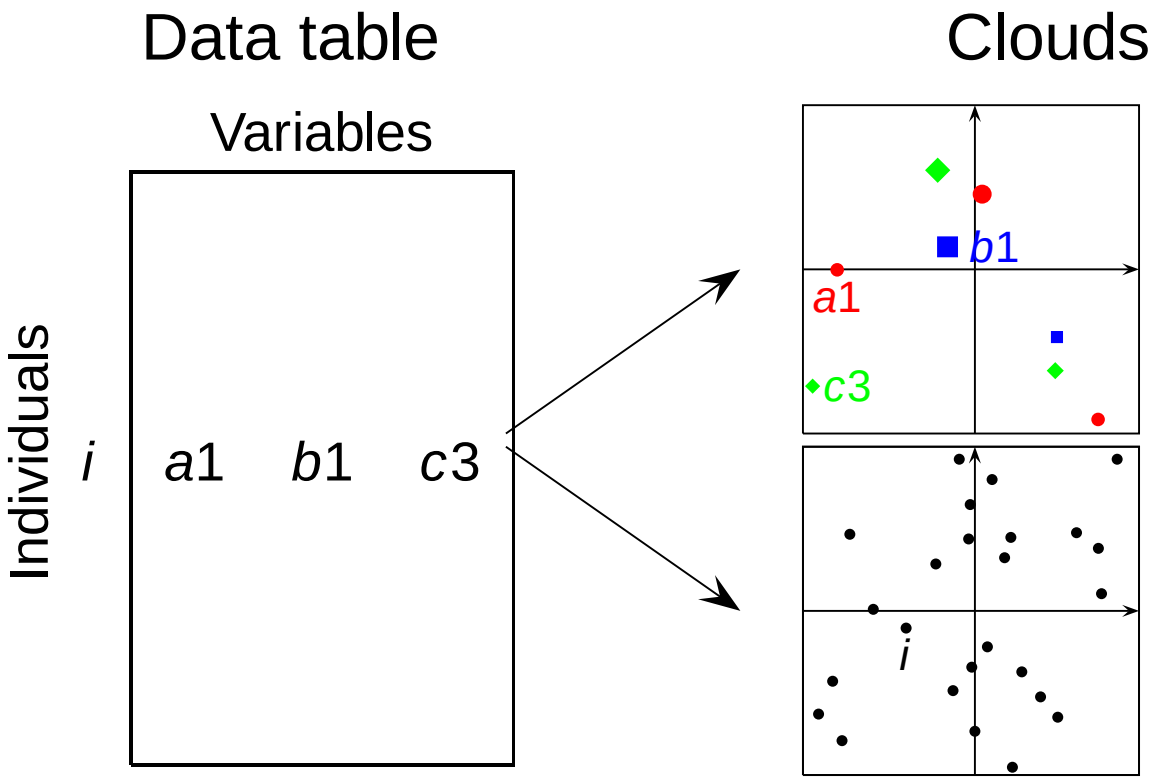
I – Introduction

Brigitte Le Roux

Brigitte.LeRoux@mi.parisdescartes.fr
www.mi.parisdescartes.fr/∼lerb/

I.1. Three Key Ideas

- *Geometric modeling*



Cloud of categories: Points represent the categories of variables.

Cloud of individuals: Points represent individuals.

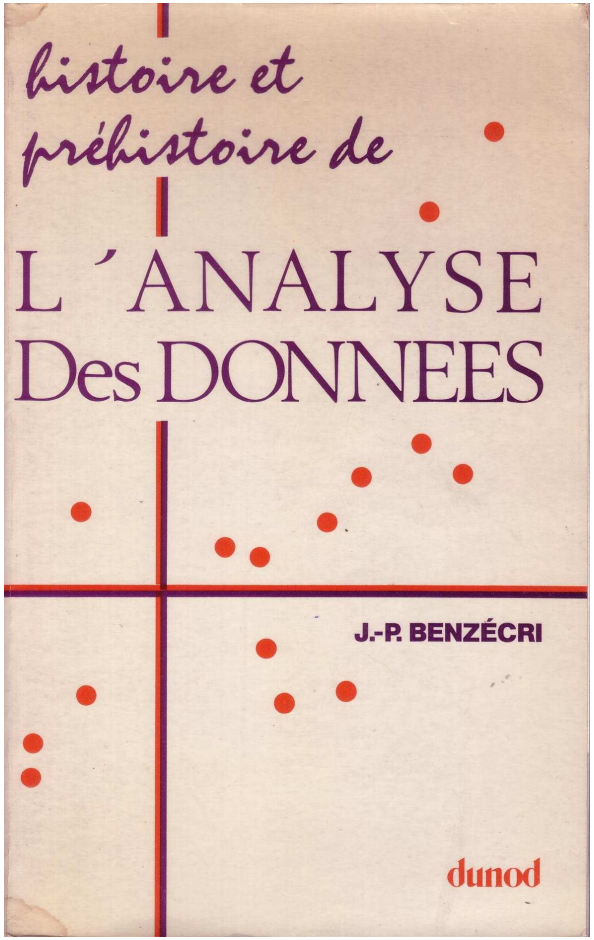
- *Formal approach.*
- *Description first!*

The model should follow the data, not the reverse!"

I.2. Three Paradigms

- *Correspondence Analysis* (CA)
  $\longrightarrow$ Contingency table
- *Principal Component Analysis* (PCA)
  $\longrightarrow$ Individuals$\times$Numerical Variables table
- *Multiple Correspondence Analysis* (MCA)
  $\longrightarrow$ Individuals$\times$Categorical Variables table

I.3. Historical Sketch



J-P. Benzécri (1982)

Precursors

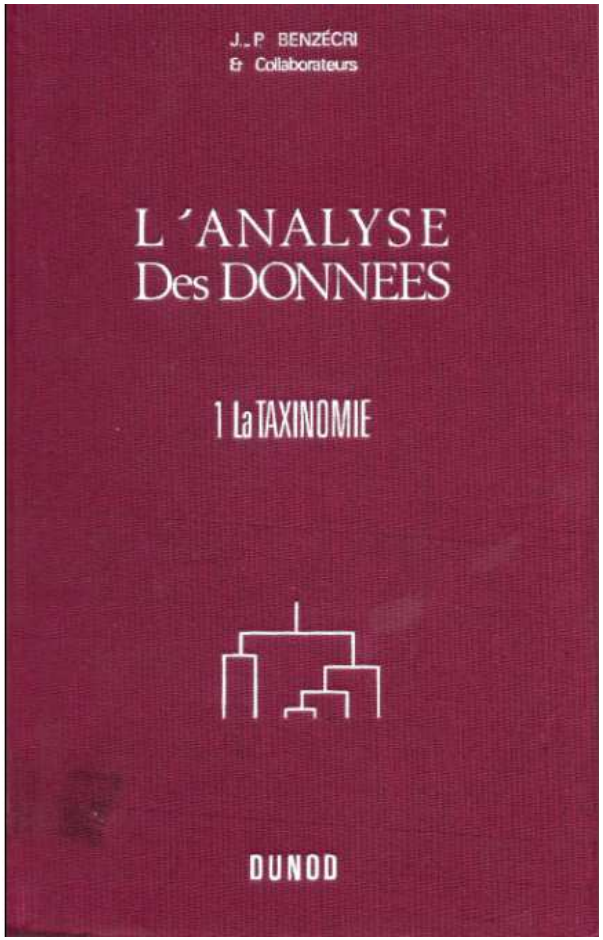
Karl Pearson (1901), Hirschfeld (1935).

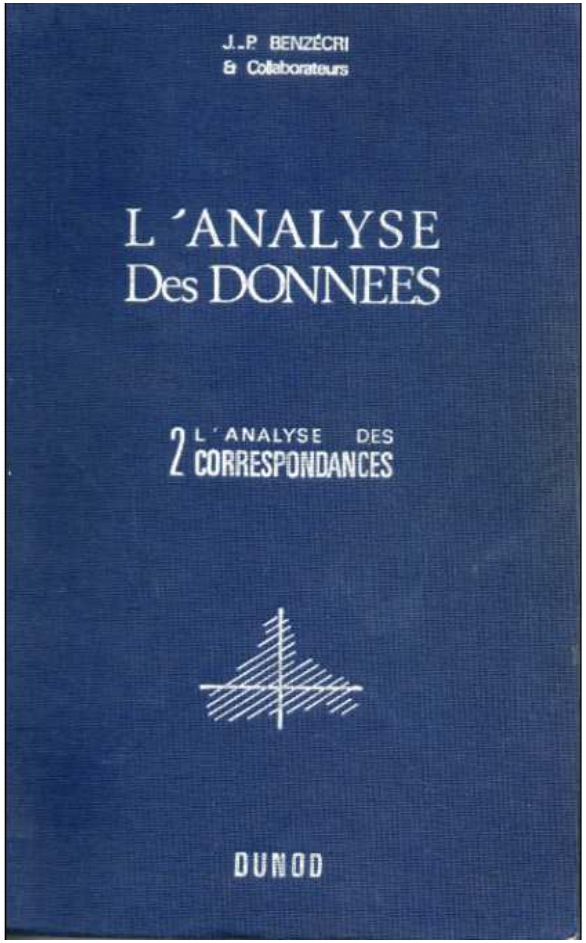
Should we need an Anglo–Saxon patronage for "Analyse des Données", we would be pleased to turn to the great Karl Pearson.

Benzécri (1982), p. 116

- Optimal scaling: Fisher (1940), Guttman (1942)
- Factor analysis: Burt (1950)
- Quantification method: Hayashi (1952)
- MDS: Shepard (1962).

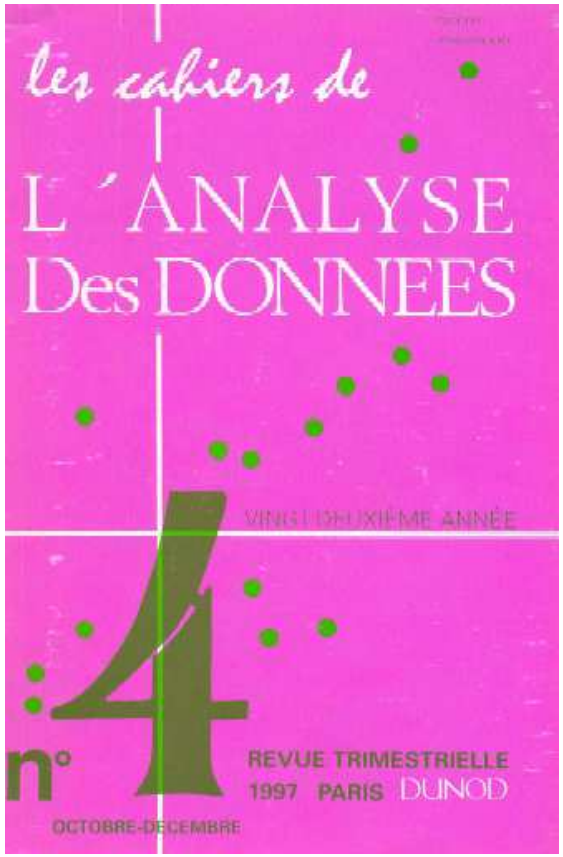
Emergence (1963-73)





Benzécri et al. (1973): L'ANALYSE Des DONNÉES
1 la TAXINOMIE 2 L'ANALYSE DES CORRESPONDANCES.

Recognition and splendid isolation (1973-1980



1977–1997

Gower (1966), Good (1969), Gabriel (1971)
Ignored in Shepard, Romney, Nerlove (1972), Kruskal & Wish (1978), Shepard (1980) and in Kendall & Stuart (1976)

International recognition (since 1981)

Greenacre (1984), Lebart & al (1984), Jambu (1991),

Benzécri (1992) (translation of the introductory book published by Dunod in 1984);

Malinvaud (1980), Deville & Malinvaud (1983):

"Econometrics without stochastic models"

Tenenhaus & Young (1985) : Psychometry;

Nishisato (1980): Dual Scaling;

Gifi (1981/1990): Homogeneity Analysis;

Carroll & Green (1988), Weller & Romney (1990): MDS group;

Gower & Hand (1996): biplot.

Where do we stand now?

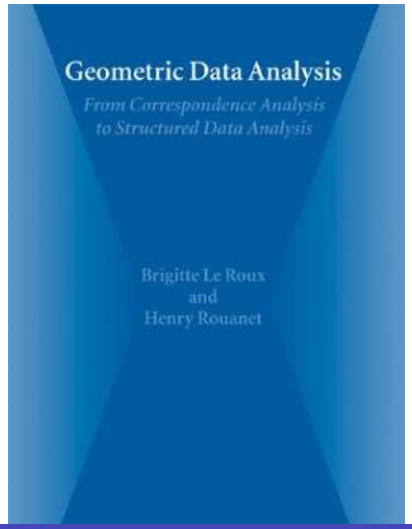
Goodman (1986, 1991), Weller & Romney (1990), Rao (1995).

CARME network: international conferences in Cologne (1991, 1995, 1999), Barcelona (2003), Rotterdam (2007), Rennes (2011), Naples (2015) . . .

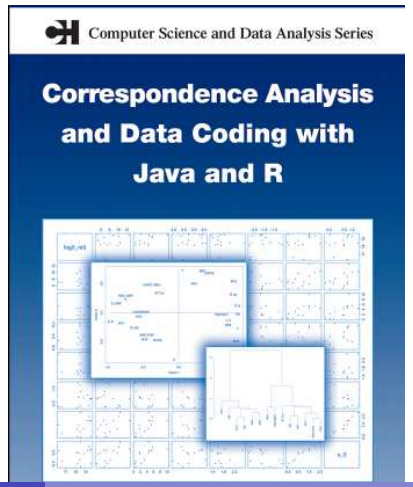
Workshops organized in Paris, Uppsala, Copenhagen, Montreux, London, Kaliningrad, Mendoza, Berkeley . . .

Recent Books:

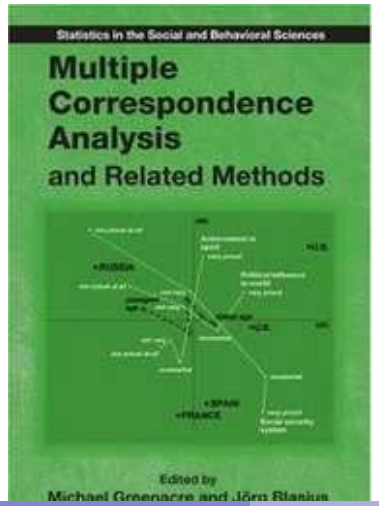
Le Roux & Rouanet 2004



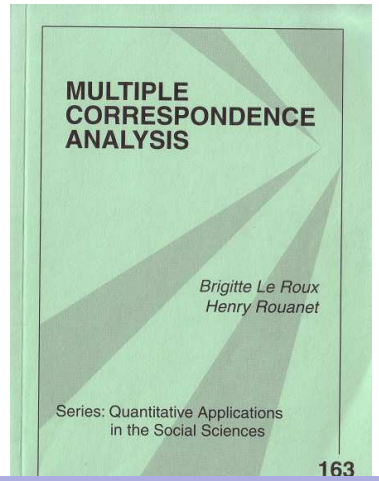
Murtagh 2005



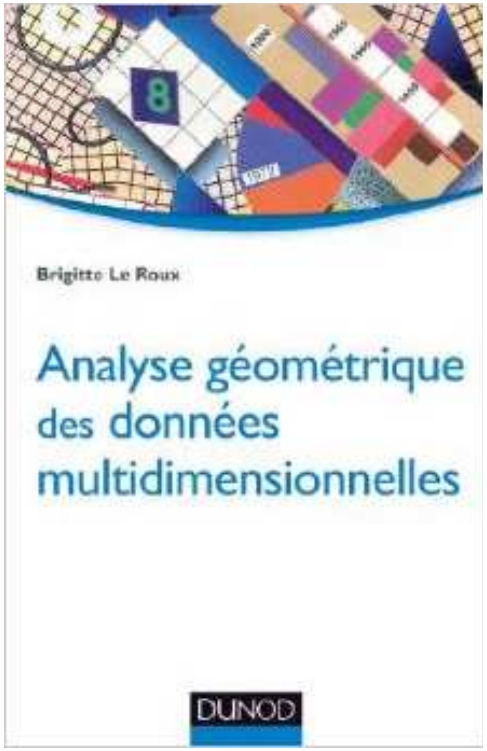
CARME 2003 (2006)



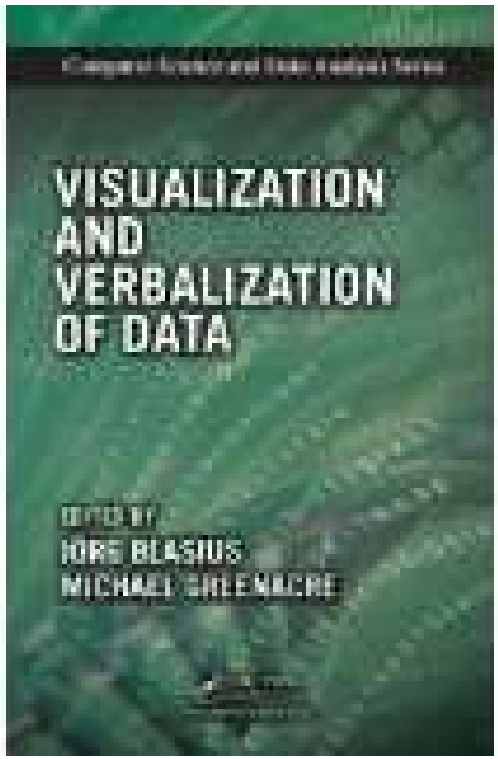
Le Roux & Rouanet 2010



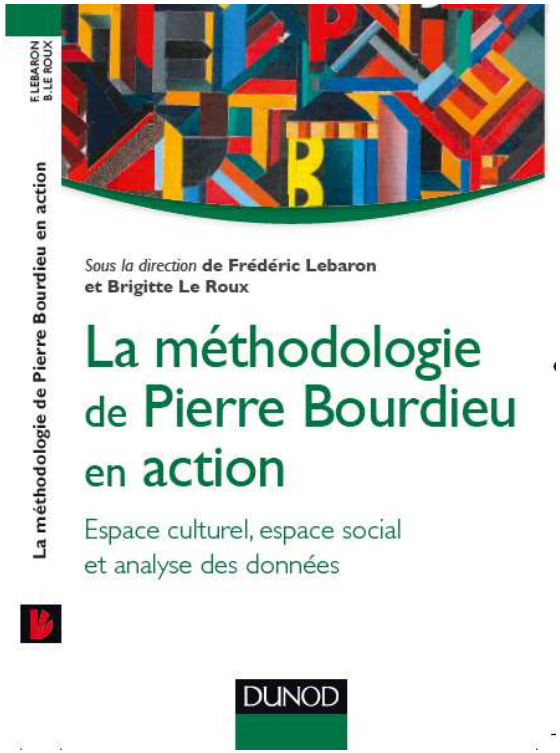
Le Roux
2014



CARME
2011 (2015)



Lebaron & Le Roux (eds)
2015



CA is now recognized and used, but GDA as a whole methodology, is waiting to be discovered by a large audience.

II — Principal Axes of a Euclidean Cloud

This text is adapted from Chapter 2 of the monograph

Multiple Correspondence Analysis

(QASS series n°163, SAGE, 2010)

II.1. Basic Geometric Notions

Elements of a geometric space: *points, line, plane*.

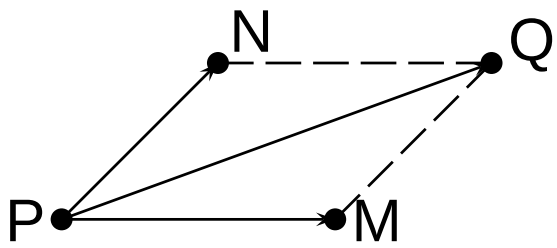
— *Affine notions*: alignment, direction and barycenter.

Couple of points (P, M), or *dipole* $\longrightarrow$ *vector* $\overrightarrow{\mathrm{PM}}$

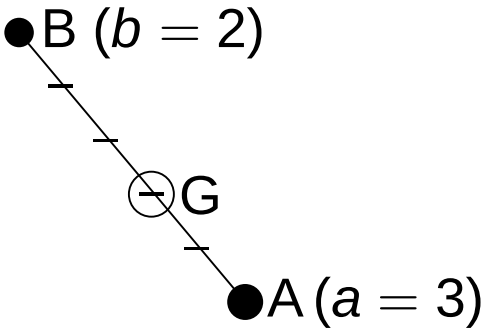
The *deviation* from point P to point M is M − P ("terminal minus initial"), that is, $\overrightarrow{\mathrm{PM}}$.

Deviations add up vectorially: sum of vectors by *parallelogram law*

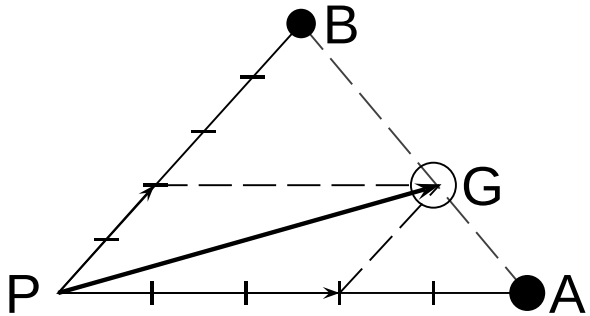
$$\overrightarrow{\mathrm{PM}} + \overrightarrow{\mathrm{PN}} = \overrightarrow{\mathrm{PQ}}$$



Barycenter of a dipole



$$G = \frac{3A+2B}{5}$$



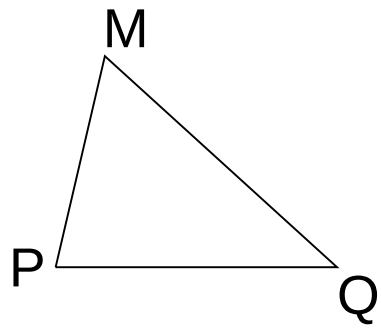
$$\overrightarrow{PG} = \frac{3}{5}\overrightarrow{PA} + \frac{2}{5}\overrightarrow{PB}$$

Barycenter = *weighted average of points*: $G = \frac{aA+bB}{a+b}$

— *Metric notions*: distances and angles.

Triangle inequality:

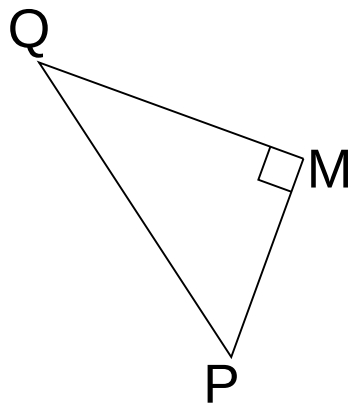
$$\mathrm{PQ} \leq \mathrm{PM} + \mathrm{MQ}$$



Pythagorean theorem:
If PM and MQ are perpendicular then:

$$(\mathrm{PM})^2 + (\mathrm{MQ})^2 = (\mathrm{PQ})^2$$

(triangle MPQ with right angle at M),



II.2. Cloud of Points

Figure 1. Target example (10 points)

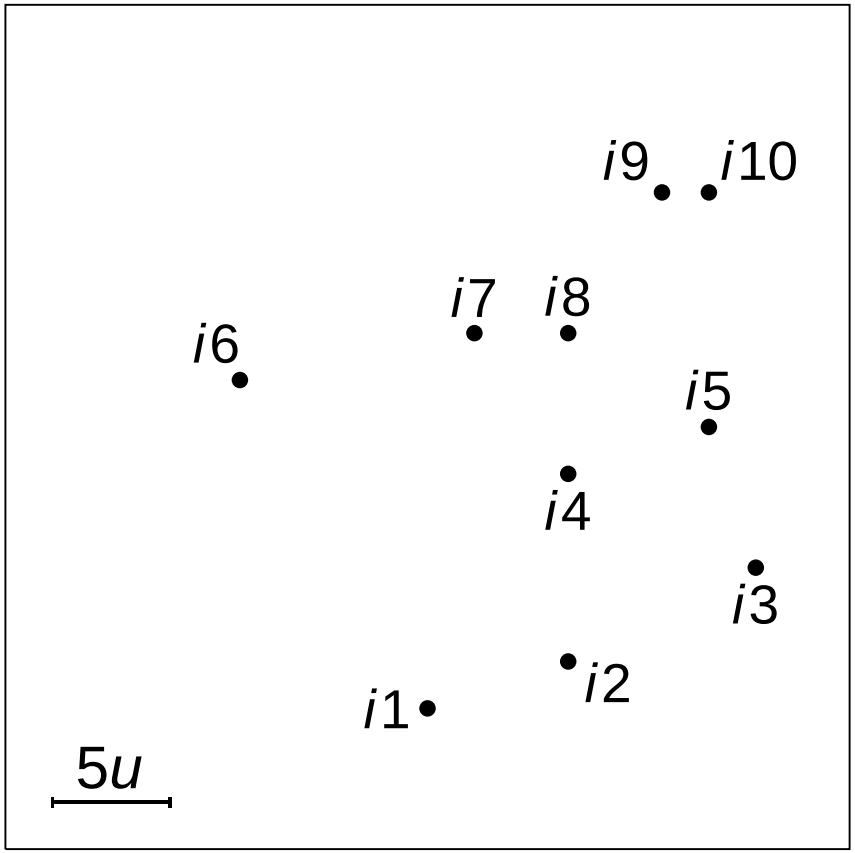
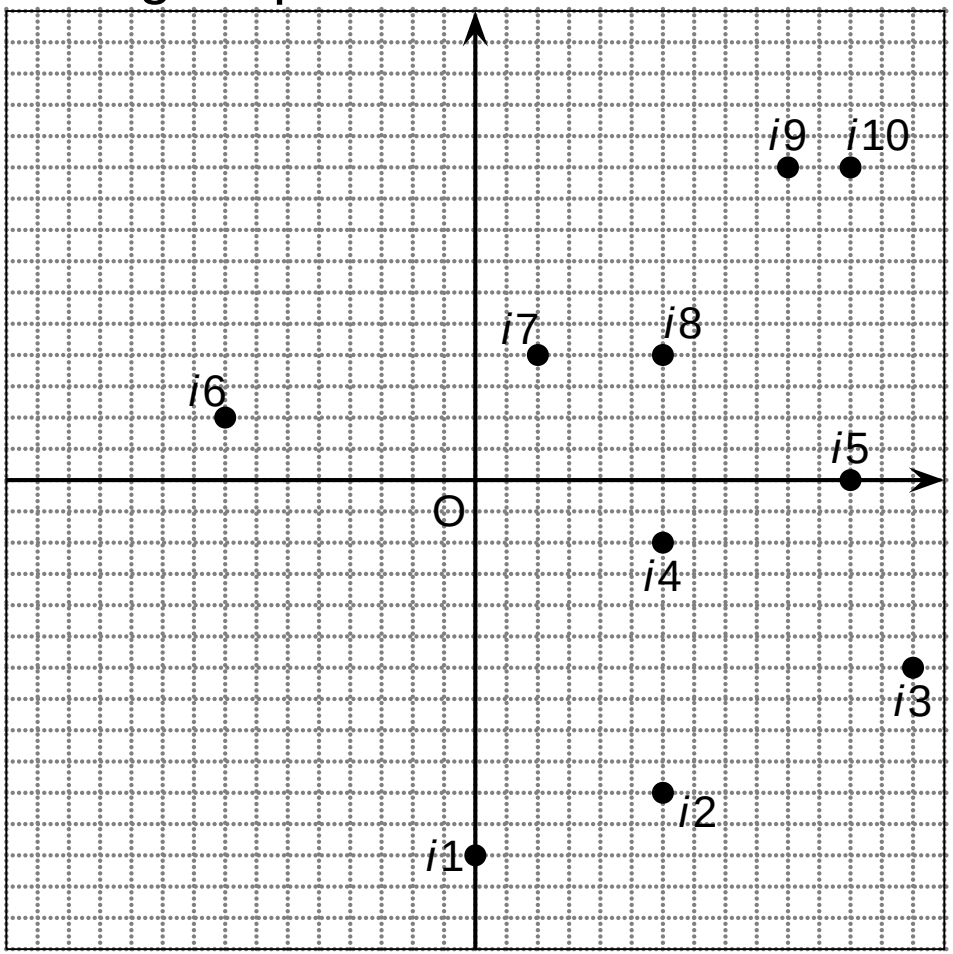


Figure 1b. Cloud of 10 points with origine-point O and initial axes



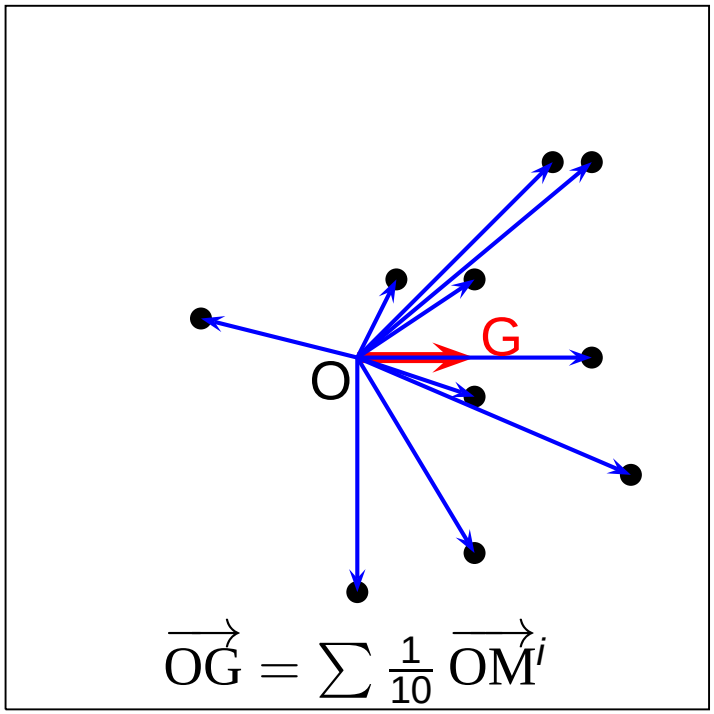
Initial coordinates

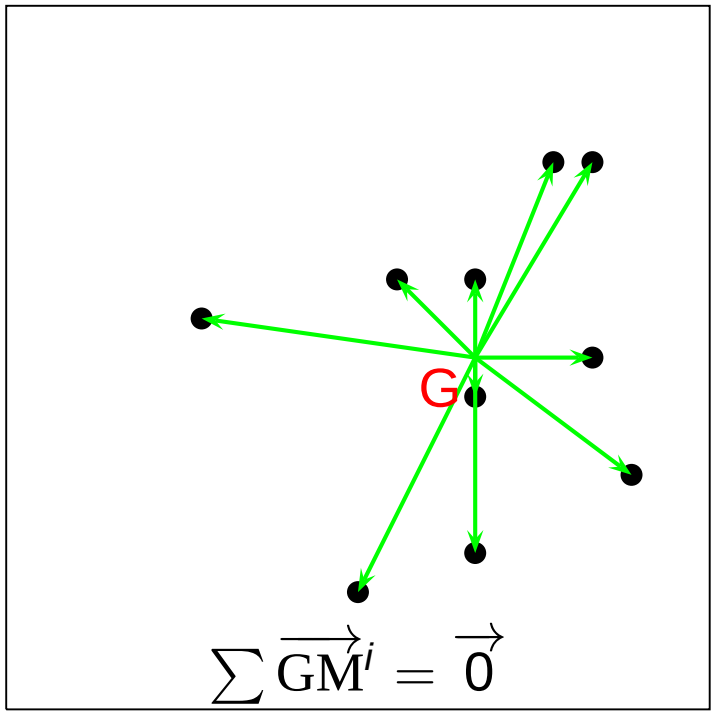
	x_1	x_2	weights
$i1$	0	−12	1
$i2$	6	−10	1
$i3$	14	−6	1
$i4$	6	−2	1
$i5$	12	0	1
$i6$	−8	2	1
$i7$	2	4	1
$i8$	6	4	1
$i9$	10	10	1
$i10$	12	10	1
Means	6	0	[10]
Variances	40	52	
Covariance		+ 8	

Mean point: point G

$$\overrightarrow{OG} = \sum p_i \overrightarrow{OM}^i \qquad \sum p_i \overrightarrow{GM}^i = \overrightarrow{0} \text{ (barycentric property)}$$

Target Example: $p_i = \frac{1}{n}$ $(p_i = \frac{1}{10})$





Variance of a cloud :

$$V_{\text{cloud}} = \sum p_i \, (\text{GM}^i)^2$$

(see Benzécri 1992, p.93)

Property

In rectangular axes, the variance of the cloud is the sum of the variances of the coordinate variables.

Contribution of point M^i :

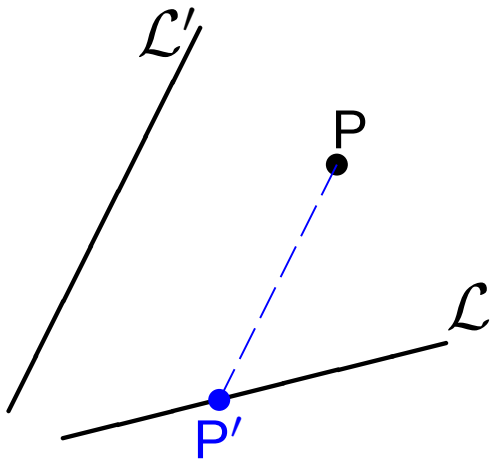
$$\text{Ctr}_i = \frac{p_i (\text{GM}^i)^2}{V_{\text{cloud}}}$$

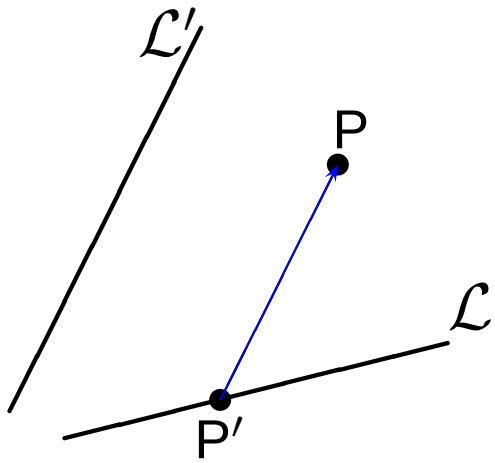
II.3. Principal Axes of a Cloud

Projection of a cloud

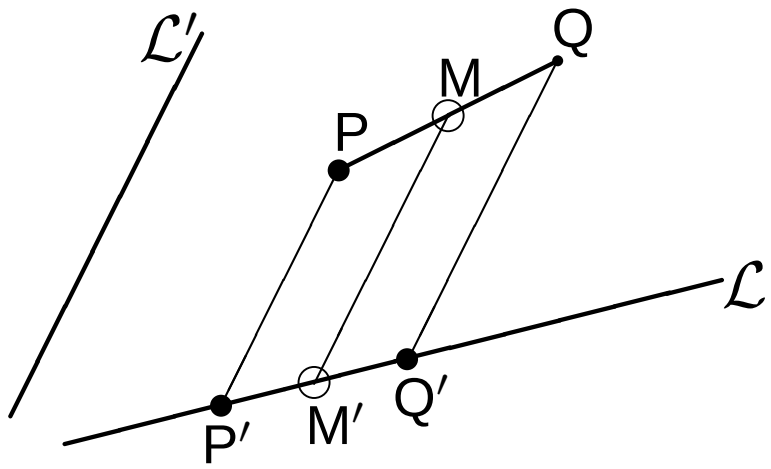
P′ = projection of point P onto $\mathcal{L}$ along $\mathcal{L}'$

$\overrightarrow{\mathrm{P'P}}$ = residual deviation





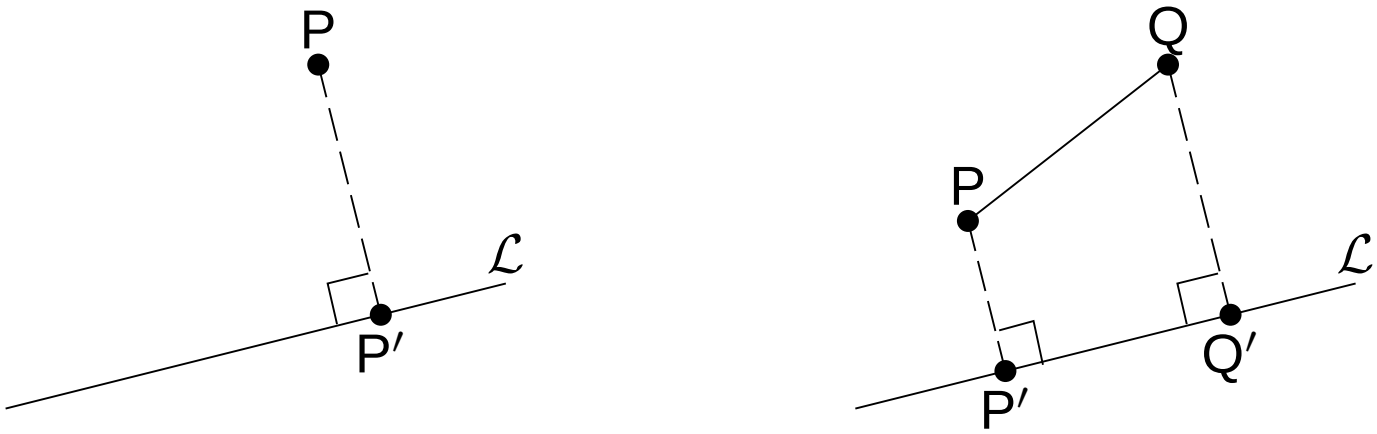
If point M is the midpoint of P and Q, the point M′, projection of M on $\mathcal{L}$, is the midpoint of P′ and Q′.



Mean point property

The mean point is preserved by projection.

Orthogonal projection: PP' is perpendicular to $\mathcal{L}$.

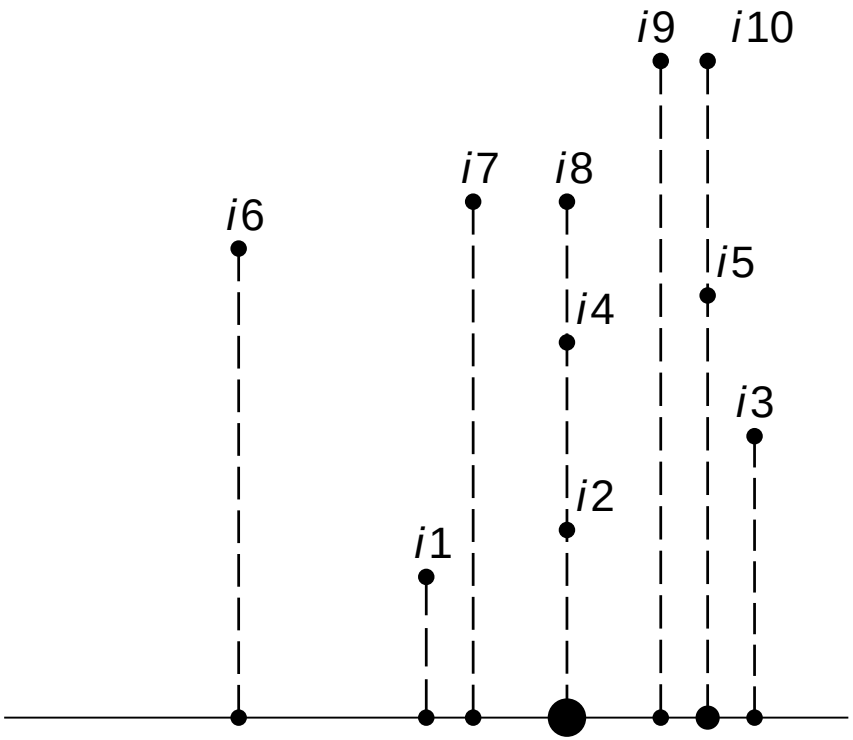


The orthogonal projection contracts distances: $\mathrm{P}'\mathrm{Q}' \leq \mathrm{PQ}$, therefore one has the

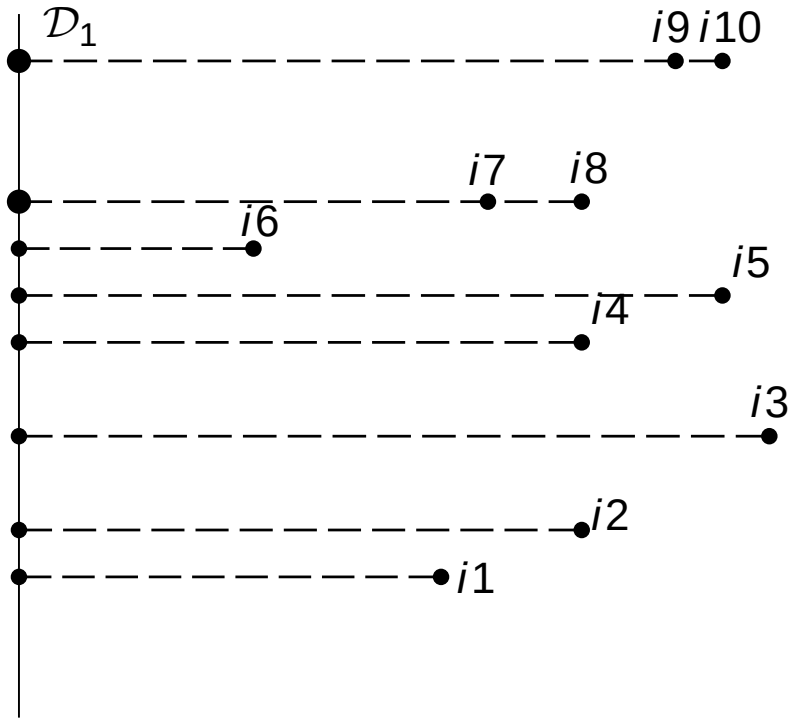
Property

variance of projected cloud $\leq$ variance of initial cloud.

Projected clouds on several lines



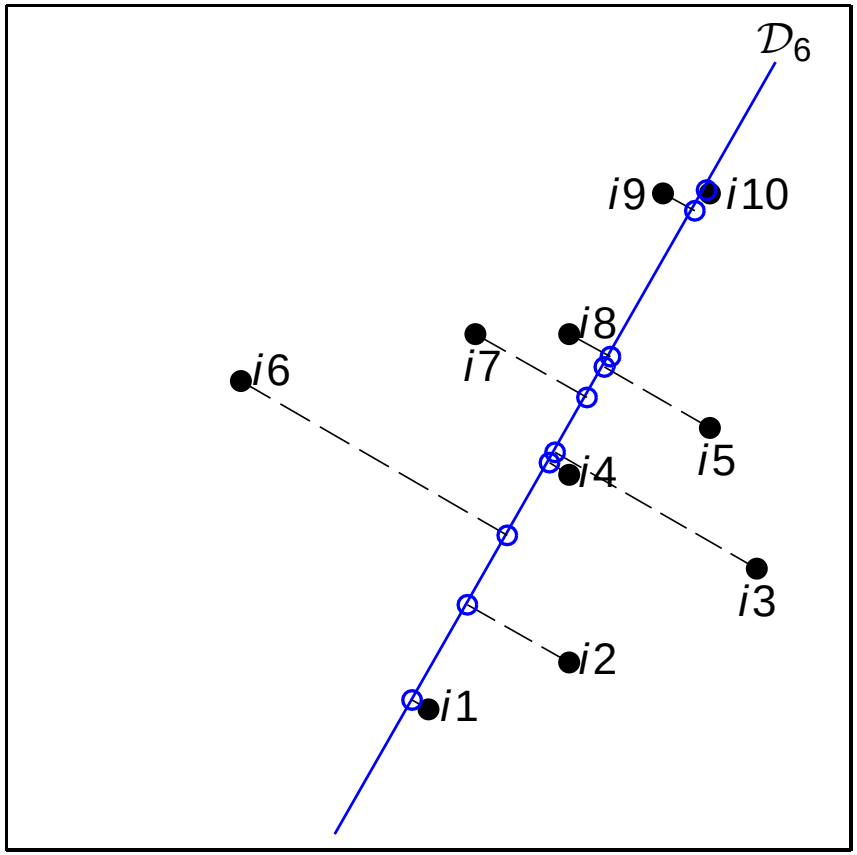
variance=40



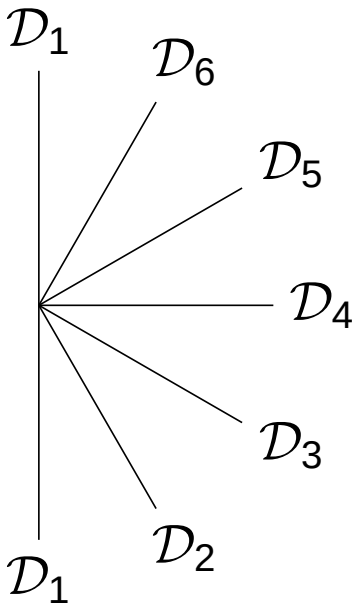
variance = 52

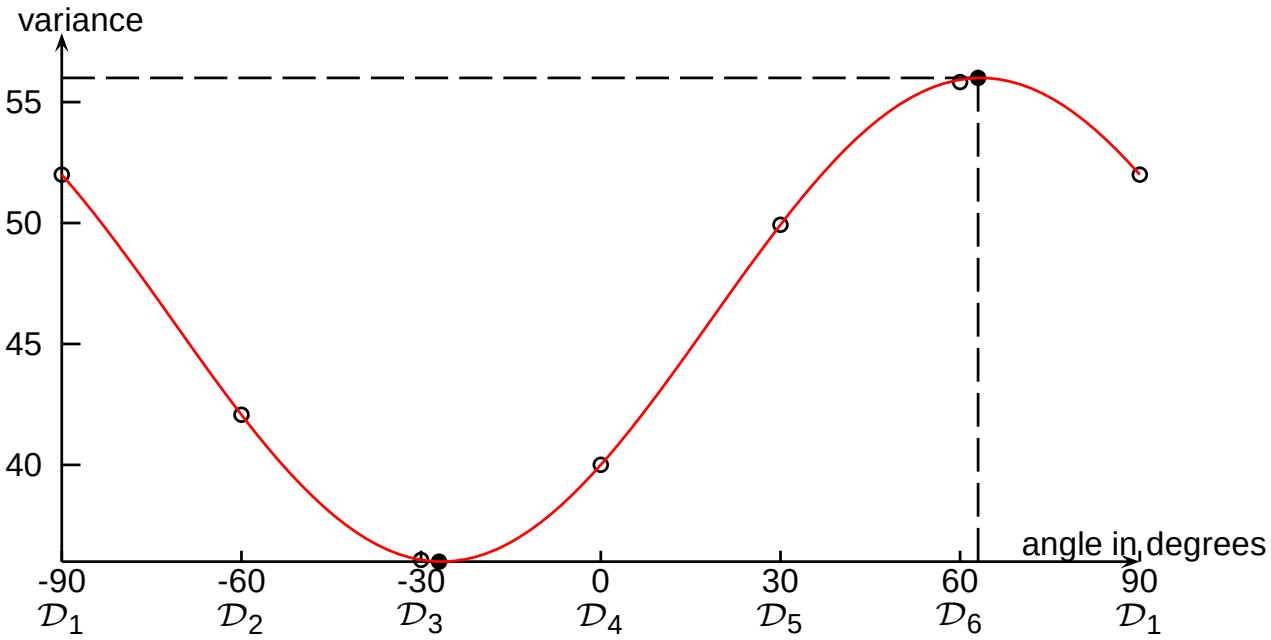
Orthogonal additive decomposition

The variance of the initial cloud is the sum of the variances of projected clouds onto perpendicular lines: $V_{\text{cloud}} = 40 + 52 = 92$.



Projection onto an oblique line (60 degrees) : variance = 55.9





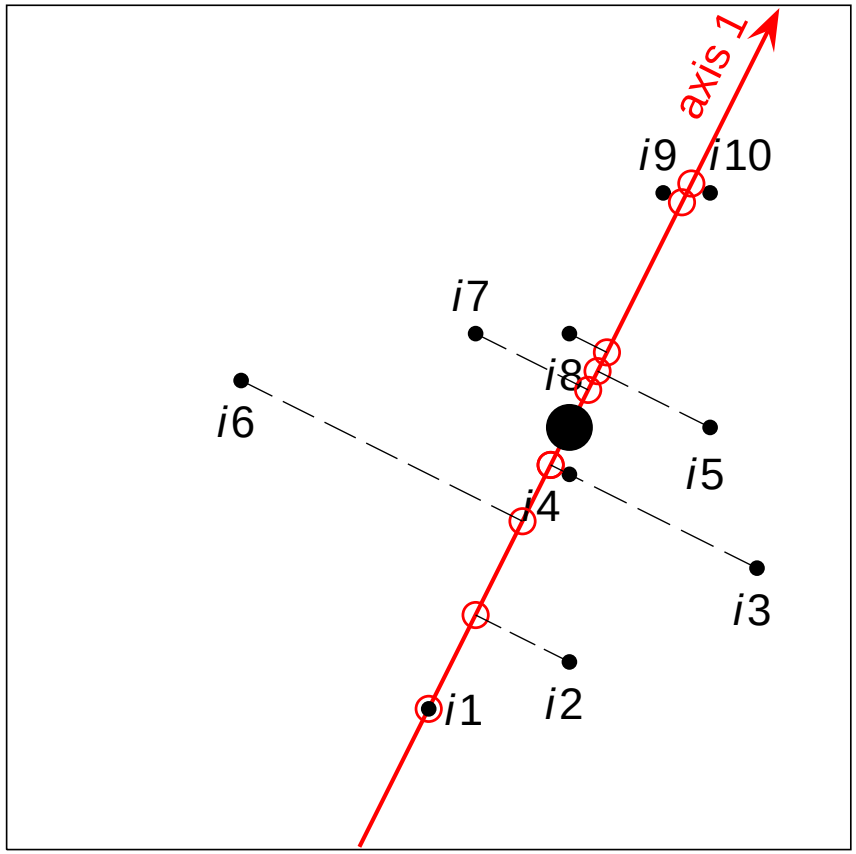
	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_1$
Variance	52	42.1	36.1	40.0	49.9	55.9	52

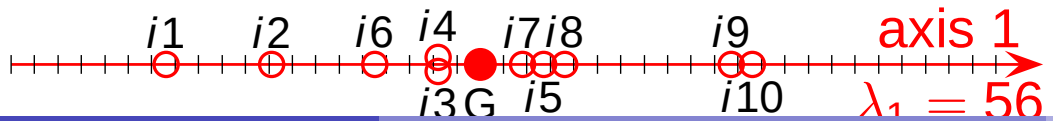
The line whose the variance of the projected cloud is maximum is called *first principal line*.
directed line $\rightarrow$ *1st principal axis*

Projected cloud = *1st principal cloud*
its variance (λ_1) = *variance of axis 1*

The first principal cloud is *the best fitting* of the initial cloud by an uni-dimensional cloud in the sense of *orthogonal least squares*

Here, $\alpha = 63°$, $\lambda_1 = 56$.

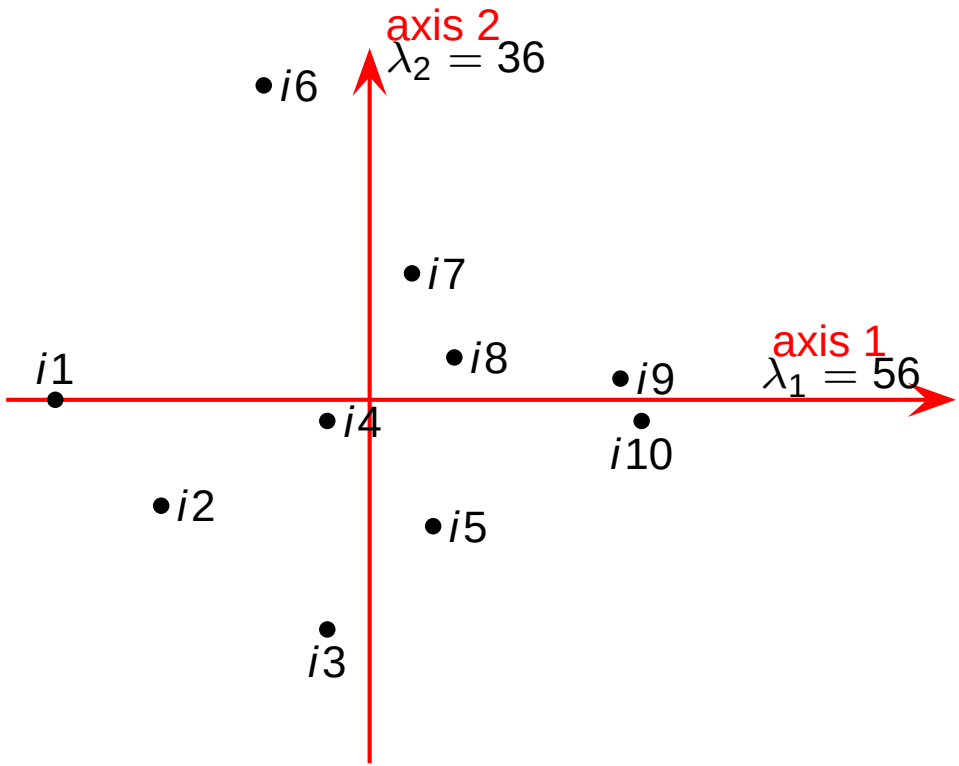




One constructs the residual cloud.

The first principal line of the residual cloud defines the *second principal line* of the initial cloud.

Here, the cloud is a plane cloud (two dimensions), hence the *second axis* is simply the perpendicular to the first axis.



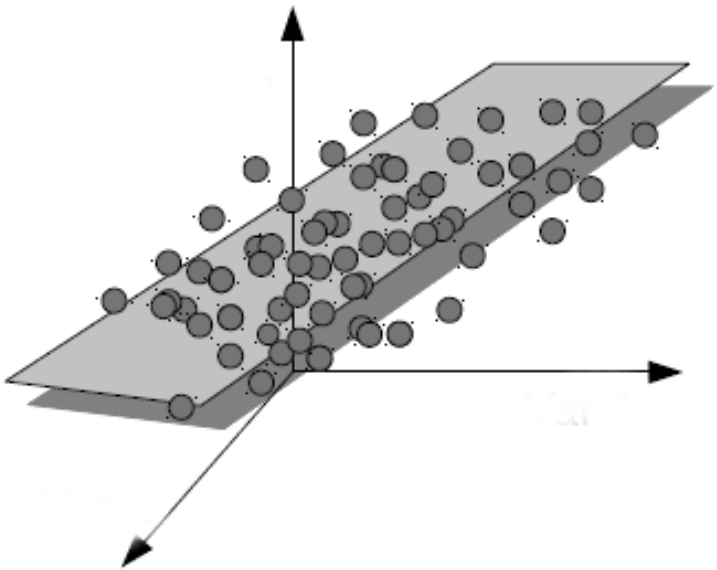
Principal representation of the cloud.

II.4. From a Plane Cloud to a Higher Dimensional Cloud

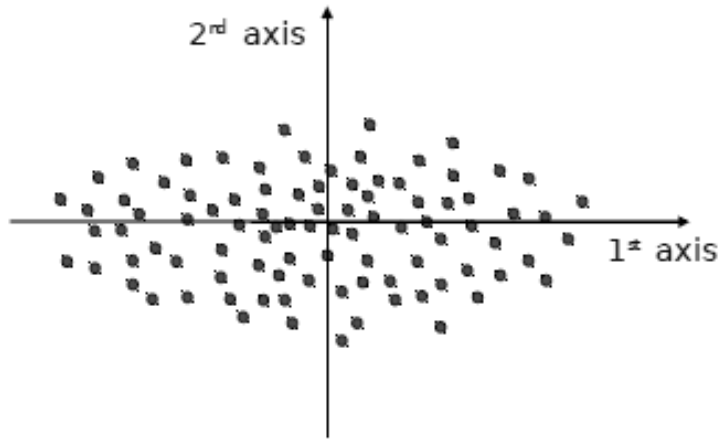
Heredity property

The plane that best fits the cloud is the one determined by the first two axes.

High dimensional cloud.



Low dimensional projection.



II.5. Properties

- Variance of cloud = sum of variances of axes: $V_{\text{cloud}} = \sum \lambda_\ell$.
- The principal axes are *pairwise orthogonal*.
  Each axis can be directed arbitrarily.
- The *principal coordinates* of points define principal variables.
  mean = 0 and variance = λ (eigenvalue)
  Principal variables are *uncorrelated* (for distinct eigenvalues).
- Reconstitution of distances between points:
  $d^2(i1, i2) = (-13.4 + 8.9)^2 + (0 - 4.47)^2 = 4.23 = (6.3)^2$

Aids to Interpretation

- Quality of fit of an axis or *variance rate*:

$$\frac{\lambda}{V_{\text{cloud}}}$$

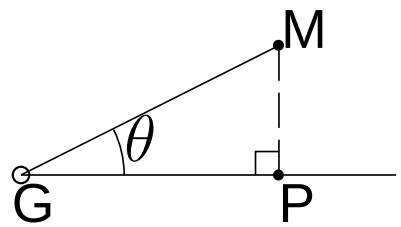
- *Contribution of point to axis*:

$$\text{Ctr} = \frac{p\,(y)^2}{\lambda} \qquad (p = \text{relative weight},\ y = \text{coordinate on axis})$$

- *Quality of representation of point onto axis*:

$$\cos^2\theta = \frac{\text{GP}^2}{\text{GM}^2}$$

Example: for $i2$, $\cos^2\theta = \frac{(-8.94)^2}{100} = 0.80$



Results of the analysis

$\lambda_1 = 56$ (variance of axis 1, eigenvalue).

Variance rate : $\dfrac{\lambda_1}{V_{\text{cloud}}} = \dfrac{56}{92} = 61\%$

		Results for axis 1			*Results for axis 2*		
		$\lambda_1 = 56$			$\lambda_1 = 36$		
	p_i	Coor-dinates	Ctr (%)	squared cosines	Coor-dinates	Ctr (%)	squared cosines
$i1$	0.1	-13.41	32.1	1.00	0.00	0	0.00
$i2$	0.1	-8.94	14.3	0.80	$+4.47$	5.6	0.20
$i3$	0.1	-1.79	0.6	0.03	$+9.84$	26.9	0.97
$i4$	0.1	-1.79	1.3	0.80	$+0.89$	0.2	0.20
$i5$	0.1	$+2.68$	3.6	0.20	$+5.37$	8	0.80
$i6$	0.1	-4.47	3.6	0.10	-13.42	50.0	0.90
$i7$	0.1	$+1.79$	0.6	0.10	-5.37	8	0.90
$i8$	0.1	$+3.58$	2.3	0.80	-1.79	0.9	0.20
$i9$	0.1	$+10.73$	20.6	0.99	-0.89	0.2	0.01
$i10$	0.1	$+11.63$	24.1	0.99	$+0.89$	0.2	0.01

III — Multiple Correspondence Analysis (MCA)

This text is adapted from Chapter 3 of the monograph

Multiple Correspondence Analysis

(QASS series n°163, SAGE, 2010)

III.1. Introduction

Language of questionnaire

Basic data set: Individuals$\times$Questions table

- Questions = categorical variables, i.e. variables with a finite number of *response categories*, or *modalities*.
- Individuals or “statistical individuals": (people, firms, items, etc.).

“*Standard format*”

for each question, each individual chooses *one and only one* response category.

$\rightarrow$ otherwise: preliminary phase of *coding*

III.2. Principles of MCA

Notations:

I: set of n individuals;

Q: set of questions

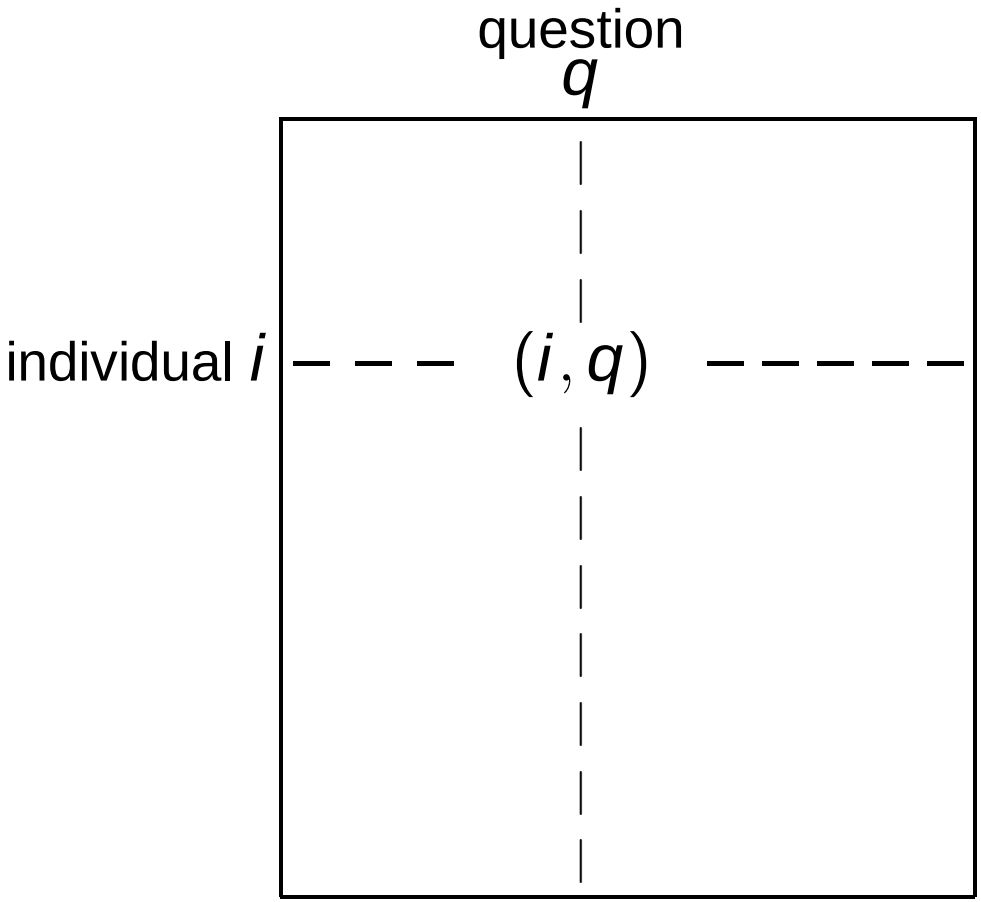
K_q: set of categories of question q ($K_q \geq 2$)

K: overall set of categories

n_k: number of individuals who have chosen category k (absolute frequency)

$f_k = \frac{n_k}{n}$ (relative frequency)

Table analyzed by MCA: $I \times Q$ table



MCA produces two clouds of points:
the *cloud of individuals* and the *cloud of categories*.

III.3. Taste example

- **Data**

$Q = 4$ active variables

Which, if any, of these different types of ... television programmes do you like the most?	n_k	f_k in %
News/Current affairs	220	18.1
Comedy/sitcoms	152	12.5
Police/detective	82	6.7
Nature/History documentaries	159	13.1
Sport	136	11.2
Film	117	9.6
Drama	134	11.0
Soap operas	215	17.7
Total	1215	100.0

Which, if any, of these different types of ... (cinema or television) films do you like the most?	n_k	f_k in %
Action/Adventure/Thriller	389	32.0
Comedy	235	19.3
Costume Drama/Literary adaptation	140	11.5
Documentary	100	8.2
Horror	62	5.1
Musical	87	7.2
Romance	101	8.3
SciFi	101	8.3
Total	1215	100.0

Which, if any, of these different types of ... art do you like the most?		n_k	f_k in %
Performance Art		105	8.6
Landscape		632	52.0
Renaissance Art		55	4.5
Still Life		71	5.8
Portrait		117	9.6
Modern Art		110	9.1
Impressionism		125	10.3
	Total	1215	100.0

Which, if any, of these different types of ... place to eat out would you like the best?	n_k	f_k in %
Fish & Chips/eat–in restaurant/cafe/teashop	107	8.8
Pub/Wine bar/Hotel	281	23.1
Chinese/Thai/**Indian Rest**aurant	402	33.1
Italian Restaurant/pizza house	228	18.8
French Restaurant	99	8.1
Traditional **Steakhouse**	98	8.1
Total	1215	100.0

$K = 8 + 8 + 7 + 6 = 29$ categories

$n = 1215$ individuals

$8 \times 8 \times 7 \times 6 = 2688$ possible response patterns, only 658 are observed.

Extract from the Individuals×Questions table

	TV	*Film*	*Art*	*Eat out*
1	Soap	Action	Landscape	SteakHouse
7	News	Action	Landscape	IndianRest
31	Soap	Romance	Portrait	Fish&Chips
235	News	Costume Drama	Renaissance	FrenchRest
679	Comedy	Horror	Modern	Indian
1215	Soap	Documentary	Landscape	SteakHouse

A row corresponds to the *response pattern* of an individual

III.4. Cloud of Individuals

Distance between 2 individuals due to question q:

— if q is an agreement question: i and i' choose the same category

$$d_q(i, i') = 0$$

— if q is a disagreement question: i chooses category k and i' chooses category k':

$$d_q^2(i, i') = \frac{1}{f_k} + \frac{1}{f_{k'}}$$

Overall distance: $d^2(i, i') = \frac{1}{Q} \sum_{q \in Q} d_q^2(i, i')$

individual $i \longrightarrow$ point M^i with relative weight $p_i = \frac{1}{n}$

G: mean point (center) of the cloud

$(GM^i)^2 = \left(\frac{1}{Q}\sum_{k \in K_i}\frac{1}{f_k}\right) - 1$ (K_i: response pattern of individual i).

Variance of the cloud of individuals

$$V_{\text{cloud}} = \frac{K}{Q} - 1$$

(average number of categories per question minus 1).

III.5. Cloud of Categories

Distance between categories k and k': $d^2(k,k') = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k\, n_{k'}/n}$

n_k = number of individuals who have chosen k (resp. $n_{k'}$);

$n_{kk'}$ = number of individuals who have chosen both categories k et k'.

category $k \longrightarrow$ category–point M^k with relative weight $p_k = f_k/Q$

Property

G is the mean point of the category–points of any question.

$(\mathrm{GM}^k)^2 = \frac{1}{f_k} - 1.$

- *Variance of the cloud of categories*: $= \frac{K}{Q} - 1.$
- *Contributions*

Contribution of *category k*	Contribution of *question q*
$\mathrm{Ctr}_k = \frac{1-f_k}{K-Q}$	$\mathrm{Ctr}_q = \frac{K_q - 1}{K-Q}$

III.6. Principal Clouds

— *Principal axes*

Fundamental properties

- The two clouds have the same variances (eigenvalues).
- $\sum_{\ell=1}^{L} \lambda_\ell = V_{\text{cloud}}$, with $\overline{\lambda} = \dfrac{V_{\text{cloud}}}{L} = \dfrac{1}{Q}$.

— *Variance rates and modified rates*

Variance rate:

$$\tau_\ell = \frac{\lambda_\ell}{V_{\text{cloud}}}$$

Modified rate:

$$\tau'_\ell = \frac{\lambda'_\ell}{S}, \text{ with } \lambda'_\ell = \left(\frac{Q}{Q-1}\right)^2 (\lambda_\ell - \overline{\lambda})^2 \text{ and } S = \sum_{\ell=1}^{\ell_{\max}} \lambda'_\ell$$

— *Principal coordinates and principal variables*

y_ℓ^i: coordinate of individual i on axis ℓ

$y_\ell^I = (y_\ell^i)_{i \in I}$: ℓ-th principal variable over I

y_ℓ^k: coordinate of category k on axis ℓ

$y_\ell^K = (y_\ell^k)_{k \in K}$: ℓ-th principal variable over K

Properties

Mean of principal variable is null:

$$\sum \tfrac{1}{n} y_\ell^i = 0 \text{ and } \sum p_k y_\ell^k = 0$$

Variance of principal variable ℓ is equal to λ_ℓ:

$$\sum \tfrac{1}{n} (y_\ell^i)^2 = \lambda_\ell \text{ and } \sum p_k (y_\ell^k)^2 = \lambda_\ell$$

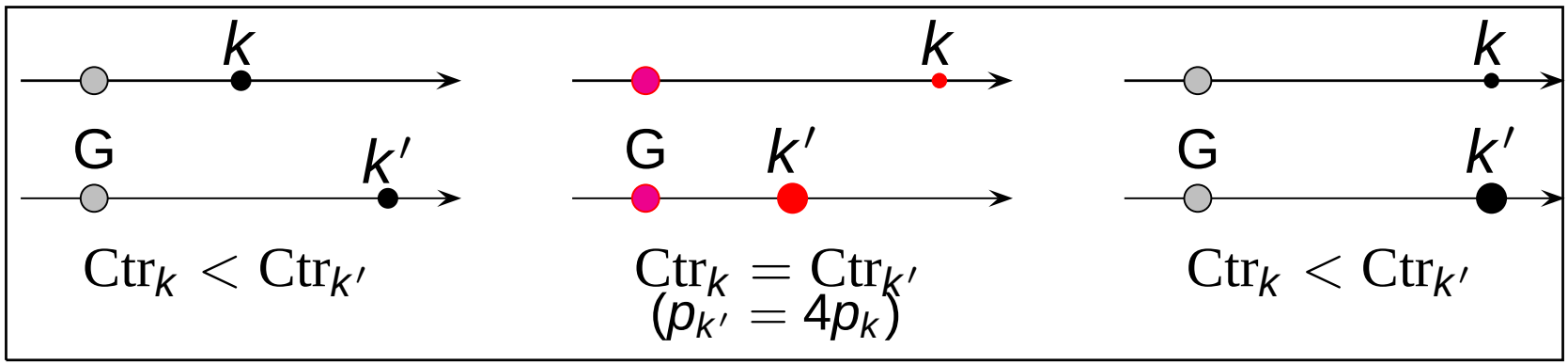
Principal variables are pairwise uncorrelated:

$$\ell \neq \ell' \quad \sum y_\ell^i y_{ell'}^i = 0 \quad \sum y_\ell^k y_{ell'}^k = 0$$

III.7. Aids to Interpretation: Contributions

Contribution of category–point k to axis ℓ: $\dfrac{p_k\,(y_\ell^k)^2}{\lambda\ell}$

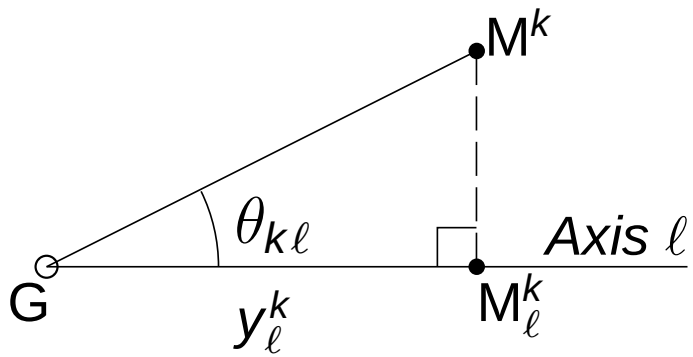
(y: coordinate of point on axis; p: relative weight; λ: variance of axis)



By grouping, contributions add up $\longrightarrow$ contribution of question...

The quality of representation of point M^k on Axis ℓ is

$$\cos^2 \theta_{k\ell} = \frac{(GM_\ell^k)^2}{(GM^k)^2} = \frac{(y_\ell^k)^2}{(GM^k)^2}$$



— Category mean points

$\overline{\mathrm{M}}^k$: category mean point for k with coordinate on axis ℓ

$$\overline{y}_\ell^k = \sqrt{\lambda_\ell}\, y_\ell^k \qquad \textit{(second transition formula)}$$

The K category mean points of question q define the

between–q cloud

.

— Supplementary elements:

individuals and/or questions

III.8. MCA of the Taste Example

Data set

The data involve:

- $Q = 4$ active variables
- $K = 8 + 8 + 7 + 6 = 29$ categories
- $n = 1215$ individuals

Overall variance of the cloud : $V_{\text{cloud}} = \frac{29}{4} - 1 = 6.25$

Contributions of questions to the overall variance:

$$\frac{8-1}{29-4} = 28\% \qquad 28\% \qquad 24\% \qquad 20\%$$

Elementary statistical results

$8 \times 8 \times 7 \times 6 = 2688$ possible response patterns; 658 are observed.

TV	n_k	f_k	Ctr_k
News	220	18.1	3.3
Comedy	152	12.5	3.5
Police	82	6.7	3.7
Nature	159	13.1	3.5
Sport	136	11.2	3.6
Film	117	9.6	3.6
Drama	134	11.0	3.6
Soap operas	215	17.7	3.3
Films	1215	100.0	28.0
Action	389	32.0	2.7
Comedy	235	19.3	3.2
Costume Drama	140	11.5	3.5
Documentary	100	8.2	3.7
Horror	62	5.1	3.8
Musical	87	7.2	3.7
Romance	101	8.3	3.7
SciFi	101	8.3	3.7
Total	1215	100.0	28.0

Art	n_k	f_k	Ctr_k
Performance	105	8.6	3.7
Landscape	632	52.0	1.9
Renaissance	55	4.5	3.8
Still Life	71	5.8	3.8
Portrait	117	9.6	3.6
Modern Art	110	9.1	3.6
Impressionism	125	10.3	3.6
Eat out	1215	100.0	24.0
Fish & Chips	107	8.8	3.6
Pub	281	23.1	3.1
Indian Rest	402	33.1	2.7
Italian Rest	228	18.8	3.2
French Rest	99	8.1	3.7
Steakhouse	98	8.1	3.7
Total	1215	100.0	20.0

Basic results of MCA

Dimensionality of the cloud $\leq K - Q = 29 - 4 = 25$.
Mean of the variances of axes: $\frac{6.25}{25} = 0.25$.

The variances of 12 axes exceed the mean.

Axes ℓ	1	2	3	4	5	6	7	8	9	10	11	12
variances (λ_ℓ)	.400	.351	.325	.308	.299	.288	.278	.274	.268	.260	.258	.251
variance rates	.064	.056	.052	.049	.048	.046	.045	.044	.043	.042	0.41	.040
modified rates	.476	.215	.118	.071	.050	.030	.017	.012	.007	.002	.001	.000

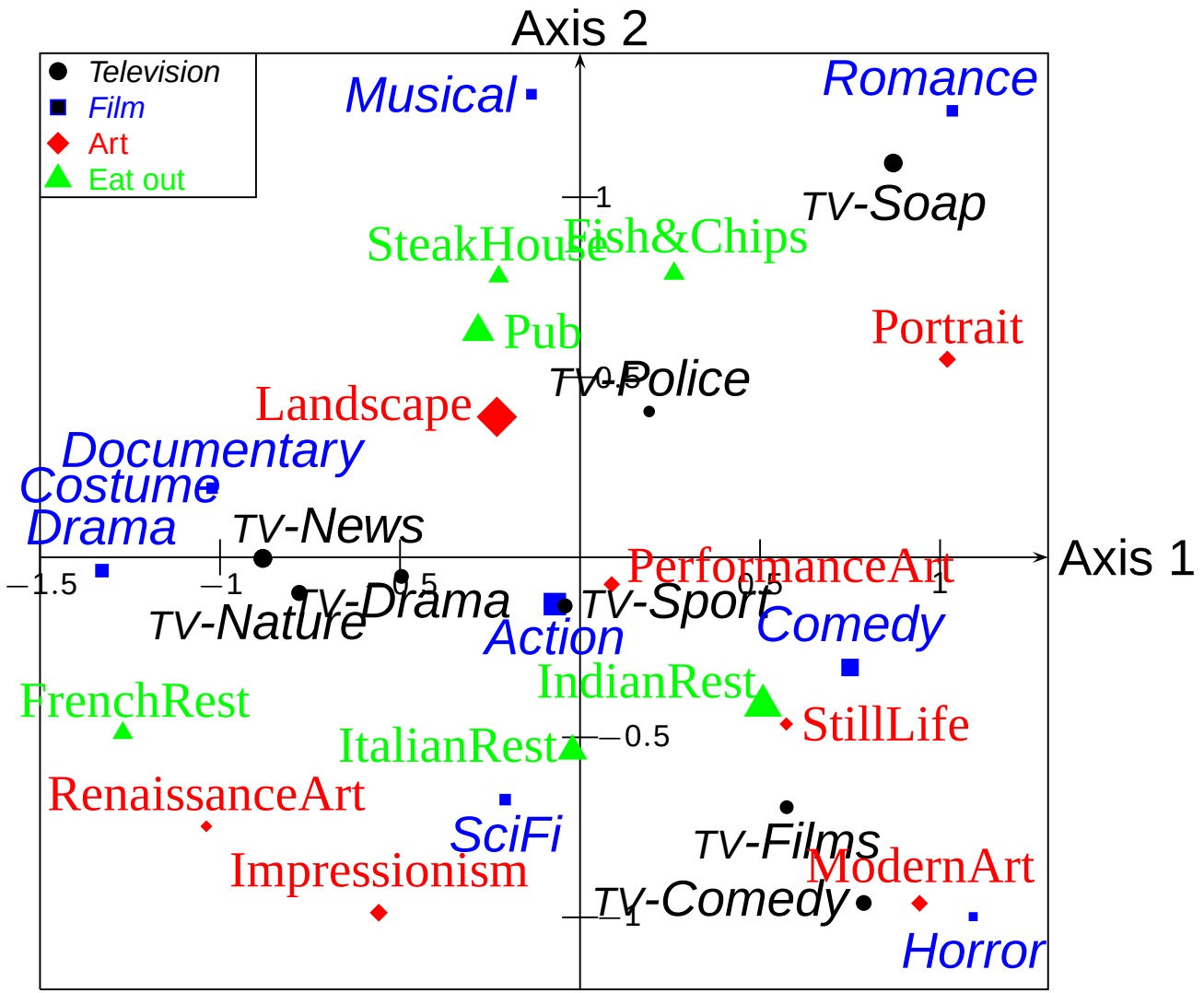
Principal coordinates and contributions of 6 individuals (in %)

	Coordinates			Contributions (in %)		
	Axis 1	Axis 2	Axis 3	Axis 1	Axis 2	Axis 3
1	+0.135	+0.902	+0.432	0.00	0.19	0.05
7	−0.266	−0.064	−0.438	0.01	0.00	0.05
31	+1.258	+1.549	−0.768	0.33	0.56	0.15
235	−1.785	−0.538	−1.158	0.65	0.07	0.34
679	+1.316	−1.405	−0.140	0.36	0.46	0.00
1215	−0.241	+1.037	+0.374	0.01	0.25	0.04

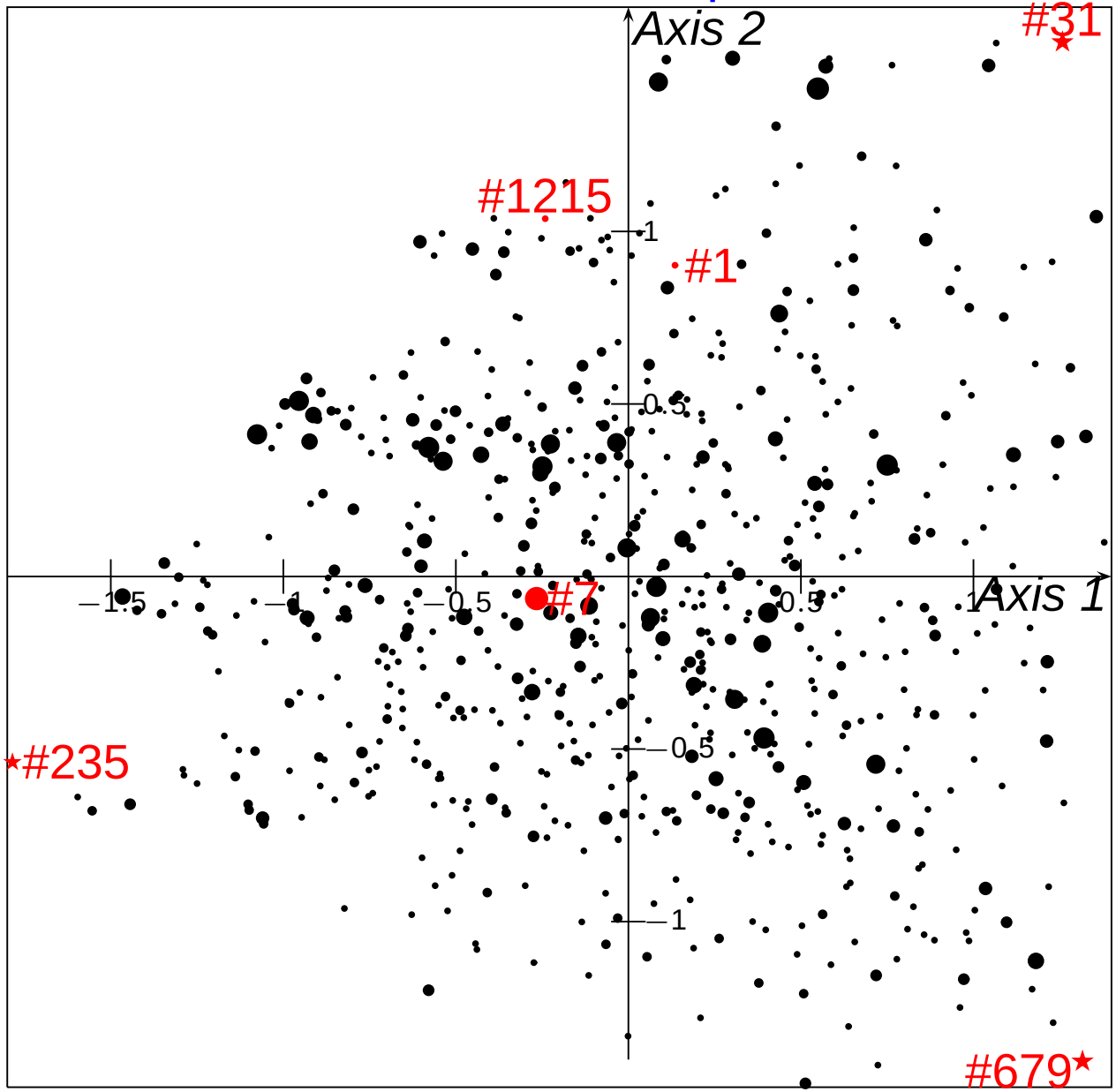
Relative weight, principal coordinates and contributions (in %) of categories

Television	p_k	Axe 1	Axe 2	Axe 3	Axe1	Axe 2	Axe 3
TV-News	.0453	−0.881	−0.003	−0.087	**8.8**	0.0	0.1
TV-Comedy	.0313	+0.788	−0.960	−0.255	**4.9**	**8.2**	0.6
TV-Police	.0169	+0.192	+0.405	+0.406	0.2	0.8	0.9
TV-Nature	.0327	−0.775	−0.099	+0.234	**4.9**	0.1	0.6
TV-Sport	.0280	−0.045	−0.133	+1.469	0.0	0.1	**18.6**
TV-Film	.0241	+0.574	−0.694	+0.606	2.0	**3.3**	2.7
TV-Drama	.0276	−0.496	−0.053	−0.981	1.7	0.0	**8.2**
TV-Soap	.0442	+0.870	+1.095	−0.707	**8.4**	**15.1**	**6.8**
Film				*Total*	*30.7*	*27.7*	*38.4*
Action	.0800	−0.070	−0.127	+0.654	0.1	0.4	**10.5**
Comedy	.0484	+0.750	−0.306	−0.307	**6.8**	1.3	1.4
CostumeDrama	.0288	−1.328	−0.037	−1.240	**12.7**	0.0	**13.6**
Documentary	.0206	−1.022	+0.192	+0.522	**5.4**	0.2	1.7
Horror	.0128	+1.092	−0.998	+0.103	**3.8**	**3.6**	0.0
Musical	.0179	−0.135	+1.286	−0.109	0.1	**8.4**	0.1
Romance	.0208	+1.034	+1.240	−1.215	**5.5**	**9.1**	**9.4**
SciFi	.0208	−0.208	−0.673	+0.646	0.2	**2.7**	2.7
Art				*Total*	*34.6*	*25.7*	*39.5*
PerformanceArt	.0216	+0.088	−0.075	−0.068	0.0	0.0	0.0
Landscape	.1300	−0.231	+0.390	+0.313	1.7	**5.6**	**3.9**
RenaissanceArt	.0113	−1.038	−0.747	−0.566	**3.0**	1.8	1.1
StillLife	.0146	+0.573	−0.463	−0.117	1.2	0.9	0.1
Portrait	.0241	+1.020	+0.550	−0.142	**6.3**	2.1	0.1
ModernArt	.0226	+0.943	−0.961	−0.285	**5.0**	**5.9**	0.6
Impressionism	.0257	−0.559	−0.987	−0.824	2.0	**7.1**	**5.4**
Eat out				*Total*	*19.3*	*23.5*	*11.2*
Fish&Chips	.0220	+0.261	+0.788	+0.313	0.4	**3.9**	0.7
Pub	.0578	−0.283	+0.627	+0.087	1.2	**6.5**	0.1
IndianRest	.0827	+0.508	−0.412	+0.119	**5.3**	**4.0**	0.4
ItalianRest	.0469	−0.021	−0.538	−0.452	0.0	**3.9**	**2.9**
FrenchRest	.0204	−1.270	−0.488	−0.748	**8.2**	1.4	**3.5**
Steakhouse	.0202	−0.226	+0.780	+0.726	0.3	**3.5**	**3.3**
				Total	*15.3*	*23.1*	*10.9*

Cloud of categories in plane 1-2



Cloud of individuals in plane 1-2.

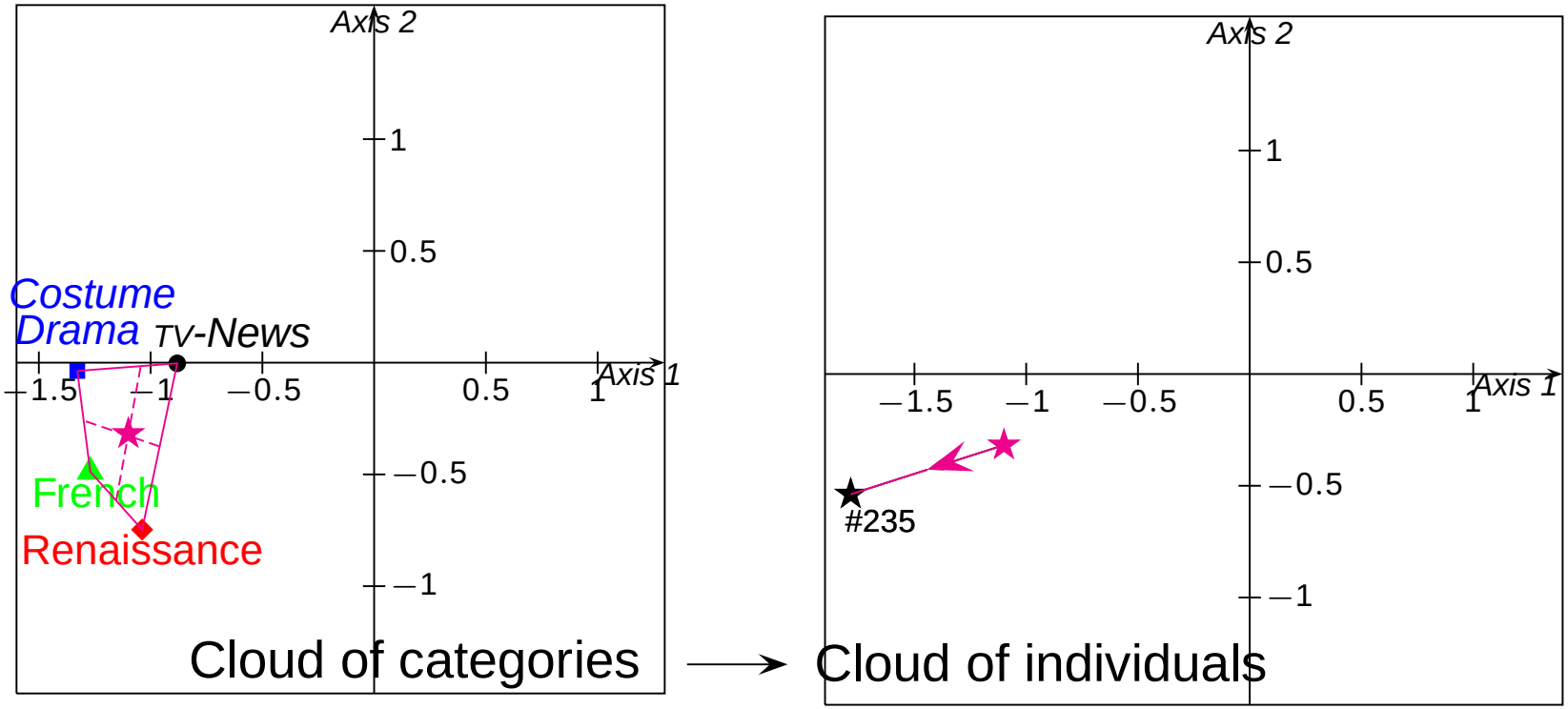


III.9. Transition Formulas

Transition formulas express the *relation* between
the *cloud of individuals*
and
the *cloud of categories*.

• First transition formula

cloud of categories $\longrightarrow$ cloud of individuals: $y^i = \frac{1}{\sqrt{\lambda}} \sum_{k \in K_i} y^k / Q$



Category–point k is located at the equibarycenter of the n_k individuals who have chosen category k, up to a stretching along principal axes.

In terms of coordinates:

1. mean of the 4 coordinates on axis 1:

$$\frac{-0.881 - 1.328 - 1.038 - 1.270}{4} = -1.12925$$

mean of the 4 coordinates on axis 2:

$$\frac{-0.003 - 0.037 - 0.747 - 0.488}{4} = -0.31875$$

2. dividing the coordinate on axis 1 by $\sqrt{\lambda_1}$:

$$y_1^i = \frac{-1.12925}{\sqrt{0.4004}} = -1.785$$

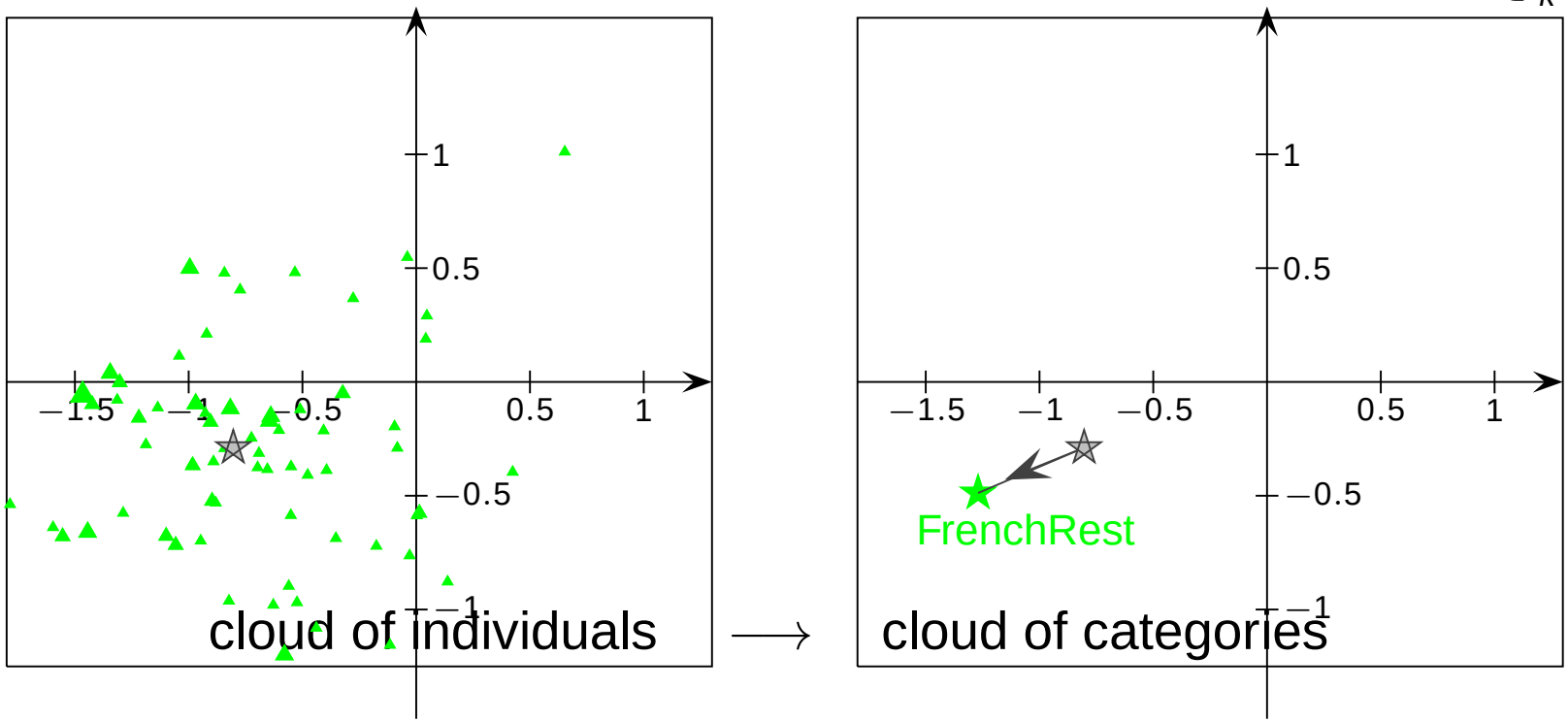
dividing the coordinate on axis 2 by $\sqrt{\lambda_2}$

$$y_2^i = \frac{-0.31875}{\sqrt{0.3512}} = -0.538$$

which are the coordinates of the *individual–point* #235 .

• Second transition formula

cloud of individuals $\longrightarrow$ cloud of categories: $y^k = \frac{1}{\sqrt{\lambda}} \sum_{i \in I_k} y^i / n_k$



Individual–point is located at the equibarycenter of the Q category–points of his response pattern, up to a stretching along principal axes.

III.10.Interpretation of the Analysis of the Taste Example

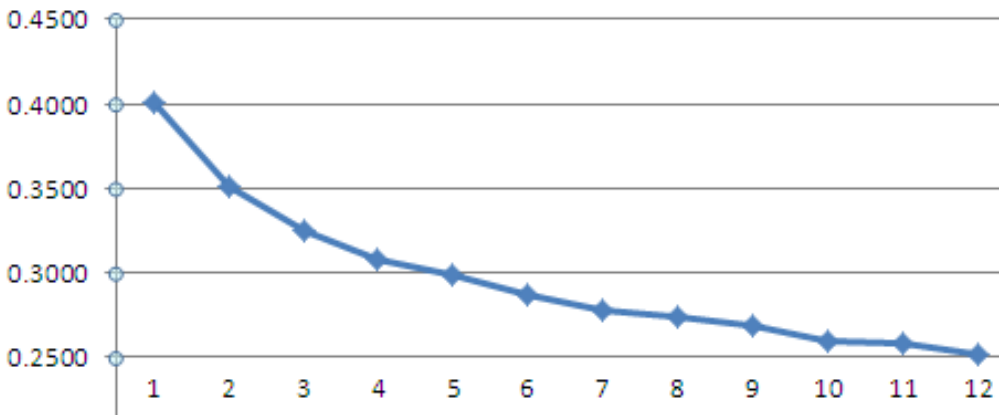
How many axes need to be interpreted?

Axis 1: ($\frac{\lambda_1-\lambda_2}{\lambda_1} = .12$); modified rate = 0.48

Axis 2: ($\frac{\lambda_2-\lambda_3}{\lambda_2} = .07$); modified rate = 0.22.

Cumulated modified rate for axes 1 and 2 = 0.70.

After axis 4, variances decrease regularly and the differences are small.



1	0.4004	6.41	0.48
2	0.3512	5.62	0.22
3	0.3250	5.20	0.12
4	0.3081	4.93	0.07
5	0.2989	4.78	0.05
6	0.2876	4.60	0.03

Cumulated modified rate for for axes 1, 2 and 3 = 82%

Guide for interpreting an axis

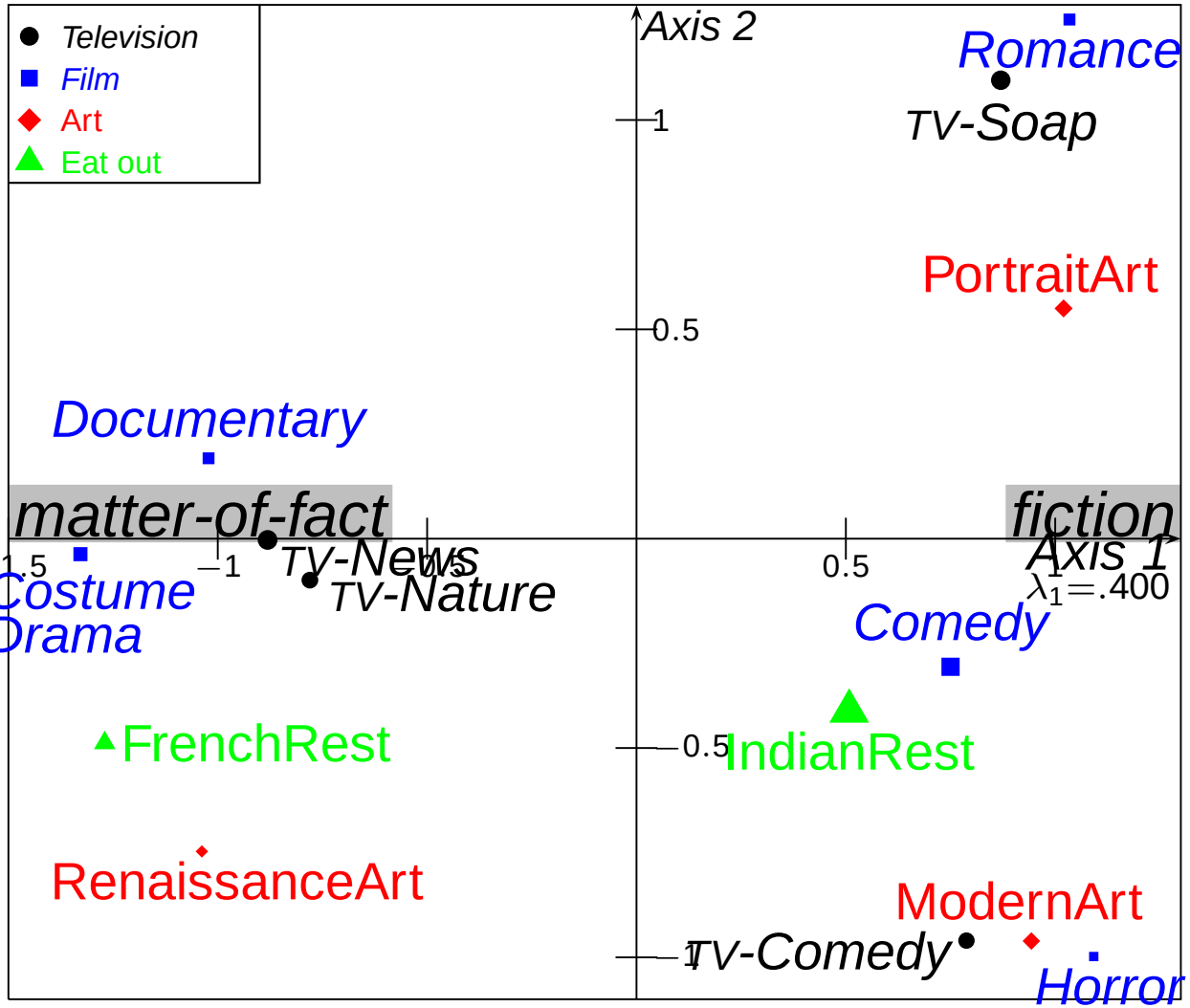
Interpreting an axis amounts to finding out what is similar, on the one hand, between all the elements figuring on the right of the origin and, on the other hand between all that is written on the left; and expressing with conciseness and precision, the contrast (or opposition) between the two extremes.

Benzécri (1992, p. 405)

For interpreting an axis, we use the method of contributions of points and deviations.
Baseline criterion = average contribution = $100/29 \rightarrow 3.4\%$

The interpretation of an axis is based on the categories whose contributions to axis exceed the average contribution.

Interpretation of axis 1

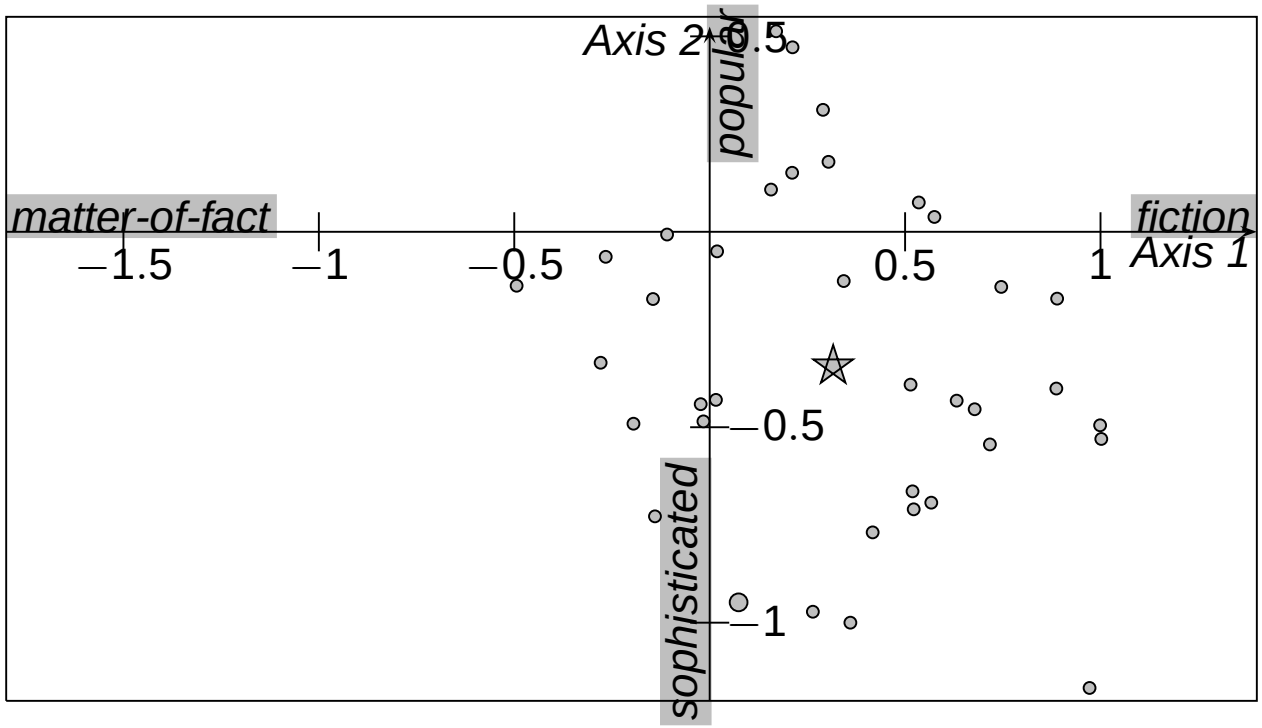


● *TV* (31%)	*left*	*right*
TV-News	8.8	
TV-Soap		8.4
TV-Nature	4.9	
TV-Comedy		4.9
■ *Film* (35%)		
Cost. Drama	12.7	
Comedy		6.8
Romance		5.5
Documentary	5.4	
Horror		3.8
♦ *Art* (19%)		
Portrait	6.3	
Modern		5.0
Renaissance	3.0	
▲*Eat out* (15%)		
French Rest.	8.2	
Indian Rest.		5.3
Total: 43.0 + 46.0 = 89.0		

14 categories selected for the interpretation of axis 1: sum of contributions = 89% $\rightarrow$ *good summary*

- Axis 1 opposes *matter–of–fact* (and traditional) tastes to *fiction world* (and modern) tastes.
- Axis 2 opposes *popular* to *sophisticated* tastes.
- Axis 3 opposes *outward dispositions* to *inward ones*.

Supplementary individuals



Plane 1-2. Cloud of 38 Indian immigrants with its mean point (⋆).

LOCATE YOURSELF

Supplementary variables

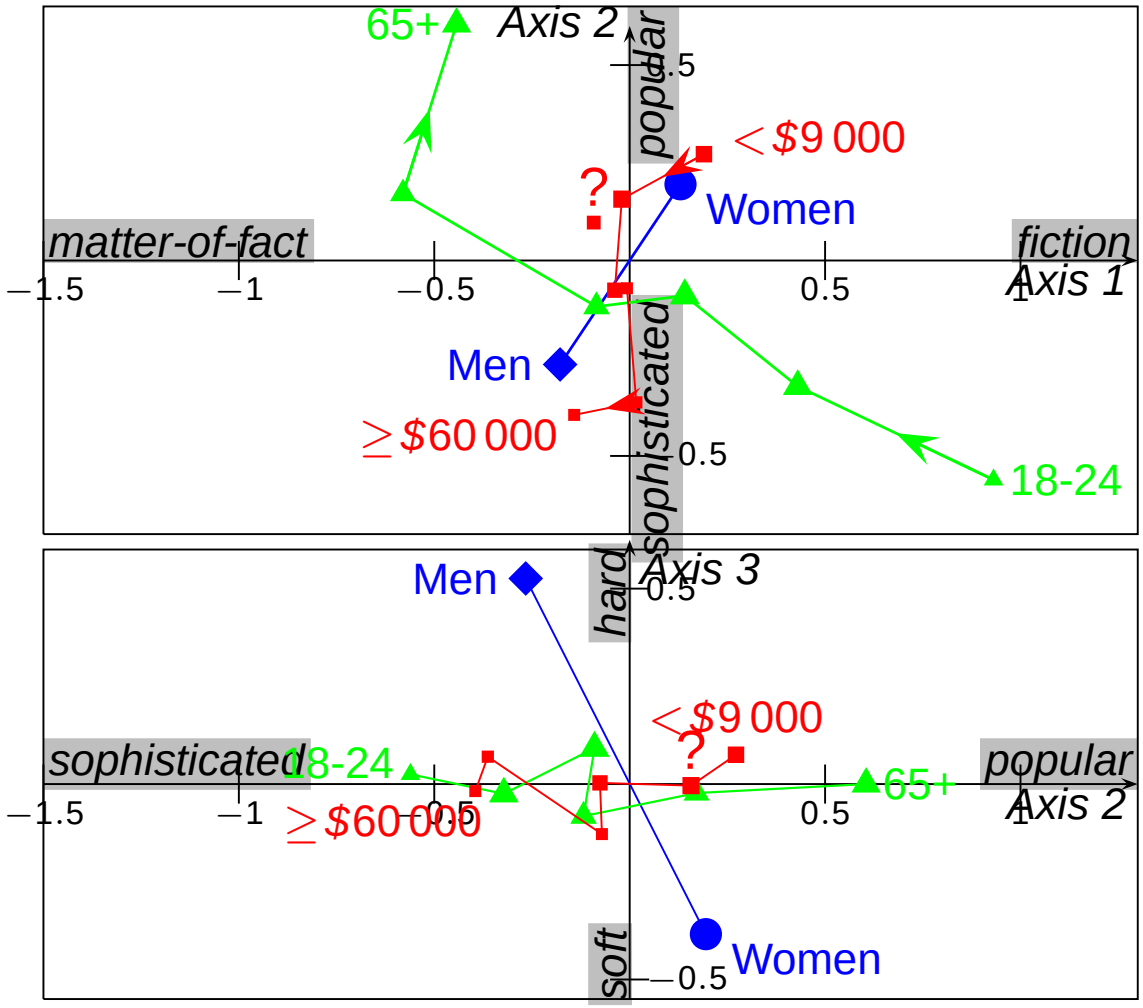
	weight	Axis 1	Axis 2	Axis 3
Men	513	−0.178	−0.266	+0.526
Women	702	+0.130	+0.195	−0.384
18-24	93	+0.931	−0.561	+0.025
25-34	248	+0.430	−0.322	−0.025
35-44	258	+0.141	−0.090	+0.092
45-54	191	−0.085	−0.118	−0.082
55-64	183	−0.580	+0.171	−0.023
≥ 65	242	−0.443	+0.605	+0.000

Income				
	weight	Axis 1	Axis 2	Axis 3
< $9 000	231	+0.190	+0.272	+0.075
$10-19 000	251	−0.020	+0.157	−0.004
$20-29 000	200	−0.038	−0.076	+0.003
$30-39 000	122	−0.007	−0.071	−0.128
$40-59 000	127	+0.017	−0.363	+0.070
> $60 000	122	−0.142	−0.395	−0.018
“unknown"	162	−0.092	+0.097	−0.050

As a *rule of thumb*:

— a deviation greater than 0.5 will be deemed to be “**notable**";

— a deviation greater than 1, definitely “**large**".



Supplementary questions in plane 1-2 (top), and in plane 2-3 (bottom) (cloud of categories).

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:

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

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 hierarchy (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 $\longrightarrow$ binary decision tree

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.

it is impossible to enumerate all the partitions of a set of n individuals into k clusters

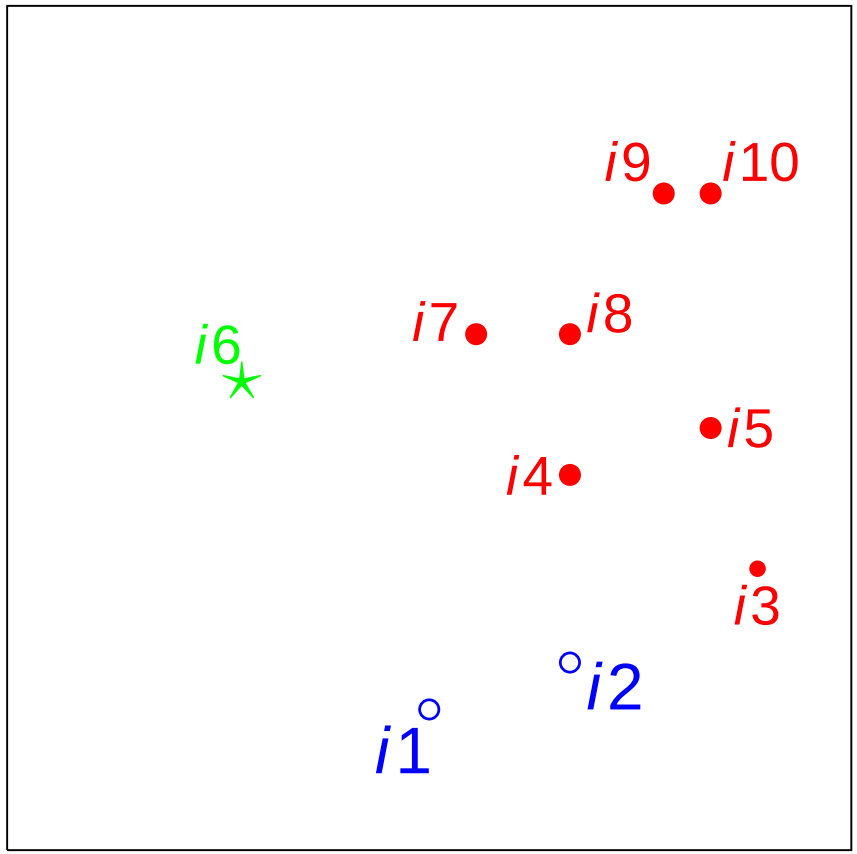
IV.2. Partition of a Cloud: Between– and Within–variance

- Subclouds

$\mathcal{A}$: subcloud of 2 points (dipole)
$\{i1, i2\}$

$\mathcal{B}$: subcloud of 1 point
$\{i6\}$

$\mathcal{C}$: subcloud of 7 points
$\{i3, i4, i5, i7, i8, i9, i10\}$

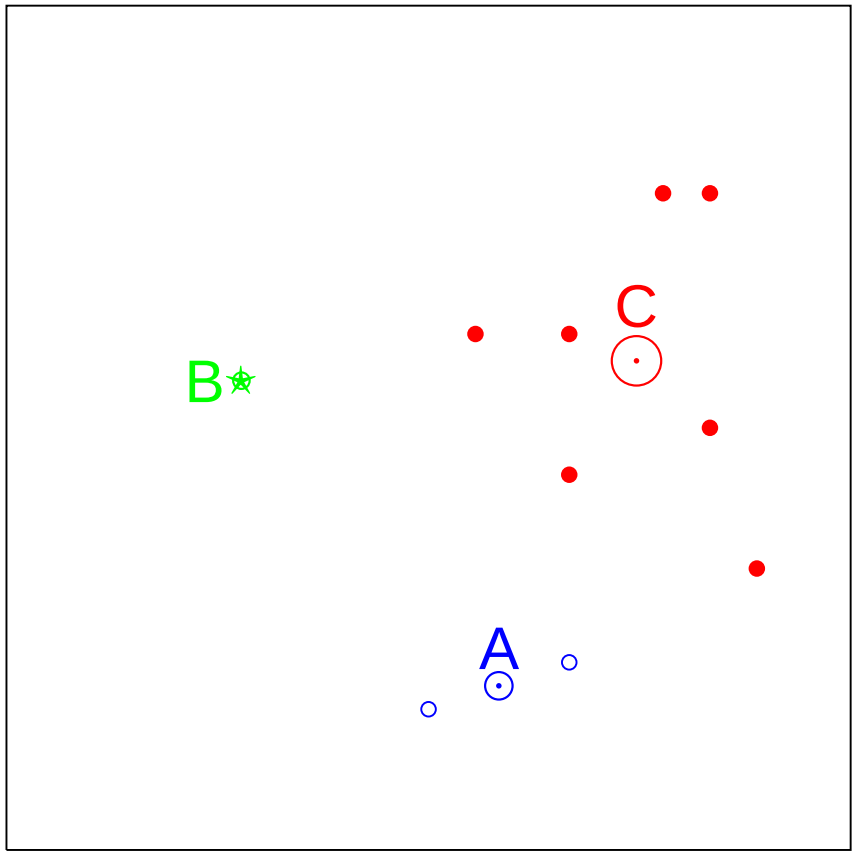


Partition of a cloud into 3 subclouds: $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$.

3 mean points A, B, C with weights 2, 1, 7.

By grouping:
— points “average up”
— weights add up

	weights	Coordinates x_1	Coordinates x_2	variances
A	$n_A = 2$	3	-11	10
B	$n_B = 1$	-8	2	0
C	$n_C = 7$	8.857	2.857	46.57
	$n = 10$	$\overline{x}_1 = 6$	$\overline{x}_2 = 0$	34.6



The mean of the variances of subclouds is the *within–variance*

Between-cloud

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

The between-cloud is a weighted cloud;

- its total weight is $n = 10$;
- its mean point is G;
- its variance is $\frac{2}{10}(\mathrm{GA})^2 + \frac{1}{10}(\mathrm{GB})^2 + \frac{7}{10}(\mathrm{GC})^2 = 57.4$ and called *between–variance*

Contributions of a subcloud

The *contribution of a subcloud* is the sum of the contributions of its points.

The *within-contribution* of a subcloud is the product of its weight by its variance and divided by V_{cloud}.

— *Example*: subcloud $\mathcal{A}$

$$\text{Ctr}_{i1} = \frac{\frac{1}{10}(\text{GM}^{i1})^2}{92} = \frac{\frac{1}{10}\times 180}{92} = \frac{18}{92}; \quad \text{Ctr}_{i2} = \frac{\frac{1}{10}(\text{GM}^{i2})^2}{92} = \frac{\frac{1}{10}\times 100}{92} = \frac{10}{92}$$

- contribution of the *subcloud*: $\text{Ctr}_{\mathcal{A}} = \frac{18}{92} + \frac{10}{92} = \frac{28}{92}$
- contribution of the *mean point*: $\text{Ctr}_{\text{A}} = \frac{\frac{2}{10}\times 130}{92} = \frac{26}{92}$
- *within–contribution*: $\frac{\frac{2}{10}\times 10}{92} = \frac{2}{92}$

Huyghens theorem

The contribution of a subcloud is the sum of the contribution of its mean point and of its within-contribution.

Example: Subcloud $\mathcal{A}$

$$\text{Ctr}_{\mathcal{A}} = \text{Ctr}_{\text{A}} + \text{within–contribution}$$

$$\frac{28}{92} = \frac{26}{92} + \frac{2}{92}$$

Between–within decomposition of variance

	Ctr $\times V_{\text{cloud}}$		
	mean points	within	subclouds
$\mathcal{A}$	26.0	2.0	28
$\mathcal{B}$	20.0	0	20
$\mathcal{C}$	11.4	32.6	44
Total	57.4	34.6	92
Variance	between	within	total

Within-variance

= sum of within–contributions $\times V_{\text{cloud}}$

= weighted mean of variances of subclouds $(\frac{2}{10} \times 10 + 0 + \frac{7}{10} \times 46.6)$

$= 34.6$

Total variance = between-variance + within-variance

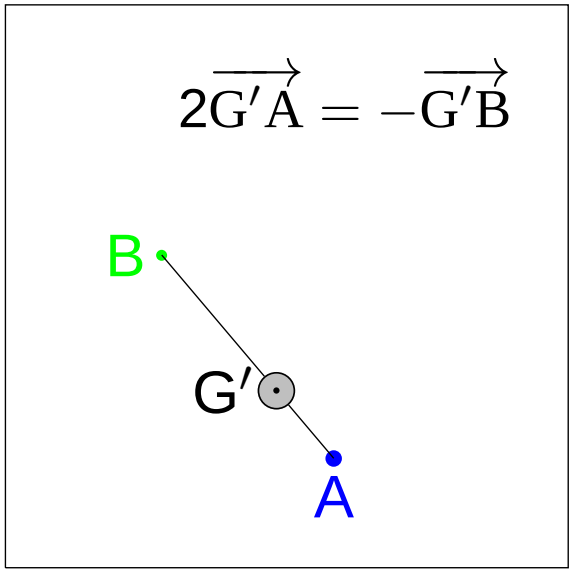
$$\eta^2 = \frac{\text{between-variance}}{\text{total variance}} \text{ (eta-square)}$$

Subcloud of 2 points (dipole)

A and B weighted by $n_A = 2$ and $n_B = 1$ with mean point G′.

Weight of dipole : $\widetilde{n_{AB}} = 1/(\frac{1}{n_A} + \frac{1}{n_B})$

Absolute contribution: $p \times d^2$ with $p = \frac{\widetilde{n_{AB}}}{n}$ (relative weight) and $d^2 = \mathrm{AB}^2$ (square of the deviation).



Example: dipole $\{\mathrm{A}, \mathrm{B}\}$.

$\mathrm{AB}^2 = 290$

$\widetilde{n_{AB}} = \frac{1}{\frac{1}{1}+\frac{1}{2}} = 2/3, p = \frac{2/3}{10} = 0.06667$

Absolute contribution:

$0.06667 \times 290 = 19.33$

Property

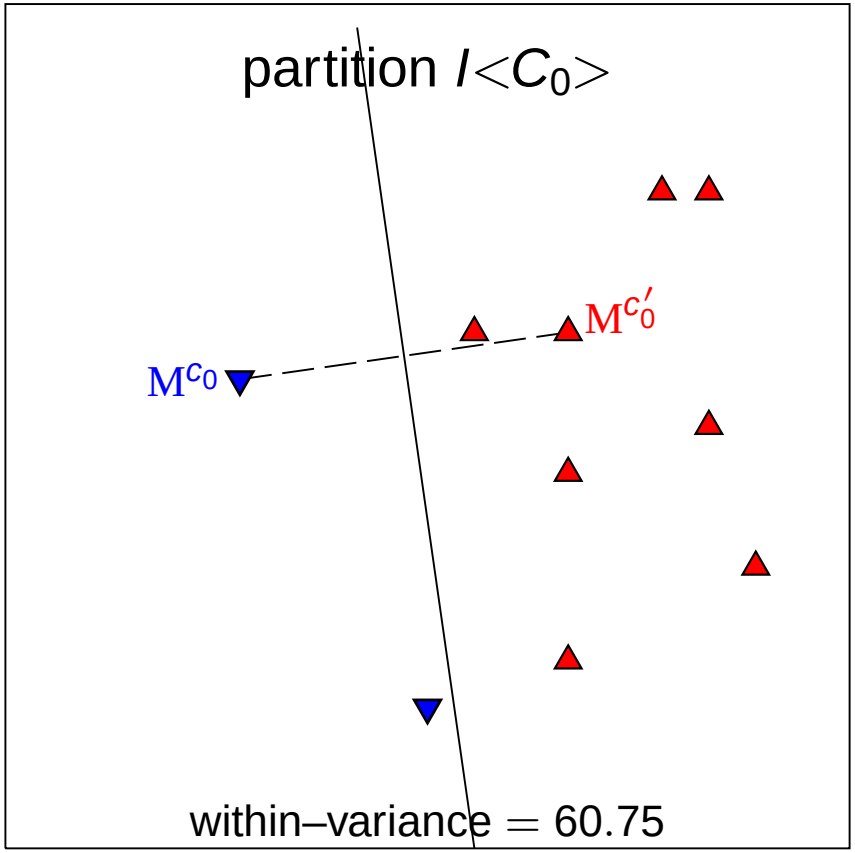
The absolute contribution of a dipole is the absolute contribution of the subcloud of its two points.

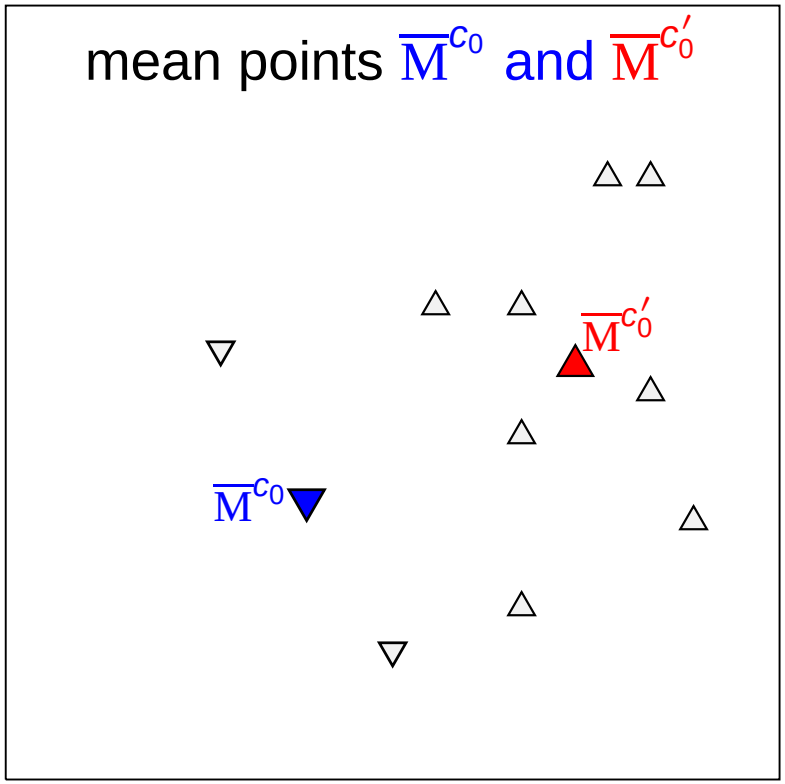
IV.3. K–means Clustering

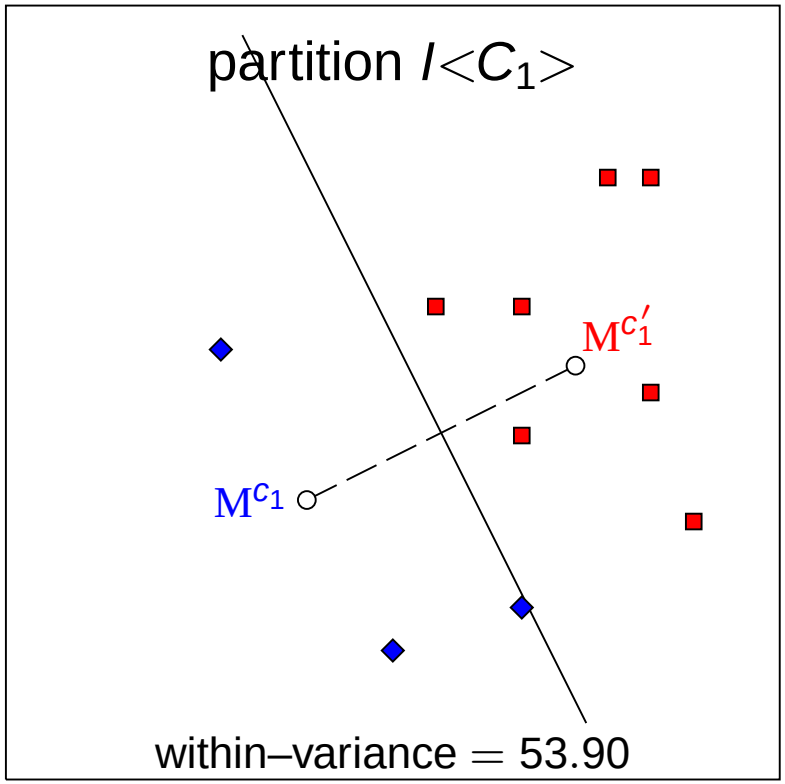
or aggregation around *moving centers*

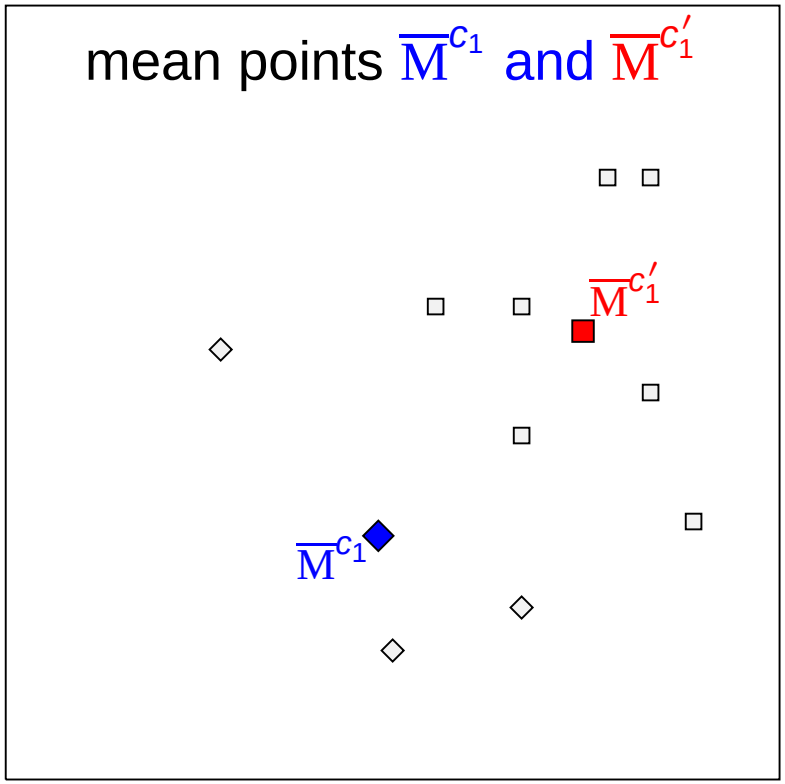
1. Fix the number of clusters, say C;
2. Choose (randomly or not) C initial class centers;
3. Assign each object to the closest center $\rightarrow$ new clusters;
4. Determine the centers of the new clusters;
5. Repeat the assignment;
6. Stop the algorithm when 2 successive iterations provide the same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

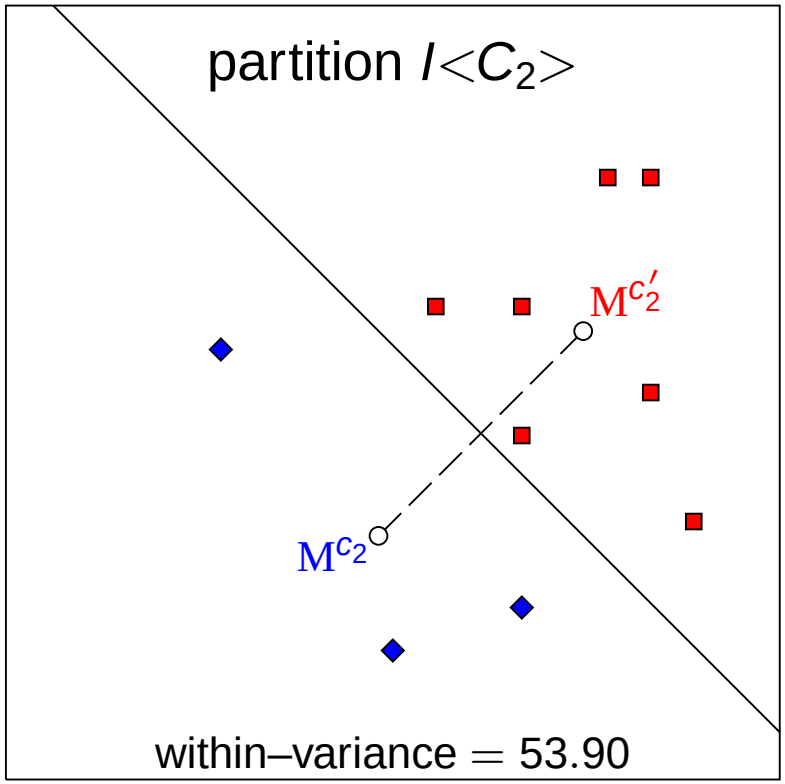
Choose 2 initial centers: M^{c_0} and $\mathrm{M}^{c_0'}$




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


mean points $\overline{M}^{c_1}$ and $\overline{M}^{c'_1}$
$\overline{M}^{c'_1}$
$\overline{M}^{c_1}$


partition $I<C_2>$
$M^{c'_2}$
M^{c_2}
within–variance = 53.90

IV.4. 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.

Ascending/agglomerative Hierarchical Clustering: starting with the basic objects (one–element clusters) proceed to successive aggregations until all objects are grouped in a single class.

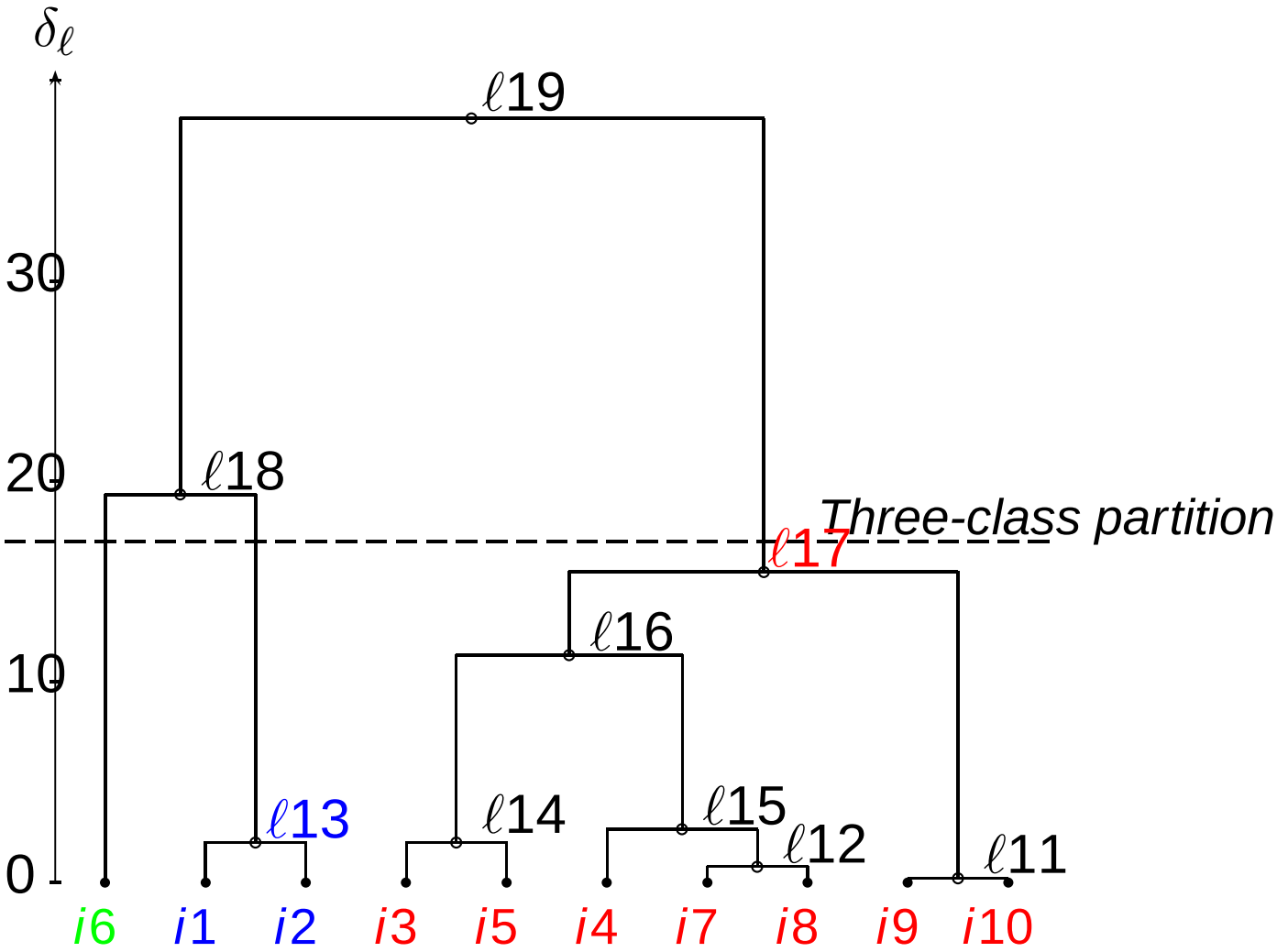
Once an aggregation index has been chosen, the *basic algorithm* of AHC is as follows:

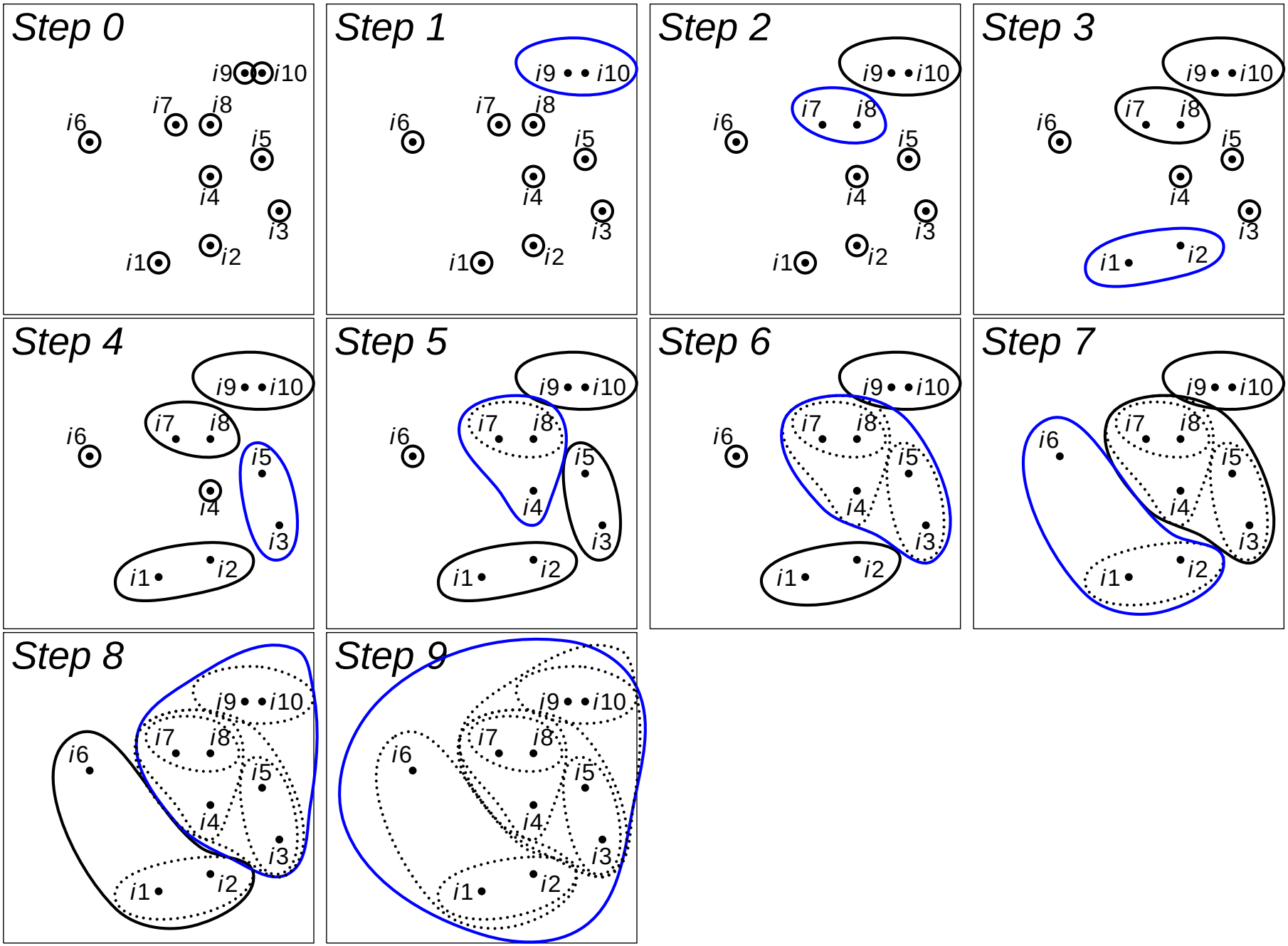
Step 1. From the table of distances between the n objects, calculate the aggregation index for the $n(n-1)/2$ pairs of one–element clusters, then aggregate a pair of clusters for which the index is minimum: hence a partition into $J-1$ clusters.

Step 2. Calculate the aggregation indices between the new class and the $n-2$ others, and aggregate a pair of clusters for which the index is minimum $\rightarrow$ second partition into $n-2$ clusters in which the first partition is nested.

Step 3. Iterate the procedure until a single class is reached.

Target example: hierarchical tree




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

IV.5. Euclidean Clustering

1. Objects = *points of Euclidean cloud*.
2. *Aggregation index* = variance index, that is, the contribution of the dipole of the class centers (Ward index).

Grouping property

If 2 clusters are *grouped*, the between–variance *decreases* from an amount equal to the contribution of the dipole constituted of the centers of the 2 grouped clusters.

Basic Algorithm

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

Example: For dipole $\{i1, i2\}$:

$\widetilde{n_{12}} = 1/(\frac{1}{1} + \frac{1}{1}) = 0.5$;

squared distance = $(0-6)^2 + (-12+10)^2 = 40$;

$\rightarrow$ absolute contribution of dipole = $\frac{0.5}{10} \times 40 = 2$.

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

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).

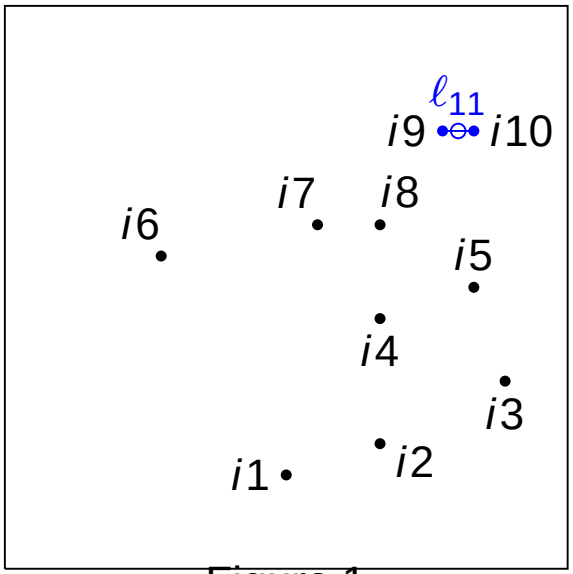
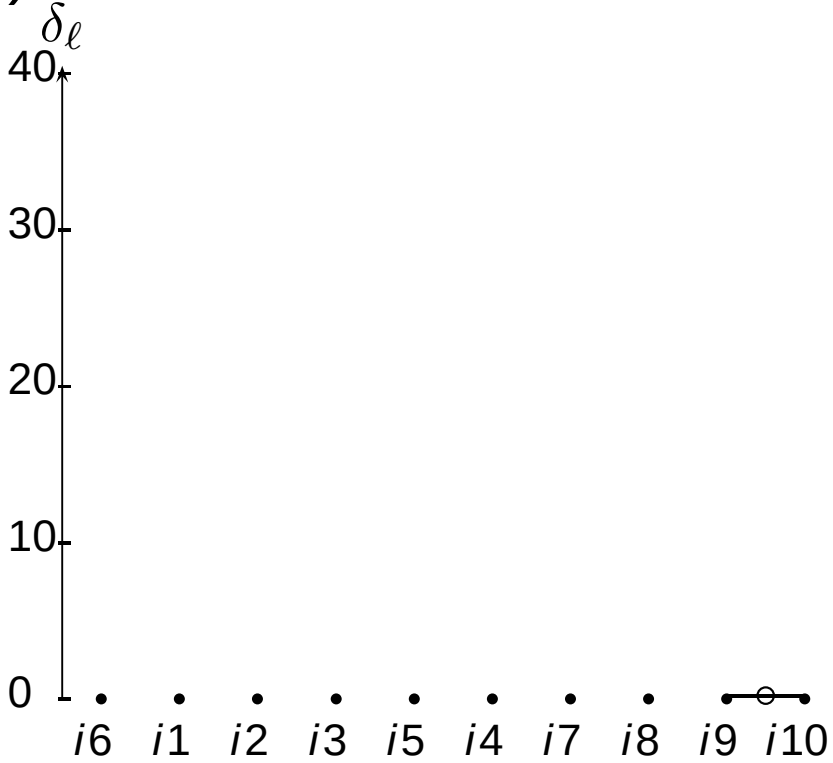


Figure 1



• **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 $\ell 12$ 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

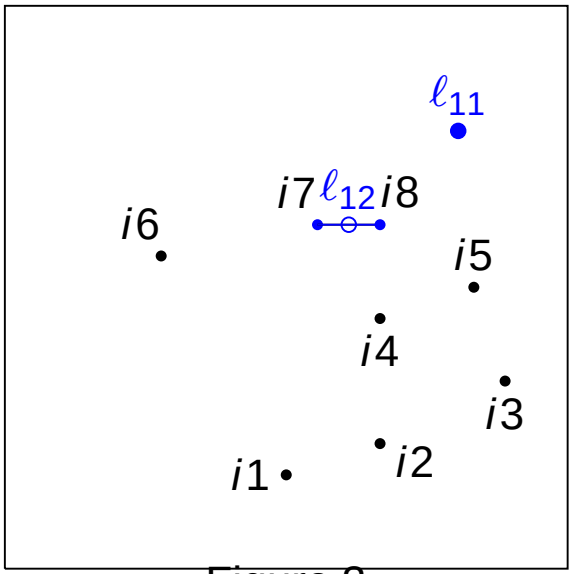
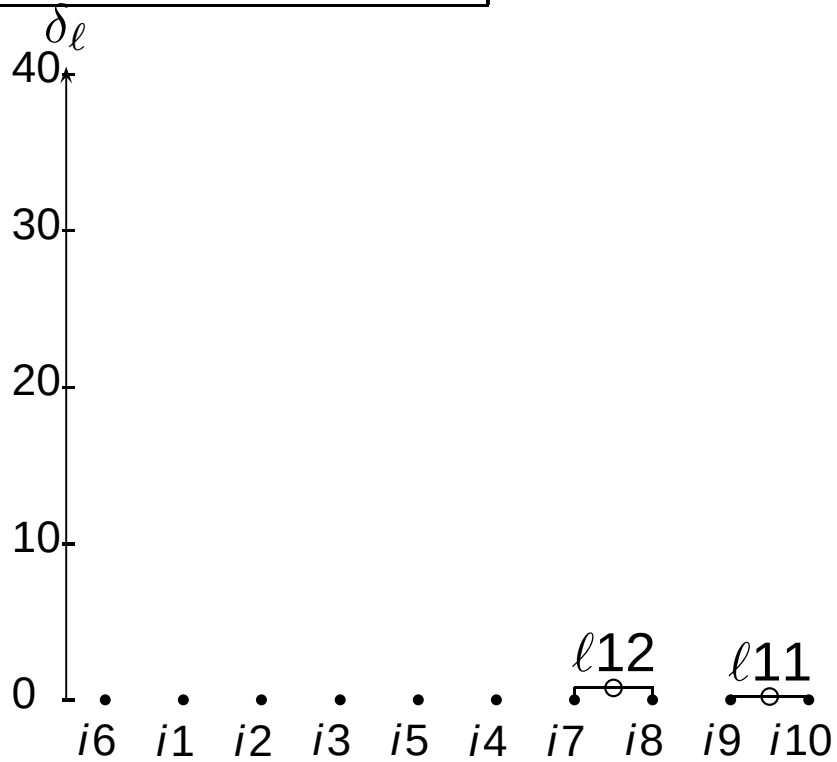


Figure 2



- **Step 3.** Iterate the procedure

Aggregation index between $\ell 12$ 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 = 2 for $\{i1, i2\}$, $\{i3, i5\}$ and $\{i4, i5\}$, aggregation of $i1$ and $i2$ (fig. 3), hence the point $\ell 13$ and a *cloud of 7 points* (fig. 4).

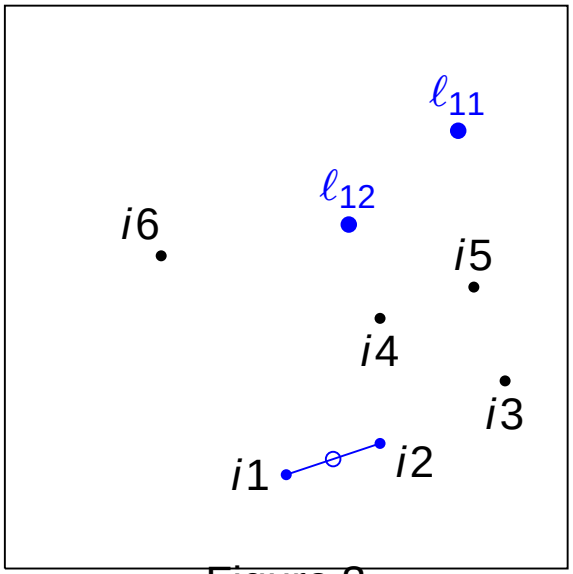
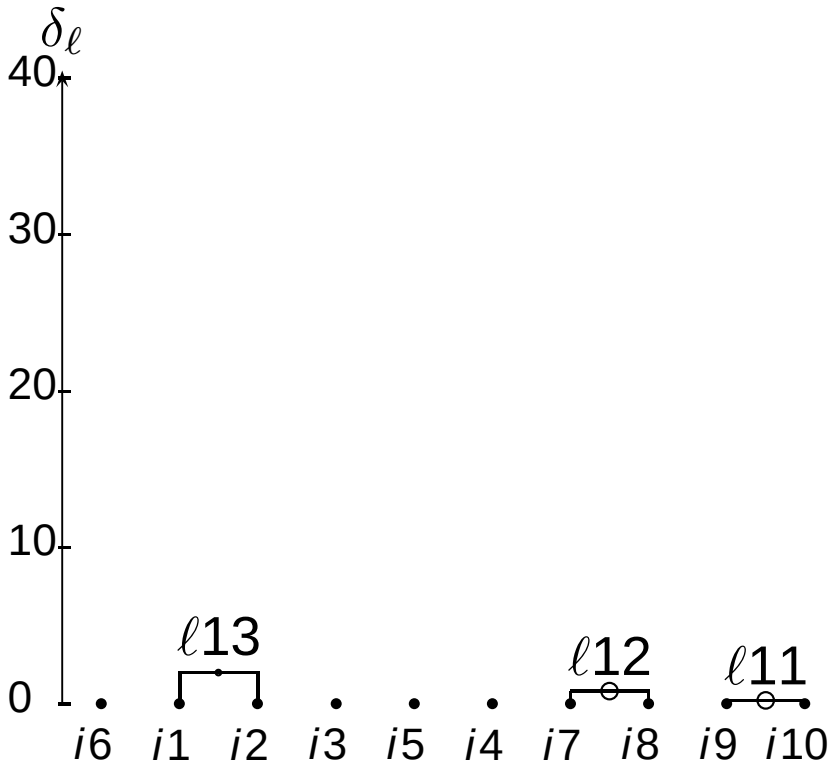


Figure 3



- **Step 4.** Iterate the procedure

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

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

Minimum of index = 2 for the two pairs $\{i3, i5\}$ and $\{i4, i5\}$.

Aggregation of $i3$ and $i5$ (fig. 4), hence the point $\ell 14$ and the *cloud of 6 points* (fig. 5).

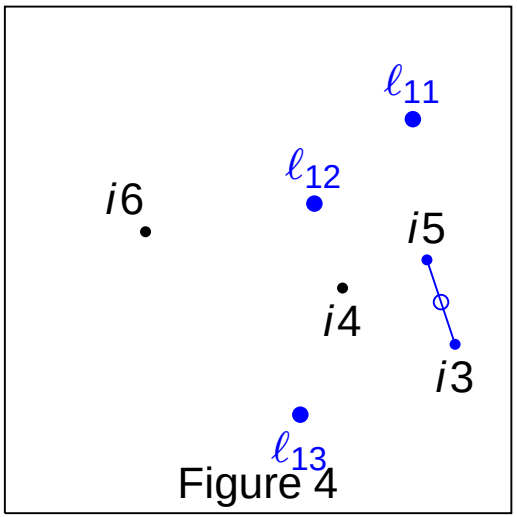
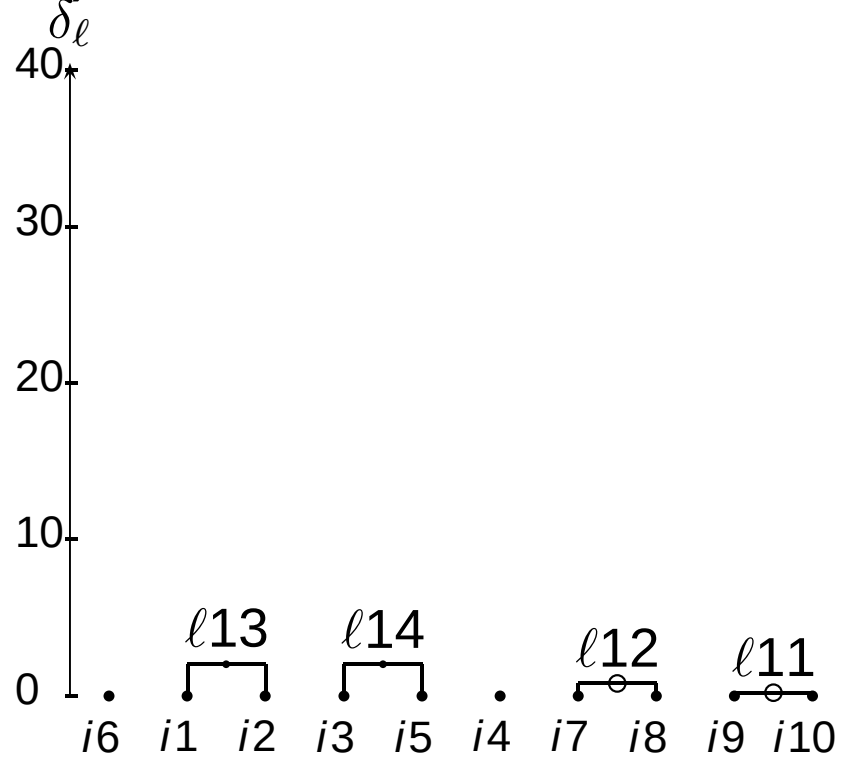


Figure 4



- **Step 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

$\rightarrow$ 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).

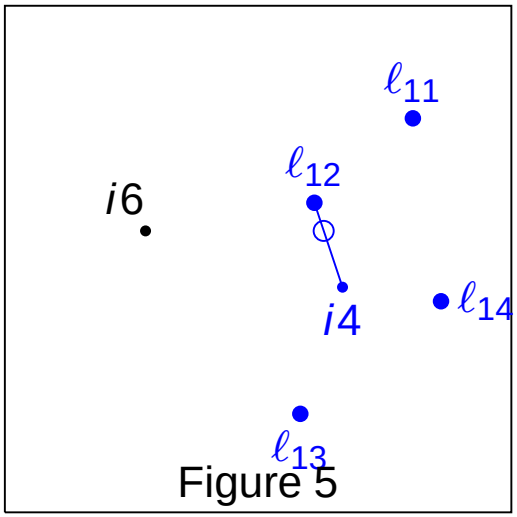
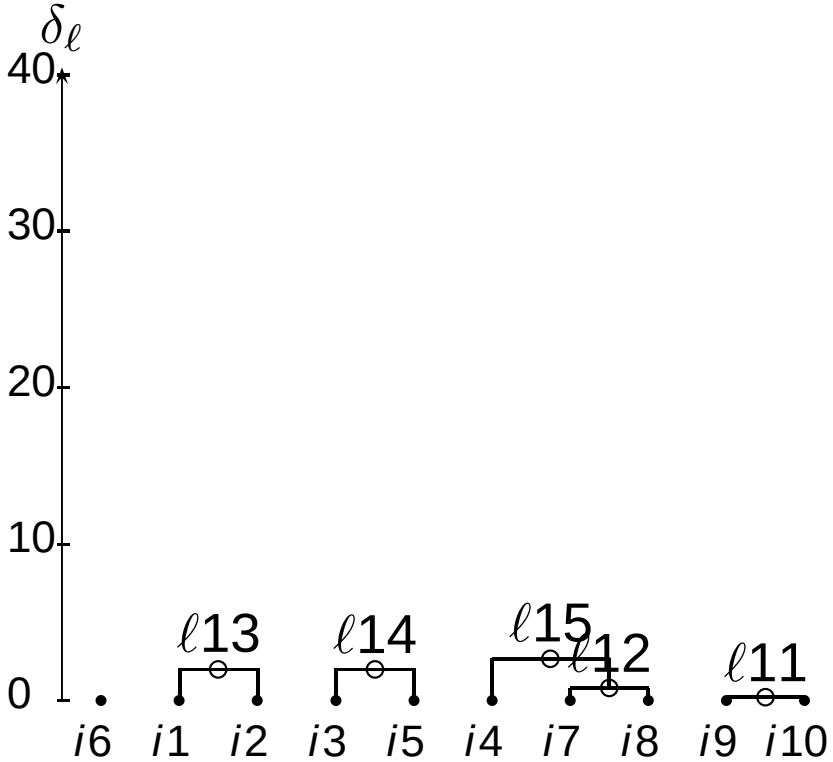


Figure 5



- **Step 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

$\rightarrow$ 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).

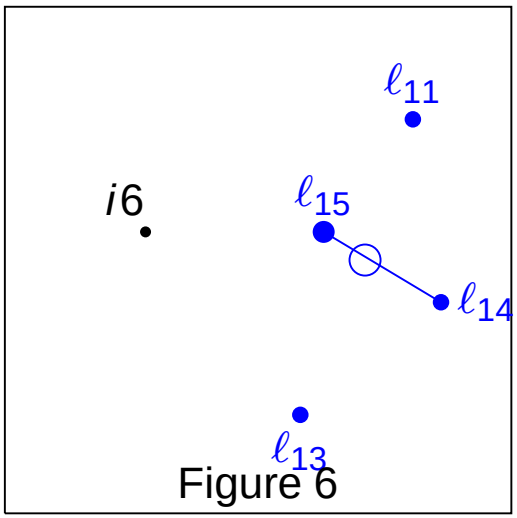
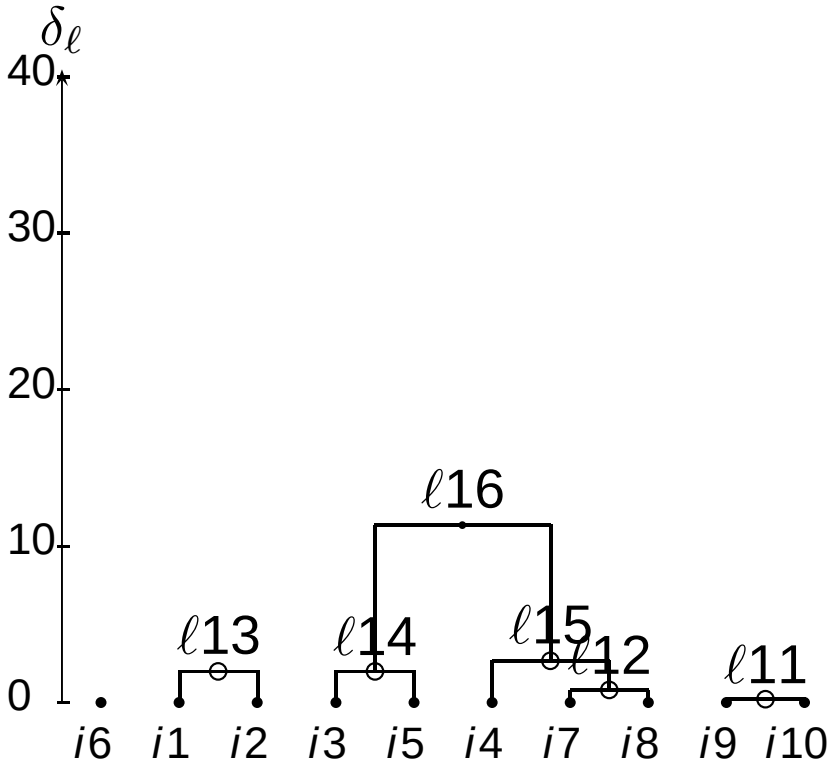


Figure 6



- **Step 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).

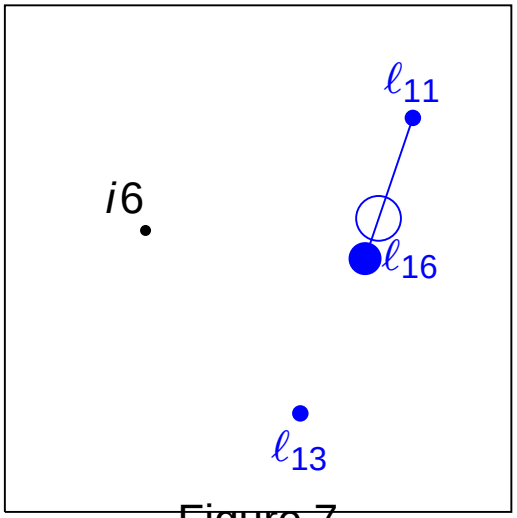
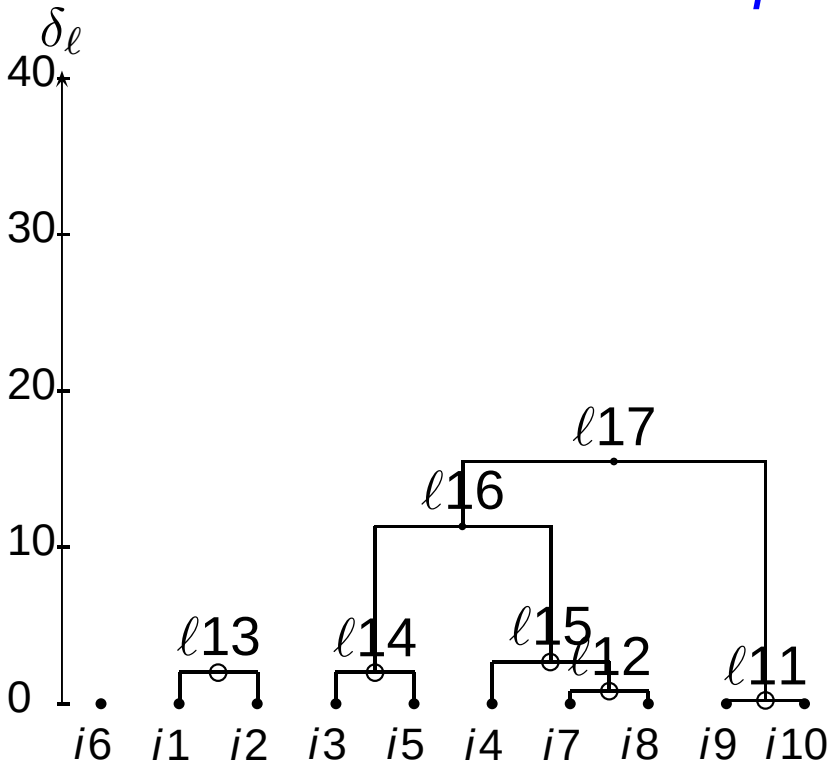


Figure 7



- **Step 8.**

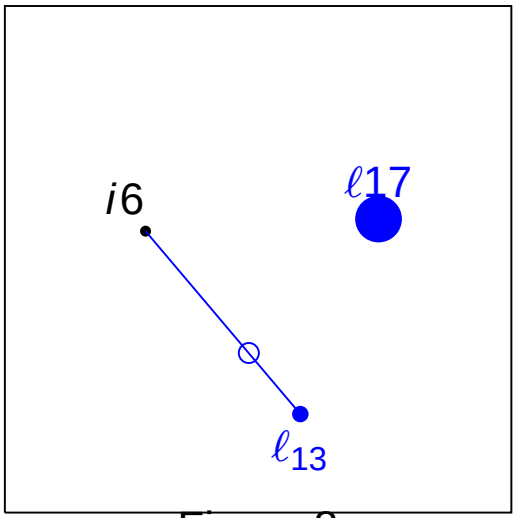
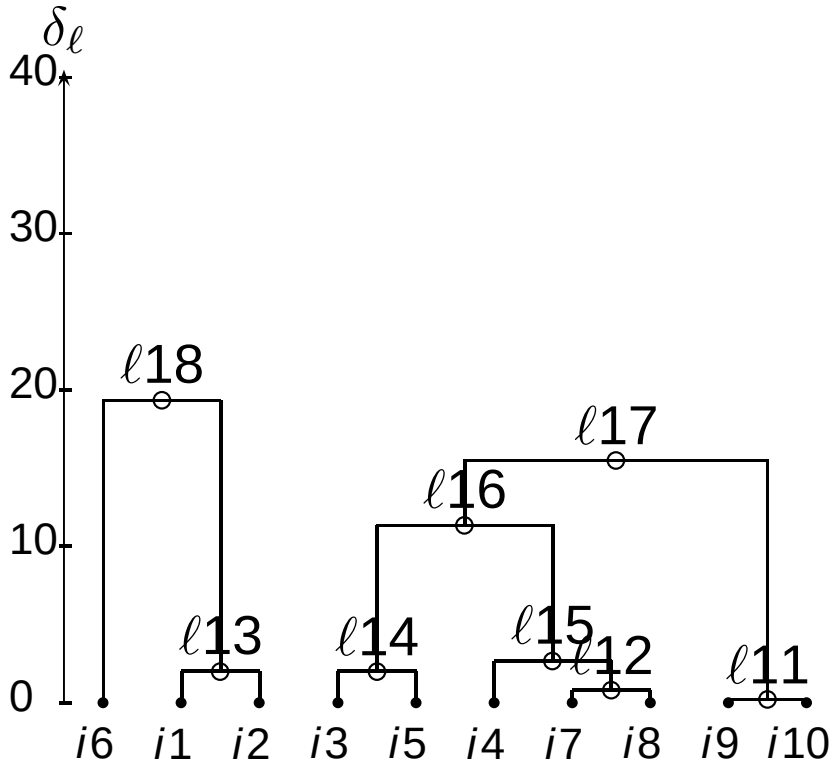


Figure 8



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.**

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

$(\text{distance})^2$	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).

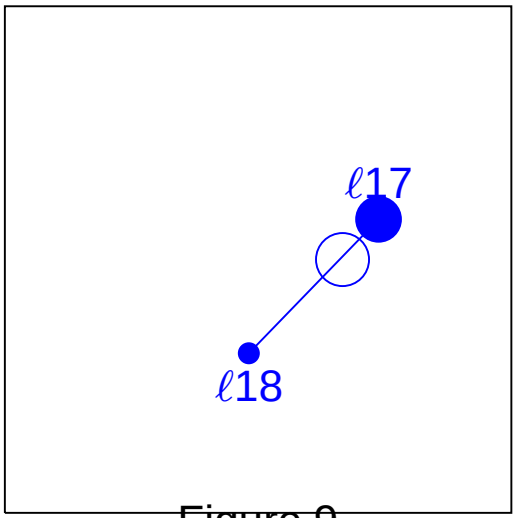
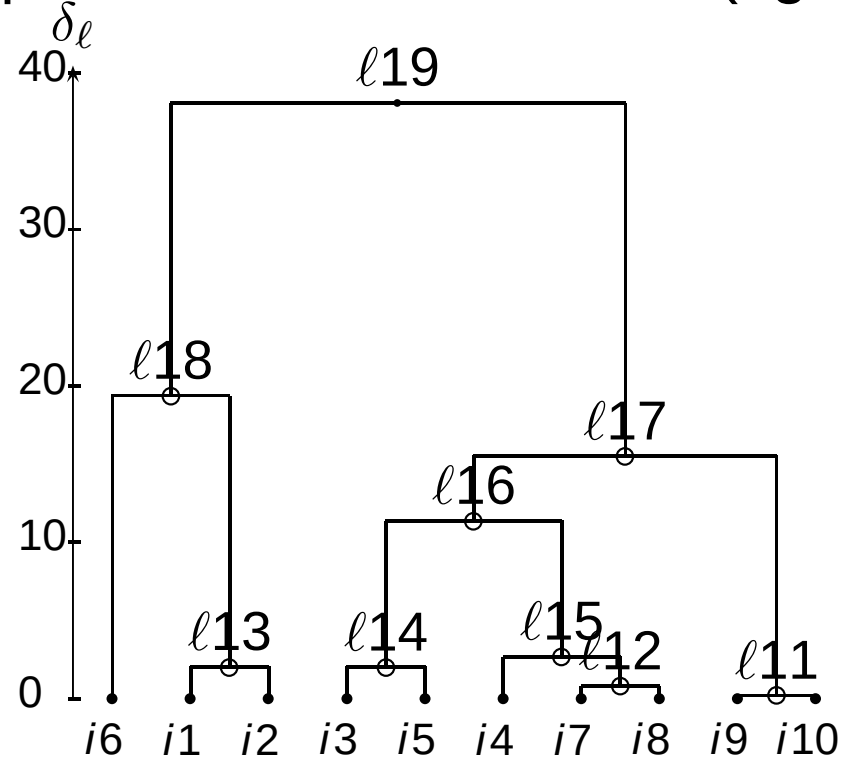


Figure 9



Successive Steps of the AHC

ℓ	δ_ℓ	clusters		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
$\ell 19$		
$\ell 18$	38.10	.414
$\ell 17$	57.43	.624
$\ell 16$	73.00	.793
$\ell 15$	84.33	.917
$\ell 14$	87.00	.957
$\ell 13$	89.00	.967
$\ell 12$	91.90	.989
$\ell 11$	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.6. Interpretation of clusters

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: $f_c - f < -0.05 \qquad f_c/f < 2$

2. inductive criterion:

The hypergeometric test of comparison of the frequency 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.

IV.7. 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.8. Divisive Hierarchical Clustering

Sart 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

V — Specific MCA

and

Class Specific Analysis (CSA)

This text is adapted from Chapter 3 (§3.3) of the monograph
Multiple Correspondence Analysis
(QASS series n°163, SAGE, 2010)

V.1. Introduction

- *Specific MCA (SpeMCA)* consists in restricting the analysis to *categories of interest*.
- *Class Specific Analysis (CSA)* consists in analyzing a *subset of individuals* by taking the whole set of individuals as a reference.

V.2. Specific MCA

The active categories are the *categories of interest*.
The excluded categories, called *passive categories*, are:

- *Infrequent categories*
  - — remote from the center
  - — contributing too much to the variance of the question
  - — too influential on the determination of axes

- *Junk categories*: categories of *no-interest*
  not representable by a single point

Cloud of individuals

If for active question q,

• i chooses active category k and i' active category k', then the distance is unchanged:

$$d_q^{2\prime} = d_q^2(i, i') = \frac{1}{f_k} + \frac{1}{f_{k'}}$$

• i chooses active category k and i' passive category k':

$$d_q^2(i, i') = \frac{1}{f_k} \text{ (dropping } \frac{1}{f_{k'}}\text{)}$$

Geometric viewpoint:

$\longrightarrow$ projection of the cloud onto a subspace of interest.

Cloud of categories

subcloud of categories of active questions with weights and distances unchanged.

K': set of *active* categories of active questions
K'': subset of *passive* modalities of active questions
K: set of *active and passive* categories of active questions
Q': set of active questions without passive categories

Properties

- Dimension of the cloud: $K' - Q'$

(number of active categories minus number of questions without passive categories).

- Specific overall variance:

$$\frac{K'}{Q} - \sum_{k \in K'} \frac{f_k}{Q} = \text{sum of eigenvalues}$$

- Modified rates:

calculate $\overline{\lambda} =$ specific variance divided by the number of dimensions of the cloud;

modified rates $= \dfrac{(\lambda - \overline{\lambda})^2}{\sum(\lambda - \overline{\lambda})^2}$ ($\sum$ over eigenvalues $> \overline{\lambda}$).

Principal axes and principal variables

- Coordinates of individuals on an axis :

  Mean = 0 Variance = specific eigenvalue

- Coordinate of categories on an axis:
  - Mean of coordinates of *active and passive* categories (weighted by the relative weight f_k/Q) = 0
  - Raw sum of squares of coordinates of *active* categories (weighted by $p_k = f_k/Q$) $= \lambda$

Fundamental properties of standard MCA are preserved:

- the principal axes of the cloud of individuals are in a one-one correspondence with those of the cloud of categories,
- the two clouds have the same eigenvalues.
- Link between the two clouds:

$$\overline{y} = \sqrt{\lambda}\, y$$

(y: principal coordinate of category k
$\overline{y}$: principal coordinate of category mean–point k)

V.3. Class Specific Analysis (CSA)

Study of a class (subset) of individuals with reference to the whole set of individuals.

We seek to
— determine the specific features of the class,
— compare the *class subcloud* with the *initial cloud*.

This is possible only if the class subcloud and the initial cloud are in the same Euclidean space.

Class specific cloud of individuals

The distance between 2 individuals of the class is the one defined from the whole cloud.

Class specific cloud of categories

The distance between two categories points depends on

- the relative frequencies of the categories in the class,
- the relative frequencies of the categories in the whole set,
- the conjoint frequency of the pairs of categories in the class.

Principal axes and principal variables

- Coordinates of individuals on an axis :

  Mean = 0 Var = specific eigenvalue

- Coordinate of categories on an axis (weighted by the relative weight in the whole set):

  Mean = 0 Var = specific eigenvalue