Graduate course

Geometric Data Analysis (GDA)

Brigitte Le Roux

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

[1]MAP5/CNRS, Université Paris Descartes

[2]CEVIPOF/CNRS, SciencesPo Paris

September 17–21, 201 Campus Gotland, Uppsala University

Table of Contents I

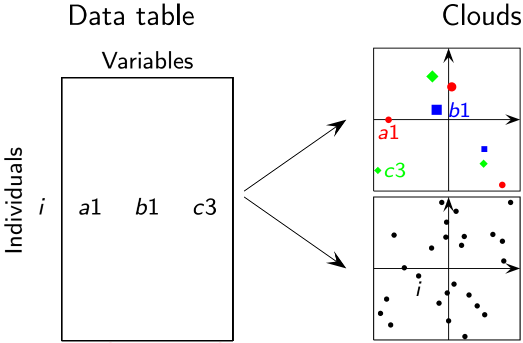
Table of Contents II

Table of Contents III

I – Introduction to Geometric Data Analysis methods

I.1. The Three Key Ideas of GDA

1. Geometric modeling



Cloud of categories: Points represent the categories of variables.

Cloud of individuals: Points represent individuals.

2. Formal approach.

Structures govern procedures!

3. Inductive philosophy

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

I.2. Three Paradigms

- *Correspondence Analysis* (CA)
  ⟶ Contingency table
- *Principal Component Analysis* (PCA)
  ⟶ Individuals×Numerical Variables table
- *Multiple Correspondence Analysis* (MCA)
  ⟶ Individuals×Categorical Variables table

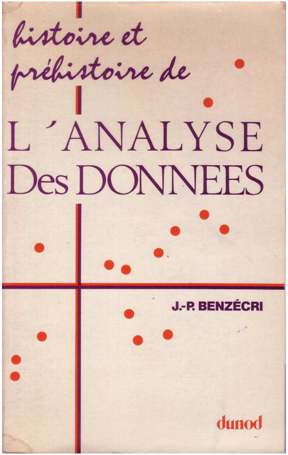
I.3. Frame model

In Geometric Data Analysis, two principles should be followed (Benzécri, 1992, pp. 382-383):

- *Homogeneity*
- *Exhaustiveness*

Benzécri, 1992, pp. 382-383

I.4. Historical Sketch



J-P. Benzécri (1982)

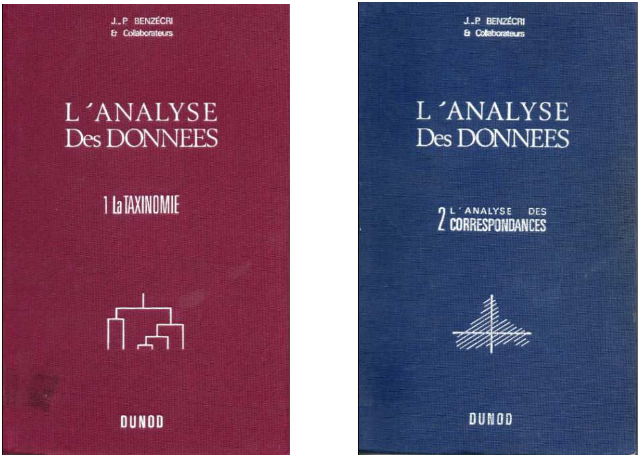
Precursors

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

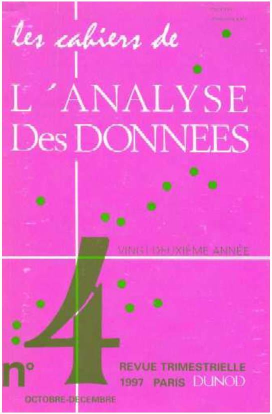
- Optimal scaling: Fisher (1940), Guttman (1942)
- Factor analysis: Burt (1950)
- Quantification method: Hayashi (1952)
- MDS: Shepard (1962).

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)



les cahiers de
L'ANALYSE
Des DONNEES

VINGT-DEUXIÈME ANNÉE

n° 4

REVUE TRIMESTRIELLE
1997 PARIS DUNOD

OCTOBRE-DÉCEMBRE

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;

Goodman (1986, 1991), Rao (1995).

Gower & Hand (1996): *Biplot*.

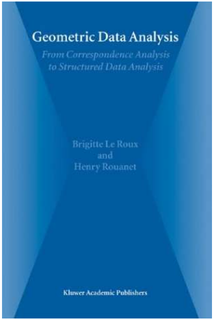
Where do we stand now?

CARME network:
international conferences in Cologne (1991,1995,1999), Barcelona (2003), Rotterdam (2007), Rennes (2011), Naples (2015), and the next one in Stellenbosch,South Africa (2019).

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

Recent Books:

Le Roux & Rouanet



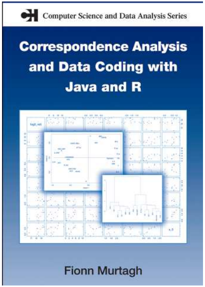
Geometric Data Analysis

From Correspondence Analysis to Structured Data Analysis

Brigitte Le Roux
and
Henry Rouanet

Kluwer Academic Publishers

Murtagh
2005

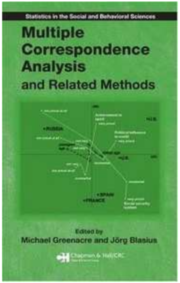


Computer Science and Data Analysis Series

Correspondence Analysis and Data Coding with Java and R

Fionn Murtagh

CARME
2003 (2006)



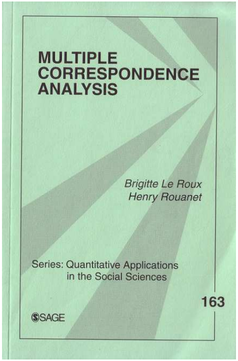
Statistics in the Social and Behavioral Sciences

Multiple Correspondence Analysis and Related Methods

Edited by
Michael Greenacre and Jörg Blasius

Chapman & Hall/CRC

Le Roux & Rouanet
2010



MULTIPLE CORRESPONDENCE ANALYSIS

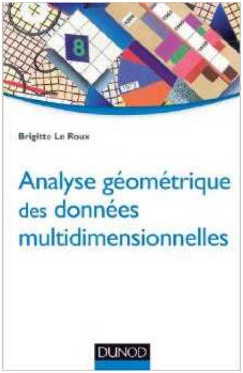
Brigitte Le Roux
Henry Rouanet

Series: Quantitative Applications in the Social Sciences

163

SAGE

Le Roux
2014

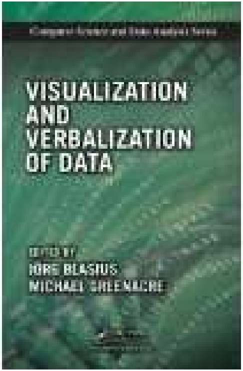


Brigitte Le Roux

Analyse géométrique des données multidimensionnelles

DUNOD

CARME
2011 (2015)

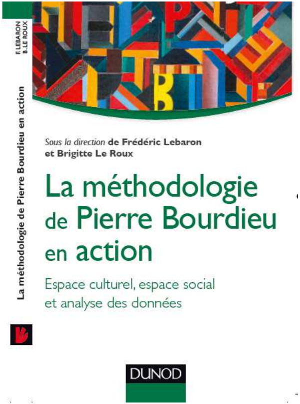


Computer Science and Data Analysis Series

VISUALIZATION AND VERBALIZATION OF DATA

EDITED BY
JÖRG BLASIUS
MICHAEL GREENACRE

Lebaron & Le Roux (eds)
2015



F. LEBARON
B. LE ROUX

La méthodologie de Pierre Bourdieu en action

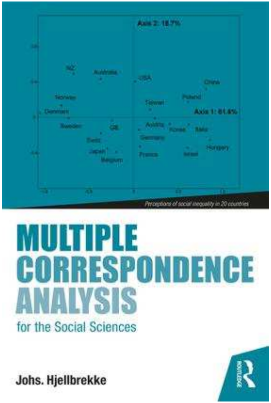
Sous la direction de **Frédéric Lebaron et Brigitte Le Roux**

La méthodologie de Pierre Bourdieu en action

Espace culturel, espace social et analyse des données

DUNOD

Hjellbrekke
2017



MULTIPLE CORRESPONDENCE ANALYSIS
for the Social Sciences

Johs. Hjellbrekke

CA is now recognized and used, but GDA as a whole methodology, is beginning 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*.

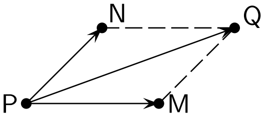
— *Affine notions*: alignment, direction and barycenter.

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

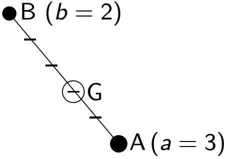
The *deviation* from point P to point M is M − P ("terminal minus initial"), that is, $\overrightarrow{\mathrm{PM}}$.

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

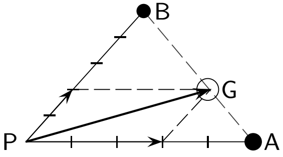
$$\overrightarrow{\mathrm{PM}} + \overrightarrow{\mathrm{PN}} = \overrightarrow{\mathrm{PQ}}$$



Barycenter of a dipole



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



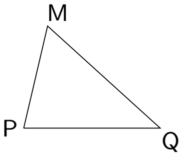
$\overrightarrow{\mathrm{PG}} = \frac{3}{5}\,\overrightarrow{\mathrm{PA}} + \frac{2}{5}\,\overrightarrow{\mathrm{PB}}$

Barycenter = *weighted average of points*: $\mathrm{G} = \dfrac{a\mathrm{A} + b\mathrm{B}}{a + b}$

— *Metric notions*: distances and angles.

Triangle inequality:

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

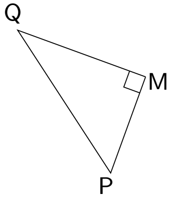


Pythagorean theorem:

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)

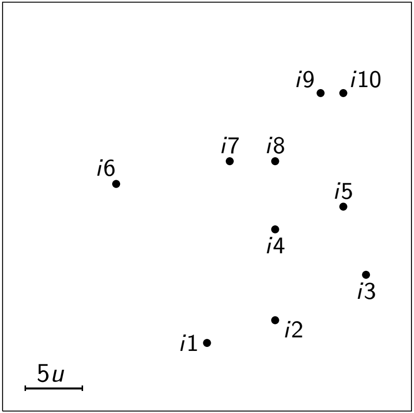
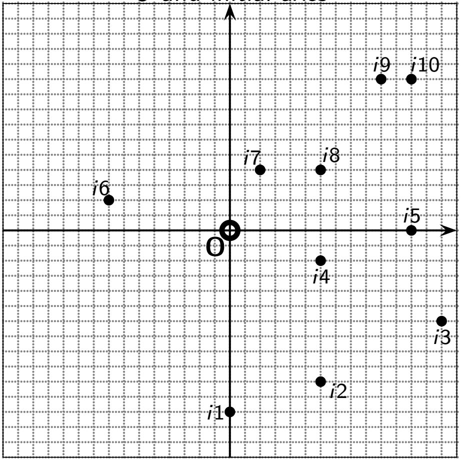


Figure 1b. Target example with origine-point O and initial axes



Initial coordinates

	x_1	x_2	weights
*i*1	0	−12	1
*i*2	6	−10	1
*i*3	14	−6	1
*i*4	6	−2	1
*i*5	12	0	1
*i*6	−8	2	1
*i*7	2	4	1
*i*8	6	4	1
*i*9	10	10	1
*i*10	12	10	1
Means	6	0	[10]
Variances	40	52	
Covariance	+8		

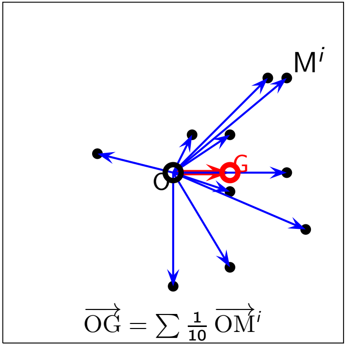
Mean Point

Cloud of points M^i with relative weights p_i

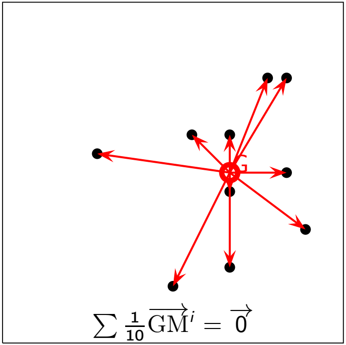
Mean point: point G

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

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



$$\overrightarrow{\mathrm{OG}} = \sum \tfrac{1}{10} \, \overrightarrow{\mathrm{OM}}^i$$



$$\sum \tfrac{1}{10} \overrightarrow{\mathrm{GM}}^i = \overrightarrow{0}$$

Variance, contribution

Variance of a cloud :

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

Property

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

Contribution of point M^i:

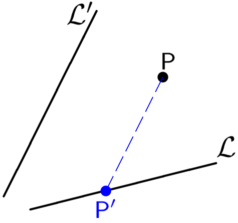
$$\mathrm{Ctr}_i = \frac{p_i(\mathrm{GM}^i)^2}{V_{\text{cloud}}}$$

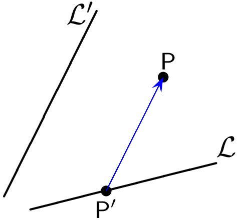
II.3. Principal Axes of a Cloud

Projection of a cloud

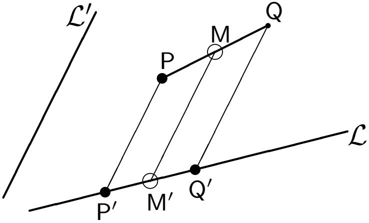
P′ = projection of point P onto $\mathcal{L}$ along $\mathcal{L}'$

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





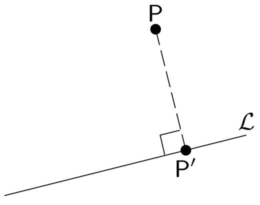
If point M is the midpoint of P and Q, the point M′, projection of M on $\mathcal{L}$, is the midpoint of P′ and Q′.

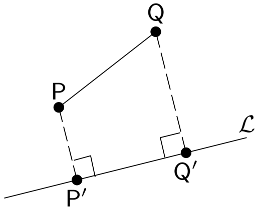


Mean point property

The mean point is preserved by projection.

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



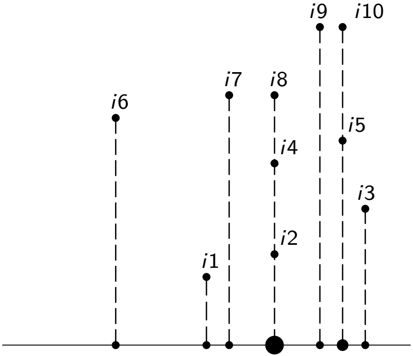


The orthogonal projection contracts distances: $\mathrm{P'Q'} \leq \mathrm{PQ}$, therefore one has the

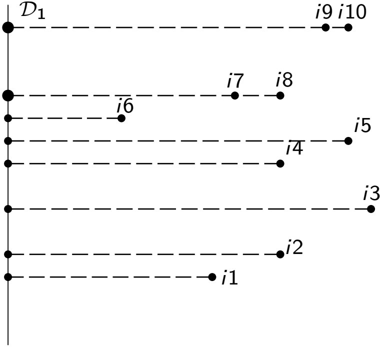
Property

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

Projected clouds on several lines



variance=40

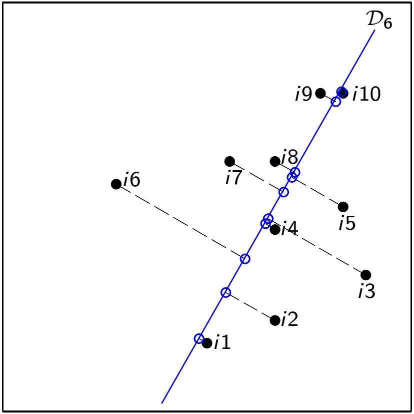


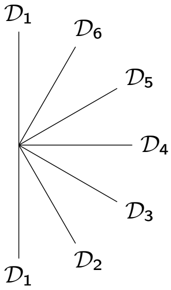
variance = 52

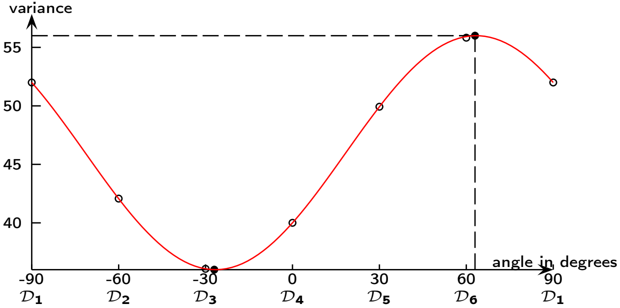
Orthogonal additive decomposition

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

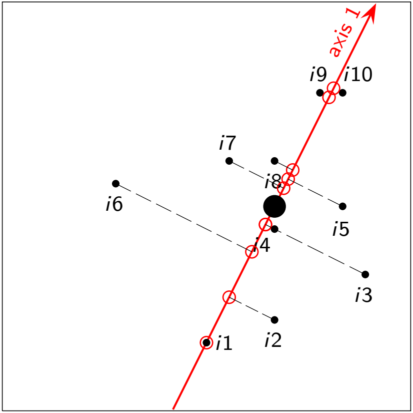
The line whose the variance of the projected cloud is maximum is called *first principal line*.

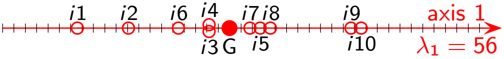
directed line → *1st principal axis*

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

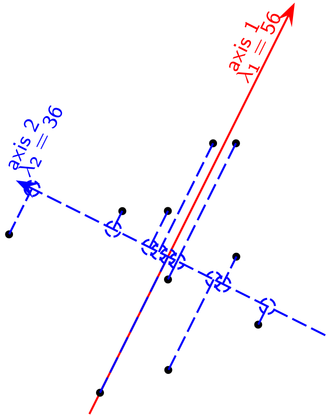
The first principal cloud is *the best fitting* of the initial cloud by an unidimensional cloud in the sense of *orthogonal least squares*

Here, angle = 63°, $\lambda_1 = 56$.





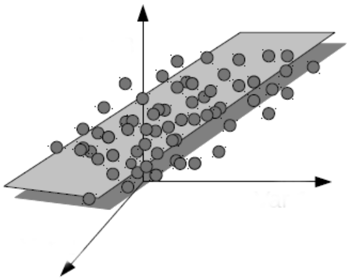
The residual cloud is constructed as the orthogonal projection of the cloud on the subspace orthogonal to the first principal line.



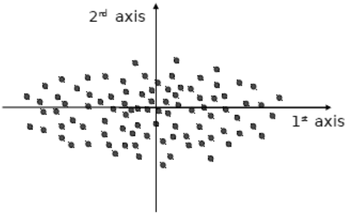
The first principal line of the residual cloud defines the *second principal line* of the initial cloud.

II.4. From Plane Cloud to High Dimensional Cloud

High dimensional cloud.



Low dimensional projection.



Heredity property

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

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.

they have mean $= 0$ and variance $= \lambda$ (eigenvalue)
they are *uncorrelated* (for distinct eigenvalues).

Aids to Interpretation

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

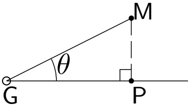
$$\frac{\lambda}{V_{\text{cloud}}}$$

- *Contribution of point to axis*:

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

- *Quality of representation of point onto axis*:

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

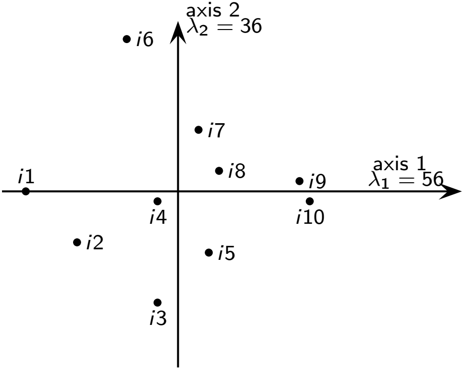


Results of the Analysis

$\lambda_1 = 56$ (variance of axis 1, eigenvalue), $\lambda_2 = 36$.

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

Principal representation of the cloud.



Results for axis 1 ($\lambda_1 = 56$)

	p_i	Coordinates	Ctr (%)	squared cosines
$i1$	0.1	-13.41	32.1	1.00
$i2$	0.1	-8.94	14.3	0.80
$i3$	0.1	-1.79	0.6	0.03
$i4$	0.1	-1.79	1.3	0.80
$i5$	0.1	$+2.68$	3.6	0.20
$i6$	0.1	-4.47	3.6	0.10
$i7$	0.1	$+1.79$	0.6	0.10
$i8$	0.1	$+3.58$	2.3	0.80
$i9$	0.1	$+10.73$	20.6	0.99
$i10$	0.1	$+11.63$	24.1	0.99

Results for axis 2 ($\lambda_2 = 36$)

Coordinates	Ctr (%)	squared cosines
0.00	0	0.00
$+4.47$	5.6	0.20
$+9.84$	26.9	0.97
$+0.89$	0.2	0.20
$+5.37$	8	0.80
-13.42	50.0	0.90
-5.37	8	0.90
-1.79	0.9	0.20
-0.89	0.2	0.01
$+0.89$	0.2	0.01

- Reconstitution of distances between points:

$$(\mathrm{M}^{i1}\mathrm{M}^{i2})^2 = (-13.41 + 8.94)^2 + (0 - 4.47)^2 = 4.23 = (6.3)^2$$

$$(\mathrm{GM}^{i2})^2 = (-8.94)^2 + (-4.47)^2 = 100$$

- Quality of representation of point M^{i2}: $\cos^2\theta = \frac{(-8.94)^2}{100} = 0.80$

III — Multiple Correspondence Analysis (MCA)

This text is adapted from Chapter 3 of the monograph

Multiple Correspondence Analysis

(QASS series n$^\circ$163, SAGE, 2010)

III.1. Introduction

Language of questionnaire

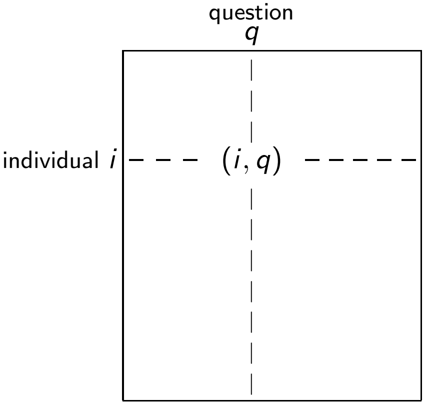
Basic data set: Individuals$\times$Questions table

- 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.
→ otherwise: preliminary phase of *coding*

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



question q

individual i

(i, q)

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-a. Cloud of Individuals

Distance between 2 individuals due to question q:

- if q is an agreement question:
  i and i' choose the same category
  $\rightsquigarrow$ the distance due to question q is null
  $$d_q = 0$$
- — if q is a disagreement question:
  i chooses category k and i' chooses category k'
  $\rightsquigarrow$ the squared distance due to question q is
  $$d_q^2 = \frac{1}{f_k} + \frac{1}{f_{k'}}$$

The squared overall distance is the mean of the squared distances due to active questions

$$d^2 = \sum d_q^2 / Q$$

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

G: mean point (center) of the cloud

Distance of an individual to the center of the cloud

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

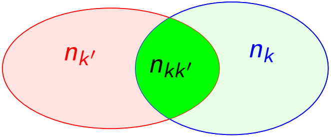
Variance of the cloud of individuals

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

(average number of categories per question minus 1).

III.4-b. Cloud of Categories

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



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

G: mean point (center) of the cloud

Property

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

Distance of a category–point to the center of the cloud

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

Variance of the cloud of individuals

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

Contributions

Contribution of *category k*

$$\mathrm{Ctr}_k = \frac{1-f_k}{K-Q}$$

Contribution of *question q*

$$\mathrm{Ctr}_q = \frac{K_q-1}{K-Q}$$

III.5. Principal Clouds

— *Principal axes*

Fundamental properties

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

— *Variance rates and modified rates*

Variance rate:

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

Modified rates $= \frac{(\lambda - \overline{\lambda})^2}{\sum(\lambda - \overline{\lambda})^2}$ (the sum is over λ such that $\lambda \geq \overline{\lambda}$)

— *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:

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

Variance of principal variable ℓ is equal to λ_ℓ:

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

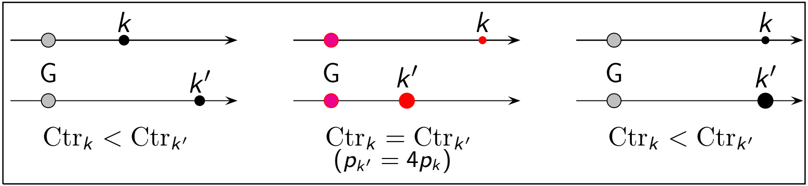
Principal variables are pairwise uncorrelated:

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

III.6. Aids to Interpretation: Contributions

Contribution of category–point k to axis ℓ: $\dfrac{p\, y^2}{\lambda}$

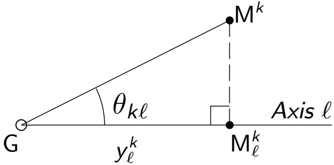
(y: coordinate of point on axis; p: relative weight; λ: variance of axis)


k
G
k′
$\mathrm{Ctr}_k < \mathrm{Ctr}_{k'}$
k
G
k′
$\mathrm{Ctr}_k = \mathrm{Ctr}_{k'}$
$(p_{k'} = 4p_k)$
k
G
k′
$\mathrm{Ctr}_k < \mathrm{Ctr}_{k'}$

By grouping, contributions add up ⟶ contribution of question...

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

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



III.7. 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: $\frac{6.25}{25} = 0.25$.

Axes whose variances exceed the mean.

Axes	1	2	3	4	5	6	7	8	9	10	11	12
variances (λ)	.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

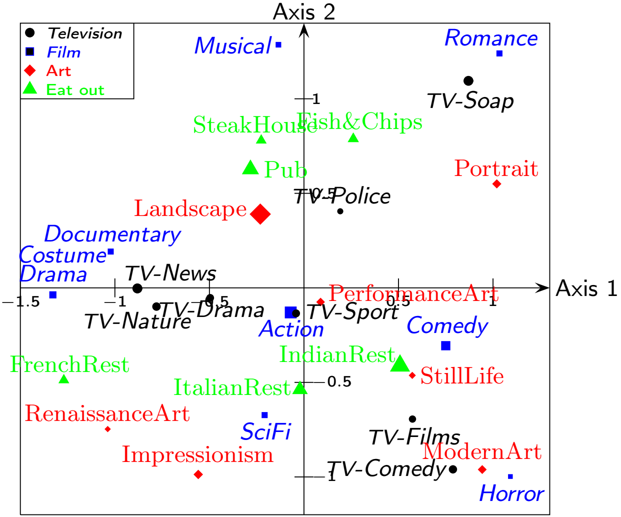
Principal coordinates and contributions (in %) of 6 individuals

	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

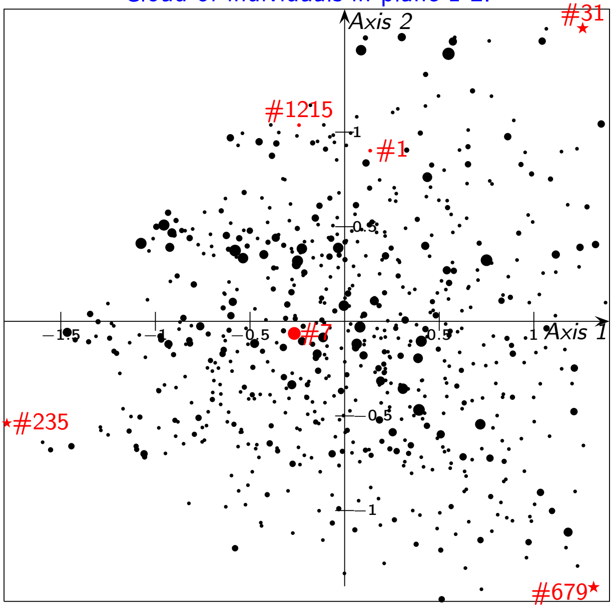
Relative weight, principal coordinates and contributions (in %) of categories

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



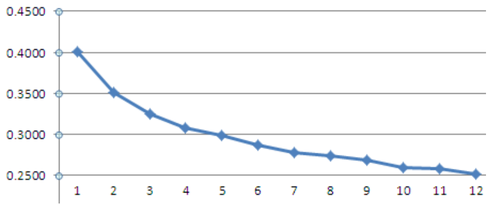
Cloud of individuals in plane 1-2.



III.8.Interpretation of Axes

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 axes 1, 2 and 3 = 82%

Guide for interpreting an axis

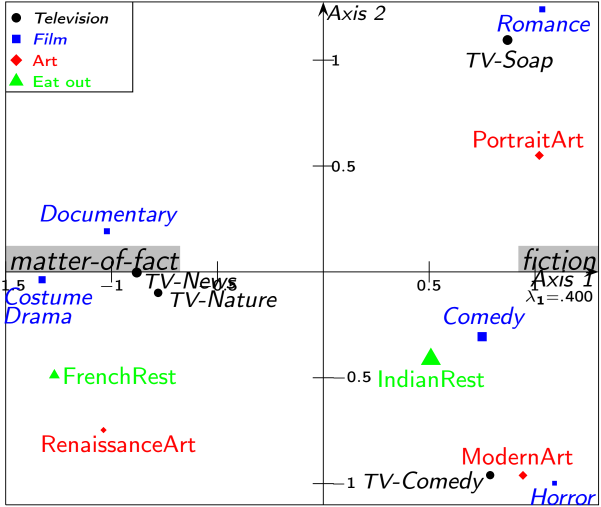
> *Interpreting an axis amounts to finding out what is similar, on the one hand, between all the elements figuring on the right of the origin and, on the other hand between all that is written on the left; and expressing with conciseness and precision, the contrast (or opposition) between the two extremes.*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

III.9. Transition formulas

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

— Category mean points

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

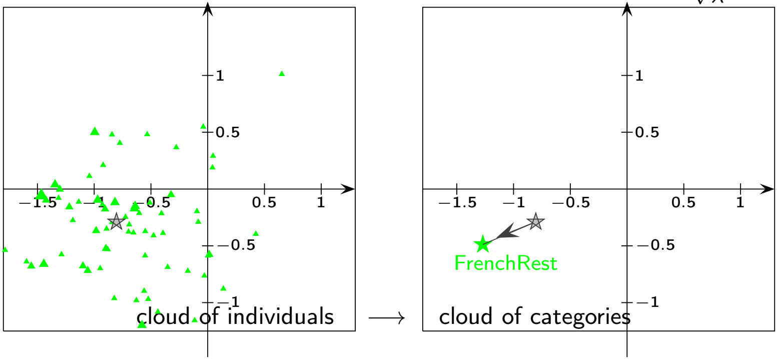
$$\overline{y}^k_\ell = \sqrt{\lambda_\ell}\, y^k_\ell \qquad \textit{(second transition formula)}$$

The K category mean points of question q define the
between–q cloud

.

• First transition formula

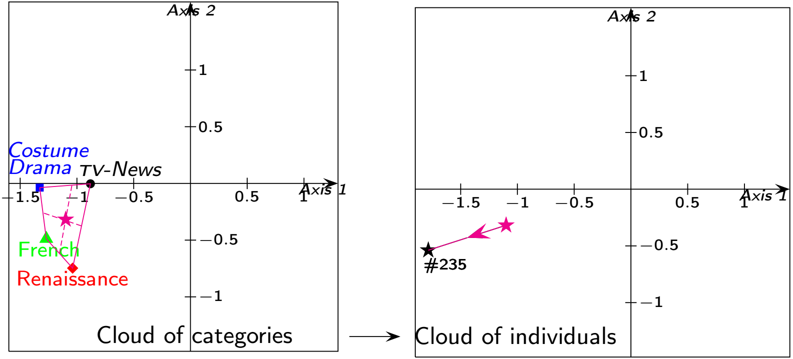
category mean point ($\overline{y}^k$) $\longrightarrow$ category point ($y^k = \frac{1}{\sqrt{\lambda}}\overline{y}^k$)



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.

- Second transition formula

mean for individual i $(\overline{y}^i = \sum_{k \in K_i} y^k/Q) \longrightarrow$ individual point $y^i = \frac{1}{\sqrt{\lambda}}\overline{y}^i$



Individual–point is located at the equibarycenter of the Q category–points of his response pattern, 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 $\sqrt{\lambda_1}$:

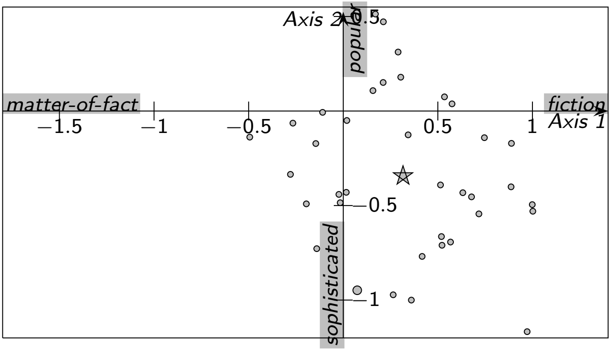
$$y_1^i = \frac{-1.12925}{\sqrt{0.4004}} = -1.785$$

   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 .

Supplementary individuals



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

LOCATE YOURSELF

III.10 Supplementary variables

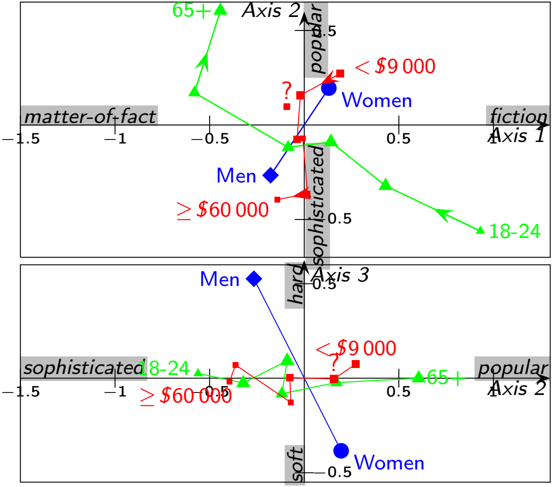
	weight	Axis 1	Axis 2	Axis 3
Men	513	−0.178	−0.266	+0.526
Women	702	+0.130	+0.195	−0.384
18-24	93	+0.931	−0.561	+0.025
25-34	248	+0.430	−0.322	−0.025
35-44	258	+0.141	−0.090	+0.092
45-54	191	−0.085	−0.118	−0.082
55-64	183	−0.580	+0.171	−0.023
≥ 65	242	−0.443	+0.605	+0.000

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).

— 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)

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.

III.11. Specific MCA

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

- *Junk categories*: categories of *no-interest*
  not representable by a single point
- *Infrequent categories*
  — remote from the center
  — contributing too much to the variance of the question
  — too influential on the determination of axes

Cloud of individuals

If for active question q,

- both i and i' choose active categories k and k': the distance is unchanged:

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

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

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

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.

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

number of active categories (K') minus number of questions without passive categories (Q').

- Specific overall variance:

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

- 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 $\lambda > \overline{\lambda}$).

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 $p_k = f_k/Q$) = λ

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 (transition formulas):

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

III.12-a. Concentration Ellipses

geometric summary of a subcloud in a principal plane

v_1 = variance of the coordinates of the subcloud on axis 1.

v_2 = variance of the coordinates of the subcloud on axis 2.

c = covariance between the 2 sets of coordinates.

$$\frac{v_2(y_1 - m_1)^2 - 2c(y_1 - m_1)(y_2 - m_2) + v_1(y_2 - m_2)^2}{v_1 v_2 - c^2} = \kappa^2$$

Properties

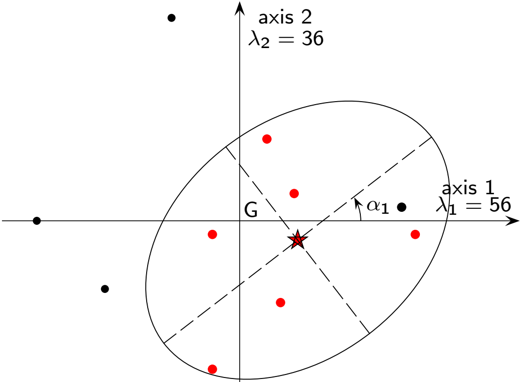
The concentration ellipse[*] of a subcloud is such that the half–axis of the ellipse is along the principal direction of the subcloud projected in the plane under study and its length is equal to $2\sqrt{\lambda'}$.

- A uniform distribution over the interior of the ellipse has the *same variance* as the subcloud.
- For a normally–shaped cloud, the concentration ellipse contains about *86% of the points* of the cloud.

Concentration ellipses are especially useful for studying families of subclouds induced by a structuring factor or a clustering procedure.

[*] see Cramér, 1946, p. 284; Le Roux & Rouanet (2010), p.69-T0

From the principal coordinates of the cloud of 10 points coordinates of the mean point C of the subcloud $\mathcal{C}$ ($m_1 = +3.8333$, $m_2 = -1.2778$), variances : $v_1 = 25.30612$, $v_2 = 21.22449$),
covariance : $c = +7.75510$
Eigenvalues of the covariance matrix: $\gamma_1^2 = 5.59$ and $\gamma_2 = 3.90$;
$\tan \alpha_1 = \frac{\gamma_1^2 - v_1}{c} = 0.7709$, hence $\alpha_1 = 37.63°$.



III.13-b. 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*.

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 :

$$\text{Mean} = 0 \qquad \text{Var} = \text{specific eigenvalue}$$

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

$$\text{Mean} = 0 \qquad \text{Var} = \text{specific eigenvalue}$$