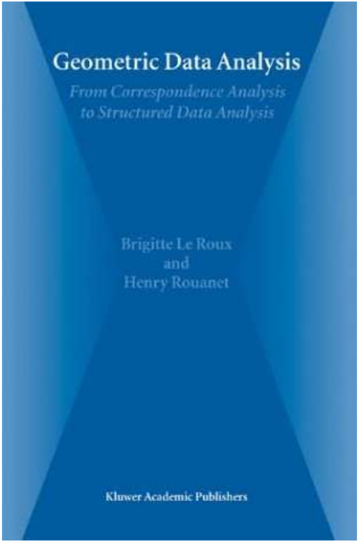
Course on GDA

Geometric Data Analysis

Brigitte Le Roux

November 6-10, 2023

What is Geometric Data Analysis (GDA)



Foreword

Geometric Data Analysis (GDA) is the name I have proposed to designate the approach to Multivariate Statistics initiated by Benzécri as Correspondence Analysis, an approach that has become more used and appreciated over the years.

PATRICK SUPPES

Stanford University

New book (to be published in 2024)
Geometric Data Analysis: Theory and Applications, Chapman & Hall (CRC Press)
by B. Le Roux, F. Cassor, F. Chanvril & J. Chiche.

I.1. Construction of clouds

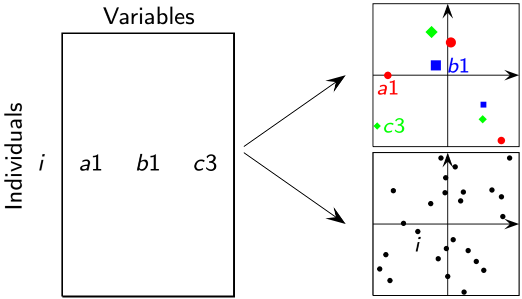
Euclidean clouds are constructed from

- Individuals×Variables tables
  - by *Principal Component Analysis* (PCA) if variables are numerical
  - by *Multiple Correspondence Analysis* (MCA) if variables are categorical
- Contingency tables by *Correspondence Analysis* (CA)
- Dissimilarity tables by *MultiDimensional Scaling* (MDS)

I.1.1. The Three Key Ideas of GDA

1. Geometric modeling

Data table *Clouds*



Cloud of categories: Points represent the categories of variables.

Cloud of individuals: Points represent individuals.

2. Formal approach.

Structures govern procedures!

3. Inductive philosophy

Descriptive analysis and geometric modelling comes prior to inductive analysis and probabilistic modelling

Priority is not exclusivity!

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

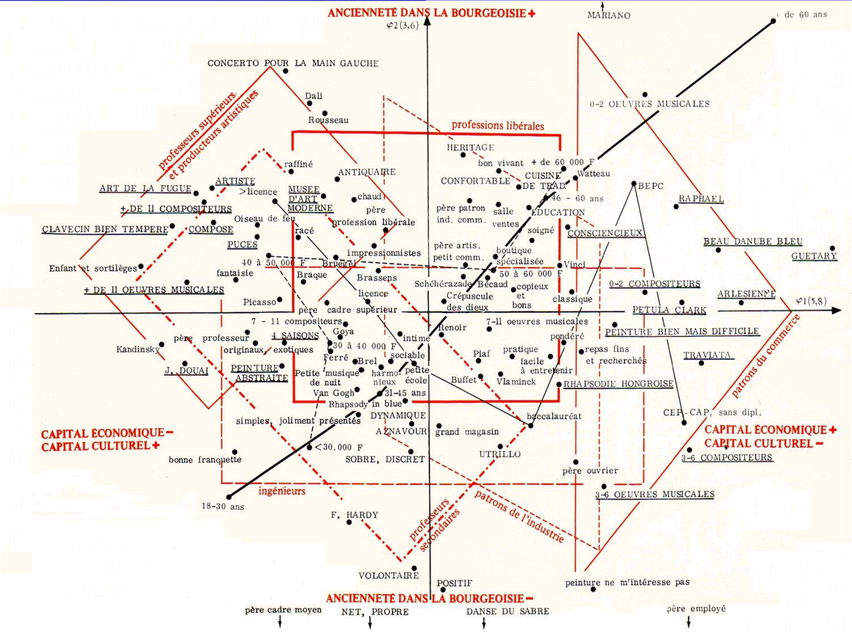
I.1.2. The Frame Model

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

- *Homogeneity*
  the topic of a study determines the fields wherefrom data are collected, but at times one has to take into account heterogeneous data collected at different levels, hence the preliminary phase of *data coding*;
- *Exhaustiveness*
  data should constitute an exhaustive or at least a *representative* inventory of the domain under study.

I.2. On Structured Data Analysis

- *supplementary variables*:
  it is the first step of structured data analysis,
  but it permits studying *mean points* but not the *dispersion* of subclouds
- *structuring factors*:
  we mean relevant variables describing the two basic sets that do not serve to construct the clouds.



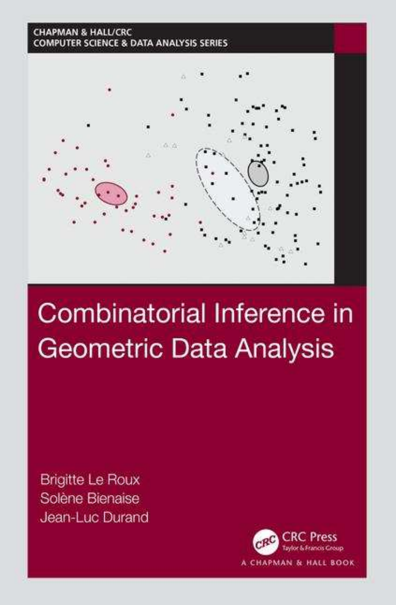
"La Distinction", Bourdieu, 1979

I.3. Inductive Data Analysis

Combinatorial Inference in GDA

Brigitte Le Roux
Solène Bienaise
Jean-Luc Durand

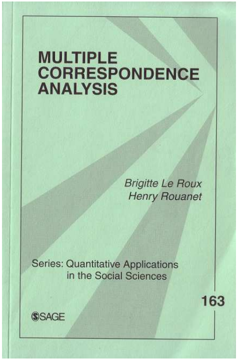
(CRC Press, 2019)

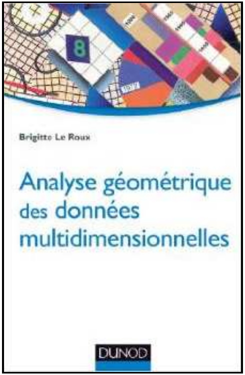


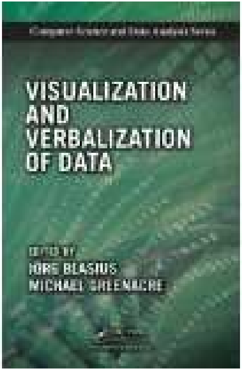
Description comes first and inference later.

Recent books

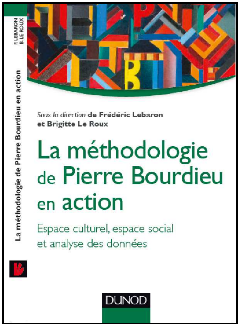
Le Roux & Rouanet 2010	Le Roux 2014	CARME 2011 (2015)

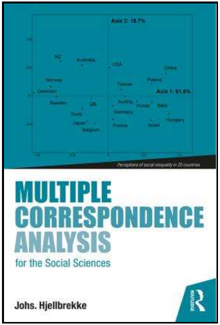


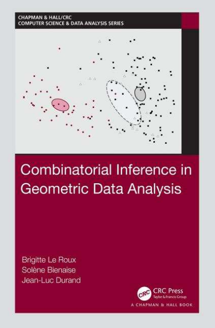




Lebaron, Le Roux (eds) 2015	Hjellbrekke 2017	Blasius, Lebaron,Le Roux, Schmitz (eds), 2019	Le Roux, Bienaise, Durand, 2019





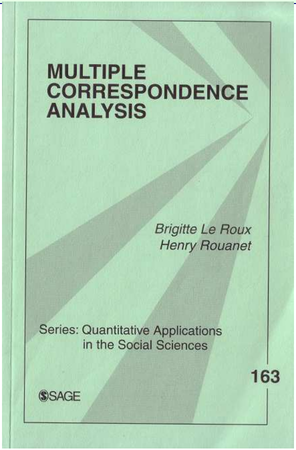


GDA, as a whole methodology, is now discovered by a large audience and largely used.

Basic Geometric Notions

Cloud of points and dimensionality reduction

This presentation is *adapted* from Chapter 2 of the monograph *Multiple Correspondence Analysis* (QASS series n°163, SAGE, 2010)



II.1. Basic Notions of Geometry

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

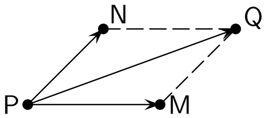
— *Affine notions*: alignment, direction and barycenter.

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

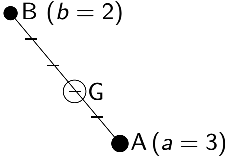
The *deviation* from point P to point M is M − P ("terminal minus initial"), that is, $\overrightarrow{\mathrm{PM}}$.

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

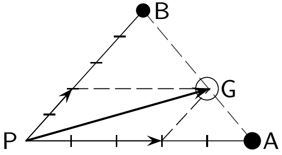
$$\overrightarrow{\mathrm{PM}} + \overrightarrow{\mathrm{PN}} = \overrightarrow{\mathrm{PQ}}$$



Barycenter of a dipole



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



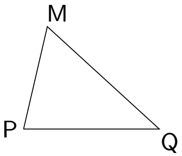
$$\overrightarrow{\mathrm{PG}} = \frac{3}{5}\,\overrightarrow{\mathrm{PA}} + \frac{2}{5}\,\overrightarrow{\mathrm{PB}}$$

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

— *Metric notions*: distances and angles.

Triangle inequality:

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

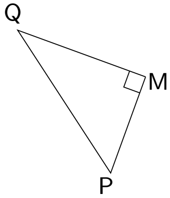


Pythagorean theorem:

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

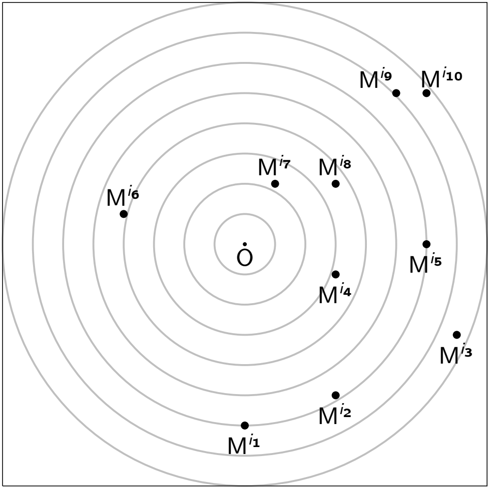


Figure 1. Target example (10 points)

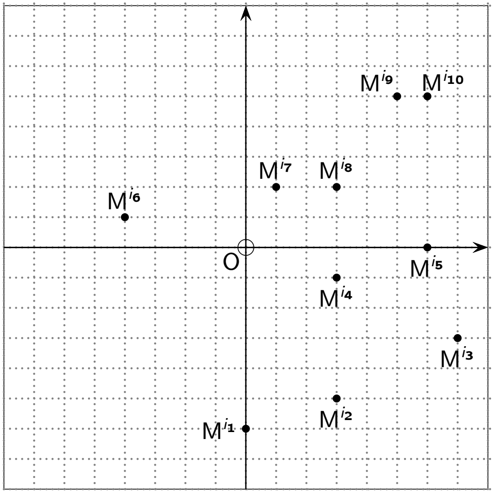


Figure 1b. Target area with two rectangular axes

Coordinates of points

	x^{I}	x'^{I}
M^{i_1}	0	−6
M^{i_2}	3	−5
M^{i_3}	7	−3
M^{i_4}	3	−1
M^{i_5}	6	0
M^{i_6}	−4	1
M^{i_7}	1	2
M^{i_8}	3	2
M^{i_9}	5	5
$M^{i_{10}}$	6	5

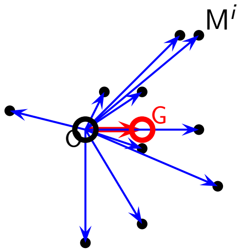
Mean Point

Cloud of points: point M^i with relative weight p_i

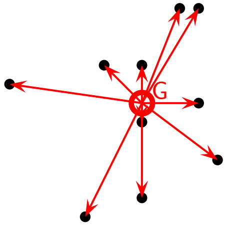
Mean point: point G

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

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



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



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

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:

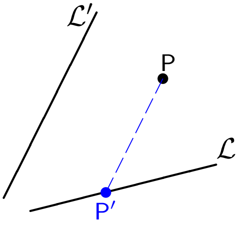
$$\mathrm{Ctr}_i = \frac{p_i(\mathrm{GM}^i)^2}{V_{\text{cloud}}}$$

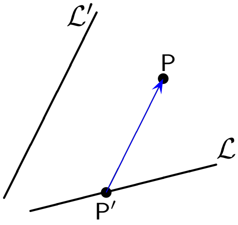
II.3. Principal Axes of a Cloud

Projection of a cloud

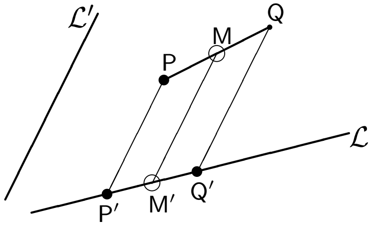
P′ = projection of point P onto $\mathcal{L}$ along $\mathcal{L}'$

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





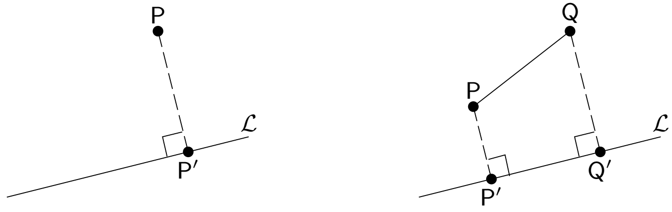
If point M is the midpoint of P and Q, the point M′, projection of M on $\mathcal{L}$, is the midpoint of P′ and Q′.



Mean point property

The mean point is preserved by projection.

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

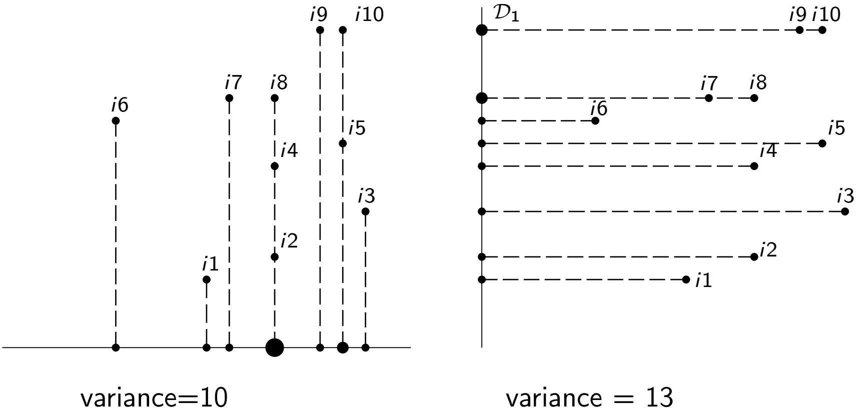


The orthogonal projection contracts distances: $\mathrm{P'Q'} \leq \mathrm{PQ}$, therefore one has the

Property

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

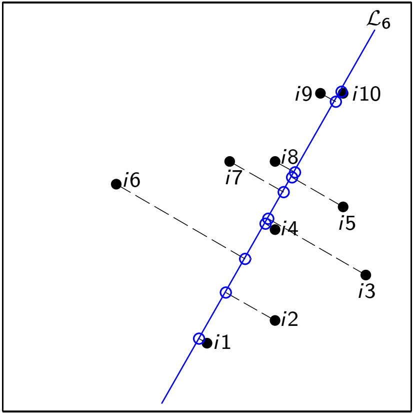
Projected clouds on several lines



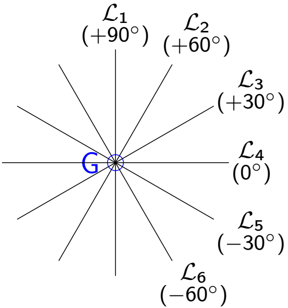
Orthogonal additive decomposition

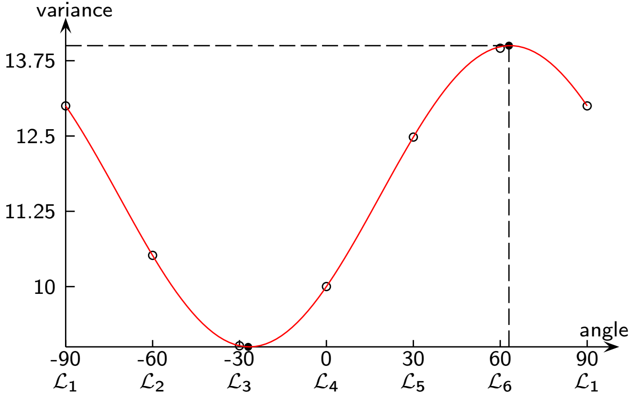
The variance of the cloud is the sum of the variances of projected clouds onto perpendicular lines: $V_{\text{cloud}} = 10 + 13 = 23$.

Projection onto an oblique line (60 degrees) : variance = 13.975



Essai





	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	$\mathcal{L}_4$	$\mathcal{L}_5$	$\mathcal{L}_6$	$\mathcal{L}_1$
Variance	13.	10.52	9.02	10.	12.48	13.98	13

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

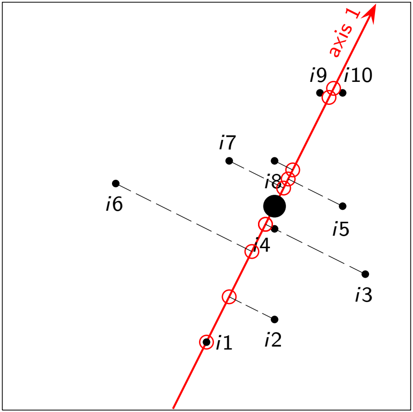
1st principal axis

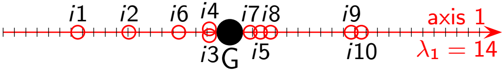
1st principal cloud

its variance (λ_1) = *variance of axis 1*

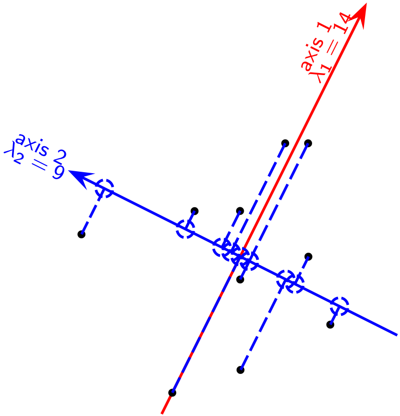
The first principal cloud is *the best fitting* of the cloud by a one-dimensional cloud in the sense of *orthogonal least squares*.

Here, angle $= 63^\circ$,
variance $= \lambda_1 = 14$.



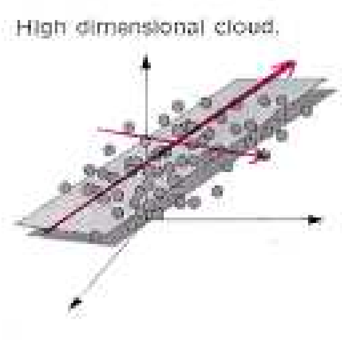


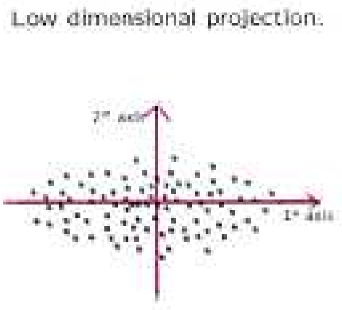
The residual cloud is constructed as the orthogonal projection of the cloud on the subspace orthogonal to the first principal axis.



The first principal axis of the residual cloud defines the *second principal axis* of the cloud.

II.4. From Plane Cloud to High Dimensional Cloud





Heredity property

The plane that best fits the cloud is the one determined by the first two principal 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, with

  mean = 0
  variance = λ (eigenvalue)
  uncorrelated for distinct eigenvalues

Aids to Interpretation

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

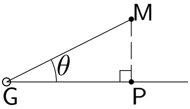
$$\frac{\lambda}{V_{\text{cloud}}}$$

- *Contribution of point to axis*:

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

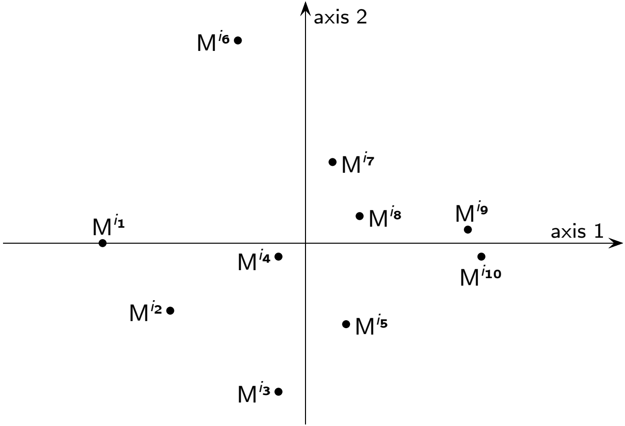
- *Quality of representation of point onto axis*:

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



Results of the Analysis

1. *Variances of axes* (eigenvalues): $\lambda_1 = 14$, $\lambda_2 = 9$.
   Variance rate : $\dfrac{\lambda_1}{V_{\text{cloud}}} = \dfrac{14}{23} = 61\%$
2. *Principal representation of the cloud.*



③ *Principal coordinates*

	weights	Axis 1	Axis 2
$i1$	0.1	−6.71	0.00
$i2$	0.1	−4.47	−2.24
$i3$	0.1	−0.89	−4.92
$i4$	0.1	−0.89	−0.45
$i5$	0.1	1.34	−2.68
$i6$	0.1	−2.24	6.71
$i7$	0.1	0.89	2.68
$i8$	0.1	1.79	0.89
$i9$	0.1	5.37	0.45
$i10$	0.1	5.81	−0.45

④ *Contributions*

	Contributions (in %) to		
	cloud	axis 1	axis 2
$i1$	19.61	32.1	0.0
$i2$	10.91	14.3	5.6
$i3$	10.91	0.6	26.9
$i4$	0.41	0.6	0.2
$i5$	3.91	1.3	8.0
$i6$	21.71	3.6	50.0
$i7$	3.51	0.6	8.0
$i8$	1.71	2.3	0.9
$i9$	12.61	20.6	0.2
$i10$	14.81	24.1	0.2

	Quality of representation onto		
	plane 1-2	axis 1	axis 2
$i1$	1.000	1.000	0.000
$i2$	1.000	0.800	0.200
$i3$	1.000	0.032	0.968
$i4$	1.000	0.800	0.200
$i5$	1.000	0.200	0.800
$i6$	1.000	0.100	0.900
$i7$	1.000	0.100	0.900
$i8$	1.000	0.800	0.200
$i9$	1.000	0.993	0.007
$i10$	1.000	0.994	0.006

Benzécri, J.-P. (1992).
Correspondence Analysis Handbook.
Dekker: New York.
(adapted from J.-P. & F. Benzécri, Paris: Dunod, 1984).

Benzécri, J.-P. & *coll*. (1973).
L'Analyse des Données. 1 Taxinomie, 2 L'analyse des correspondances.
Paris: Dunod.

Blasius, J. and Greenacre, M. (2014).
Visualization and verbalization of data.
CRC Press.

Blasius, J., Lebaron, F., Le Roux, B., and Schmitz, A. (2020).
Empirical investigations of social space, volume 15.
Springer.

Bourdieu, P. (1979).
La distinction. Critique sociale du jugement.
Paris: Minuit.

Hjellbrekke, J. (2018).
Multiple correspondence analysis for the social sciences.
Routledge.

Le Roux, B. (2014).
Analyse géométrique des données multidimensionnelles.
Paris: Dunod.

Le Roux, B., Bienaise, S., and Durand, J.-L. (2019).
Combinatorial Inference in Geometric Data Analysis.
CRC Press.

Le Roux, B. and Rouanet, H. (2004).
Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis (Foreword by P. Suppes).
Dordrecht: Kluwer.

Le Roux, B. and Rouanet, H. (2010).
Multiple Correspondence Analysis, 163.
QASS. Thousand Oaks (CA): Sage Publications.

Lebaron, F. (2006).
L'enquête quantitative en sciences sociales. Recueil et analyse de données.
Paris: Dunod.