Course on GDA

Geometric Data Analysis

Brigitte Le Roux

MATHÉMATIQUES ET INFORMATIQUE
Sciences
Université Paris Cité

SciencesPo. CEVIPOF CNRS

November 20–22, 2023

UPPSALA
UNIVERSITET

Multiple Correspondence Analysis (MCA)

This text is adapted from Chapter 3 of the monograph

Multiple Correspondence Analysis

(QASS series n°163, SAGE, 2010)

I.1. Introduction

Language of questionnaire

Basic data set: Individuals×Questions table

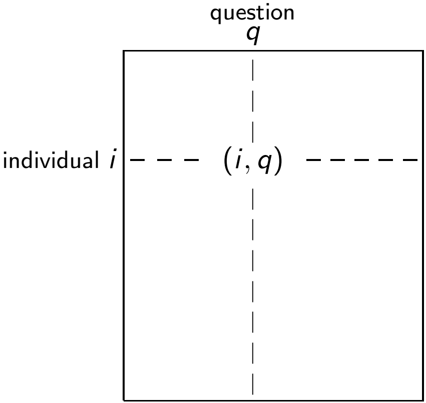
- Questions = categorical variables, i.e. variables with a (finite) number of *response categories* (or *modalities*).
- Individuals or "statistical individuals": (people, firms, items, etc.).

"*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



MCA produces two clouds of points:

the *cloud of individuals* and the *cloud of categories*.

I.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

The data source is the ESRC project "Cultural Capital and Social Exclusion: A Critical Investigation". The data were collected in 2003–2004. The research team (Open University and Manchester University, UK) included T. Bennett, M. Savage, E. Silva, A. Warde, D. Wright and M. Gayo-Cal. Details of the survey can be found in *Culture, Class, Distinction*, 2009, Bennett & al, (p. 262).

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+burger barcafe+cafe or 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

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

The original sample size was 1564 (stratified, clustered random sample from 111 postcode sectors), with in addition several groups of people belonging to minority ethnic groups in Britain (Indians, Pakistani, Afro-Carribbean).

$K = 8 + 8 + 7 + 6 = 29$ categories
$n = 1215$ individuals with response pattern without "other" category.
$8 \times 8 \times 7 \times 6 = 2688$ possible response patterns, only 658 are observed.

I.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' (other than 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

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

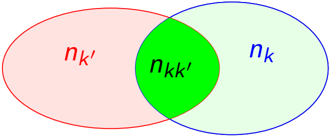
Variance of the cloud of individuals

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

(average number of categories per question minus 1).

I.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.

Squared distance of a category–point to the center of the cloud

$$\frac{1}{f_k} - 1$$

Variance of the cloud of categories

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

Contributions

Contribution of *category k*

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

Contribution of *question q*

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

I.5. Principal Clouds

— *Principal axes*

Fundamental properties

- The variances of principal axes (eigenvalues) of the 2 clouds are equal.
- $\sum \lambda = V_{\text{cloud}}$, with $\overline{\lambda} = \dfrac{V_{\text{cloud}}}{L} = \dfrac{1}{Q}$.

— *Variance rates* and *modified rates* (importance index)

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 ℓ 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:

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

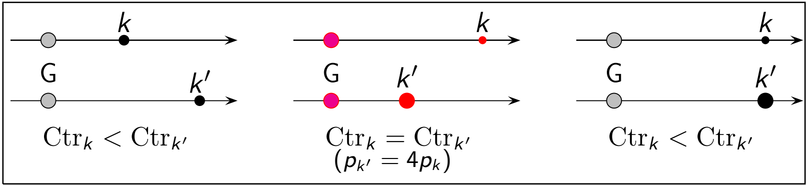
Principal variables ℓ and ℓ' ($\ell \neq \ell'$) are pairwise uncorrelated:

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

I.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)

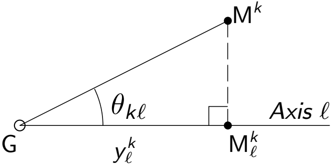


k
G
k'
$\mathrm{Ctr}_k < \mathrm{Ctr}_{k'}$
k
G
k'
$\mathrm{Ctr}_k = \mathrm{Ctr}_{k'}$
$(p_{k'} = 4p_k)$
k
G
k'
$\mathrm{Ctr}_k < \mathrm{Ctr}_{k'}$

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

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

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



I.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

Dimensionality of the cloud $\leq K - Q = 29 - 4 = 25$

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

Contributions of questions to the overall variance:

$$\frac{8-1}{29-4} = 28\% \qquad \frac{8-1}{29-4} = 28\% \qquad \frac{7-1}{29-4} = 24\% \qquad \frac{6-1}{29-4} = 20\%$$

Elementary statistical results

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

n_k: absolute frequency, f_k: relative frequency (in %), Ctr$_k$: contribution to cloud (in %)

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

Recall that modified rates *are not* rates of variance

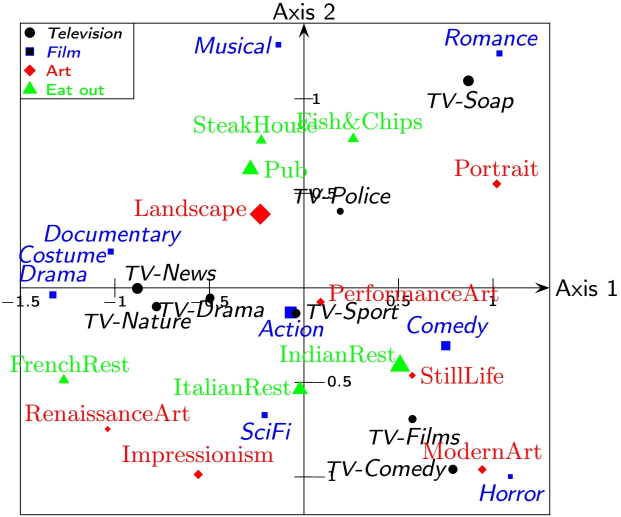
Principal coordinates and contributions of 6 individuals

	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

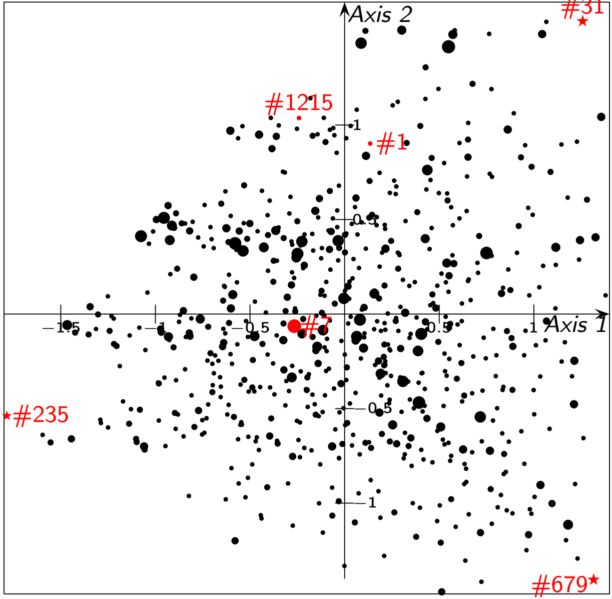
Relative weight, principal coordinates and contributions (in %) of categories

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.

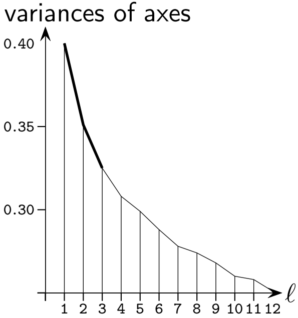


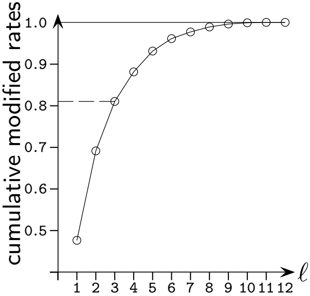
I.8.Interpretation of axes

How many axes need to be interpreted?

Variances;cumulative variance rates (in %); cumulative modified rates (%)

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	6.4	12.0	17.2	22.2	26.9	31.5	36.0	40.4	44.7	48.8	53.0	57.0
modified rates	47.6	69.1	81.0	88.1	93.1	96.1	97.7	98.9	99.6	99.9	100.0	100.0





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 $=0.70$.
Axis 3: $(\frac{\lambda_3-\lambda_4}{\lambda_3}=.05)$; modified rate $=0.12$.

After axis 4, variances decrease regularly and the differences are small.
Cumulated modified rate for axes 1, 2 and 3 : $0.48+0.22+0.12=82\%$

Guide for interpreting an axis

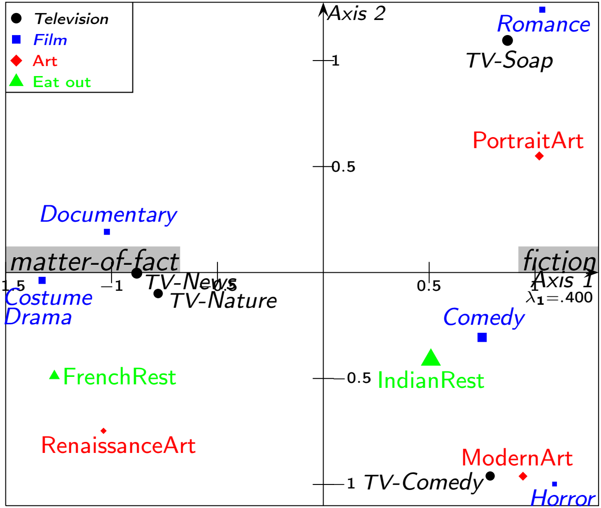
Interpreting an axis amounts to finding out what is similar, on the one hand, between all the elements figuring on the right of the origin and, on the other hand between all that is written on the left; and expressing with conciseness and precision, the contrast (or opposition) between the two extremes.

Benzécri (1992, p. 405)

For interpreting an axis, we use the method of contributions of points and deviations
(Le Roux & Rouanet (1998), Interpreting Axes in Multiple Correspondence Analysis:
`http://helios.mi.parisdescartes.fr/~lerb/publications/Int_Axes.pdf`)
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 ;

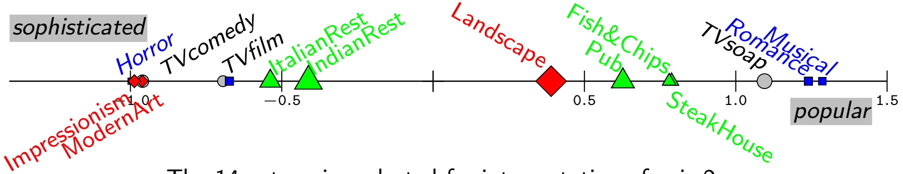
sum of contributions of retained categories = 89% → *good summary*

Interpretation of axis 2

	left	right	*deviation*
● *TV* (30.7)			
TVsoap		15.1	
TVcomedy	8.2		*26.3*
TVfilm	3.3		
■ *Film* (25.7)			
Romance		9.1	
Musical		8.4	*13.9*
Horror	3.6		

Total contribution: 32.8 + 52.1 = 84.9

	left	right	*deviation*
◊ *Art* (23.5)			
Impressionism	7.1		
Modern	5.9		*18.7*
Landscape		5.6	
▲ *Eat out (23.1)*			
green Pub		6.5	
IndianRest	4.0		
ItalianRest	3.9		*21.3*
Fish&Chips		3.9	
SteakHouse		3.5	



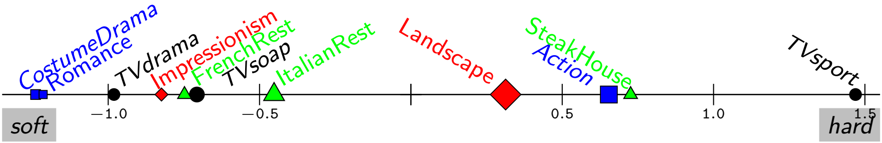
The 14 categories selected for interpretation of axis 2.

Interpretation of axis 3

	left	right	*deviation*
● *TV* (38.4)			
tvsport	18.6		
tvdrama		8.2	*32.2*
tvsoap		6.8	
◇ *Art* (11.2)			
Impressionism		5.4	
Landscape	3.9		*8.5*

Total contribution; 36.3 + 49.9 = 86.2

	left	right	*deviation*
■ *Film* (39.5)			
CostumeDrama		13.6	
Action	10.5		*33.4*
Romance		9.4	
▲ *Eat out* (10.9%)			
FrenchRest		3.5	
SteakHouse	3.3		*7.7*
ItalianRest		2.9	



The 11 categories selected for the interpretation of axis 3.

- 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*.

I.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

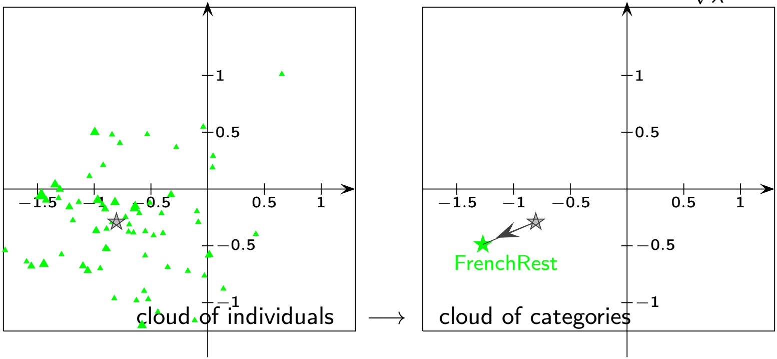
$$\overline{y}_\ell^k = \sqrt{\lambda_\ell}\, y_\ell^k \qquad \textit{(second transition formula)}$$

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

.

• First transition formula

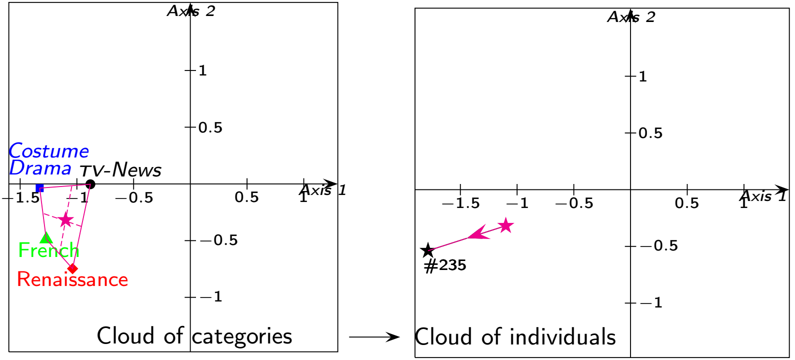
category mean point ($\overline{y}^k$) $\longrightarrow$ category point ($y^k = \frac{1}{\sqrt{\lambda}}\overline{y}^k$)



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 $\sqrt{\lambda_1}$:

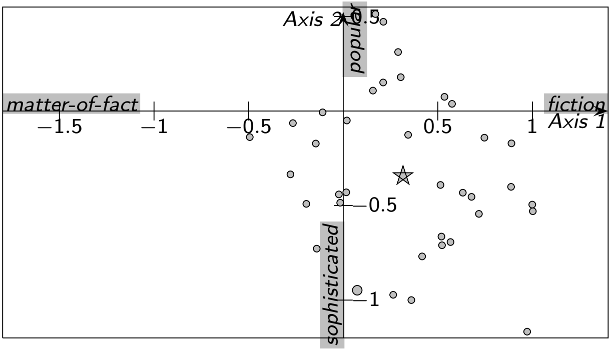
$$y_1^i = \frac{-1.12925}{\sqrt{0.4004}} = -1.785$$

   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 (⋆).

I.10 Supplementary variables

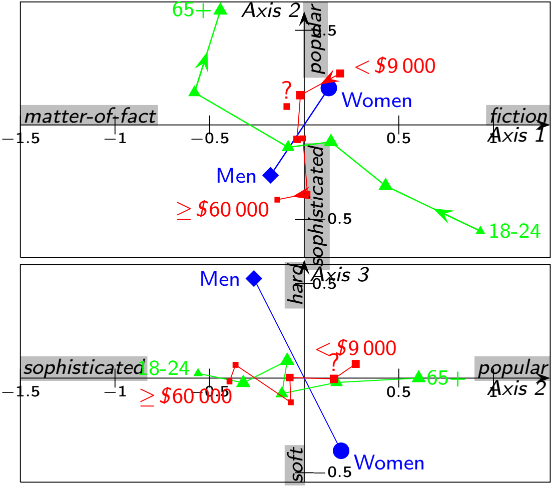
	weight	Axis 1	Axis 2	Axis 3
Men	513	−0.178	−0.266	+0.526
Women	702	+0.130	+0.195	−0.384
18-24	93	+0.931	−0.561	+0.025
25-34	248	+0.430	−0.322	−0.025
35-44	258	+0.141	−0.090	+0.092
45-54	191	−0.085	−0.118	−0.082
55-64	183	−0.580	+0.171	−0.023
≥ 65	242	−0.443	+0.605	+0.000

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 0.4 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).

I.10. Subclouds and Concentration Ellipses

Geometric summary of a subcloud in a principal plane is given by its concentration ellipse.

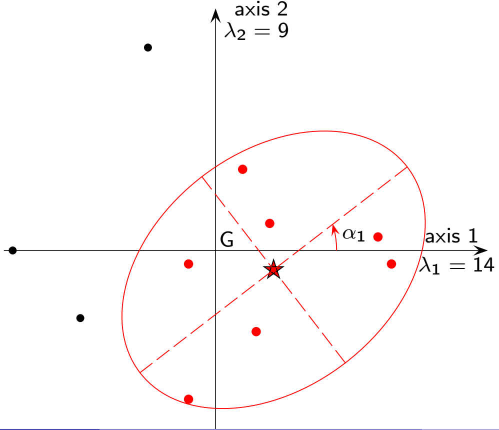
Properties

- The concentration ellipse* of a subcloud is such that the half–axis of the ellipse is along the principal line of the subcloud projected in the plane under study and its length is equal to 2 times the standard deviation of the subcloud along the principal line.
- 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 $\mathcal{C}$ ($m_1 = +1.917$, $m_2 = -0.639$), variances : $v_1 = 6.327$, $v_2 = 5.306$),
covariance : $c = +1.939$
Eigenvalues of the covariance matrix: $\gamma_1^2 = 7.821$ and $\gamma_2 = 3.812$;
$\tan \alpha_1 = \frac{\gamma_1^2 - v_1}{c}$ $\alpha_1 = 37.63$.



I.11. Specific MCA

Specific MCA (SpeMCA) consists in restricting the analysis to *categories of interest*.

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 of the cloud

  — contributing too much to the variance of the question

  — too influential on the determination of axes

References

Escofier, B. (1987) Traitement des questionnaires avec non-réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte,*Pub. Inst. Stat.Univ. Paris*, XXXII, fasc. 3, p. 33-69.

Le Roux, B. (1999) Analyse spécifique d'un nuage euclidien : application à l'étude des questionnaires, *Mathématiques et sciences humaines*, 146, 65–83.

The two Clouds

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: projection of the cloud onto a subspace of interest.

Cloud of categories

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

Property of Clouds

- Dimension: $K' - Q'$

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

- Specific overall variance:

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

$(\sum_{k \in K'} p_k = \text{sum of relative weights of active categories}, < 1)$

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 (transition formulas):

$$\overline{y} = \sqrt{\lambda}\, y \qquad \begin{array}{l}(y\text{: principal coordinate of category } k \\ \ \overline{y}\text{: principal coordinate of category mean–point } k)\end{array}$$

I.13. Class Specific Analysis (CSA)

CSA consists in analyzing a *subset of individuals* by taking the whole set of individuals as a reference.

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*.

Reference

Le Roux B., Rouanet H. (2004) *Geometric Data Analysis*, Kluwer

CSA: The Clouds

Cloud of individuals

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

Cloud of categories

The distance between 2 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

Methodology of MCA

- selecting active and supplementary individuals and their weights
- Selecting active variables, supplementary variables and structuring factors
- Coding of data
  Missing data
  "Junk" categories
  Rare categories
- Interpretation strategy
  1. Examination of clouds
  2. How many axes?
  3. Interpretation of axes

     Step 1 Important variables
     Step 2 Important categories
     Step 3 Landmarks response patterns
     Step 4 Geometric summary

Methodology of MCA

- Supplementary individuals
- Supplementary variables
- Joint use of MCA and Clustering

Some references I

Bennett, T., Savage, M., Silva, E., Warde, A., Gayo-Cal, M., and Wright, D. (2009).
Culture, class, distinction.
Routledge.

Benzécri, J.-P. (1977).
Sur l'analyse des tableaux binaires associés à une correspondance multiple.
Les cahiers de l'analyse des données, 2(1):55–71.
D'après un texte ronéotypé de 1972.

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

Some references II

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.

Börjesson, M., Broady, D., Le Roux, B., Lidegran, I., and Palme, M. (2016).
Cultural capital in the elite subfield of swedish higher education.
Poetics, 56:15–34.

Some references III

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

Bourdieu, P. (1984).
Homo academicus.
Paris: Minuit.

Bourdieu, P. (1989).
La noblesse d'État. Grandes écoles et esprit de corps.
Paris: Minuit.

Bourdieu, P. (1999).
Une révolution conservatrice dans l'édition.
Actes de la recherche en sciences sociales, 126–127:3–28.

Some references IV

Bourdieu, P. and de Saint-Martin, M. (1978).
Le patronat.
Actes de la recherche en sciences sociales, 20–21:3–82.

Chiche, J., Le Roux, B., Perrineau, P., and Rouanet, H. (2000).
L'espace politique des électeurs français à la fin des années 1990.
Revue française de sciences politiques, pages 463–487.

Denord, F., Lagneau-Ymonet, P., and Thine, S. (2011).
Le champ du pouvoir en france.
Actes de la recherche en sciences sociales, 190:24–57.

Ferguson, J. L. (2022).
The great refusal: The west, the rest, and the new regulations on homosexuality, 1970–2015.
American Journal of Sociology, 128(3):680–727.

Some references V

Gounelle, C. and Le Roux, B. (2007).
Etude de la structure d'un test de jugement pratique par l'analyse geometrique des donnees.
European review of applied psychology, 57(2):107–117.

Hayashi, C. (1956).
Theory and example of quantification (ii).
Proceedings of the Institute of Statistical Mathematics, 4:19–30.

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

Hjellbrekke, J., Le Roux, B., Korsnes, O., Lebaron, F., Rosenlund, L., and Rouanet, H. (2007).
The Norwegian Field of Power Anno 2000.
European Societies, 9(2):245–273.

Some references VI

Le Roux, B. (1999).
Analyse spécifique d'un nuage euclidien: application à l'étude des questionnaires.
Mathématiques et sciences humaines, 146:6–83.

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.

Le Roux, B., Rouanet, H., Savage, M., and Warde, A. (2008).
Class and cultural division in the uk.
Sociology, 42:1049–1071.

Some references VII

Lebaron, F. (2000).
La croyance économique.
Paris: Seuil.

Lebaron, F. and Bonnet, P. (2020).
Class-specific analysis: Methodological and sociological reflections.
In Blasius, J., Lebaron, F., Le Roux, B., and Schmitz, A., editors, *Empirical investigations of social space*, chapter 21, pages 359–376. San diego: Academic Press.

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

Some references VIII

Lebart, L. (1975).
L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples.
Consommation, 2:73–96.

Roose, H. (2015).
Signs of emerging cultural capital? analysing symbolic struggles using class specific analysis.
Sociology, 49(3):556–573.

Sapiro, G. (1996).
La raison littéraire: Le champ littéraire français sous l'occupation (1940–1944).
Actes de la recherche en sciences sociales, 111-112:3–35.

ETC.