III — Multiple Correspondence Analysis (MCA)

Brigitte Le Roux <Brigitte.LeRoux@mi.parisdescartes.fr>
Frédéric Lebaron <frederic.lebaron@u-picardie.fr>
Johannes Hjellbrekke <Johs.Hjellbrekke@sos.uib.no>

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

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*

3.1 Principles of MCA

Notations:

I: set of n individuals;

Q: set of questions

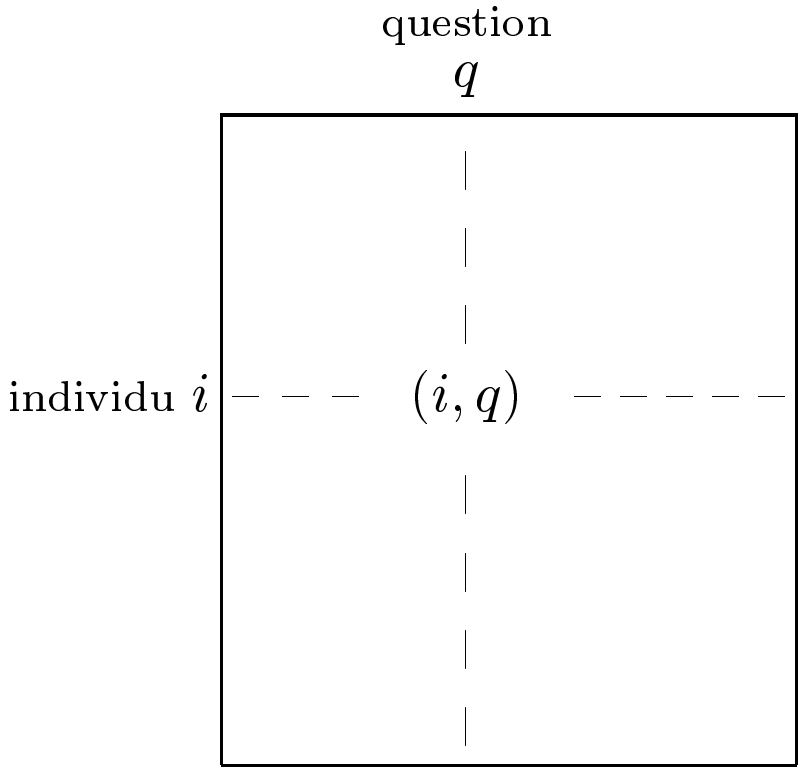
K_q: set of categories of question q ($K_q \geq 2$)

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.***

3.2 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

$8 \times 8 \times 8 \times 7 \times 6 = 2688$ possible response pattens, 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

3.3 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

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

Variance of the cloud: $V_{\text{cloud}} = \sum \frac{1}{n} (\mathrm{GM}^i)^2 = \frac{K}{Q} - 1$
(average number of categories per question minus 1).

3.4 Cloud of categories

Distance between 2 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
$p_k = f_k/Q = n_k/nQ$

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

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

- *Variance*: $= \sum p_k (\mathrm{GM}^k)^2 = \frac{K}{Q} - 1.$
- *Contributions*

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

3.5 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: $\tau_\ell = \dfrac{\lambda_\ell}{V_{\text{cloud}}}$

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

— Principal coordinates and principal variables

y_ℓ^i: coordinate of individual i on axis ℓ

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

y_ℓ^k: coordinate of category k on axis ℓ

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

Means of principal variables are null:

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

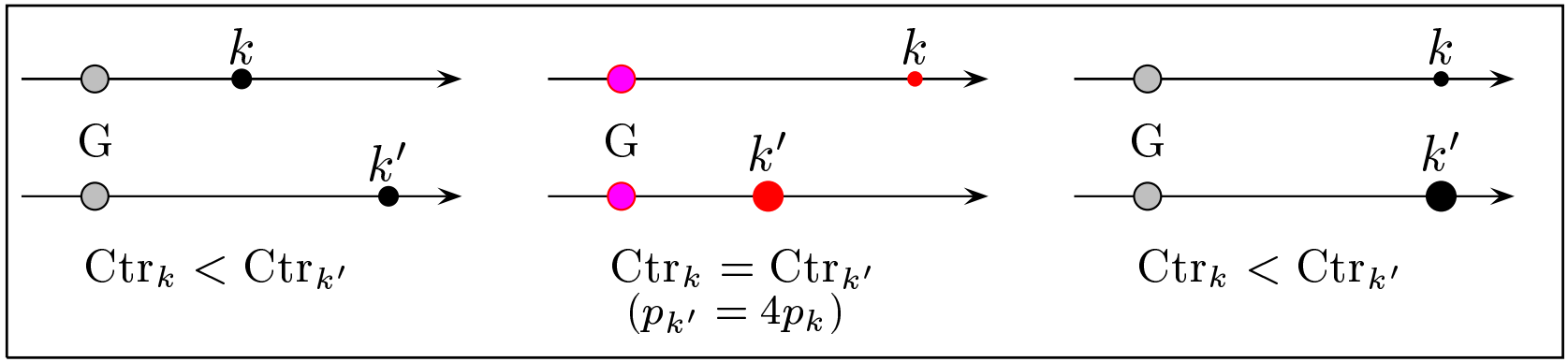
Variances of principal variables are equal to eigenvalues:

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

—Aids to interpretation: Contributions

Contribution of point M^k to axis: $\dfrac{p_k\,(y^k)^2}{\lambda}$

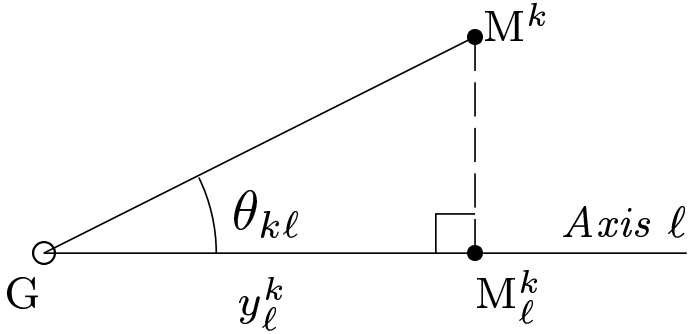
(y: coordinate of point on axis; p: relative weight; λ: eigenvalue)



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

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

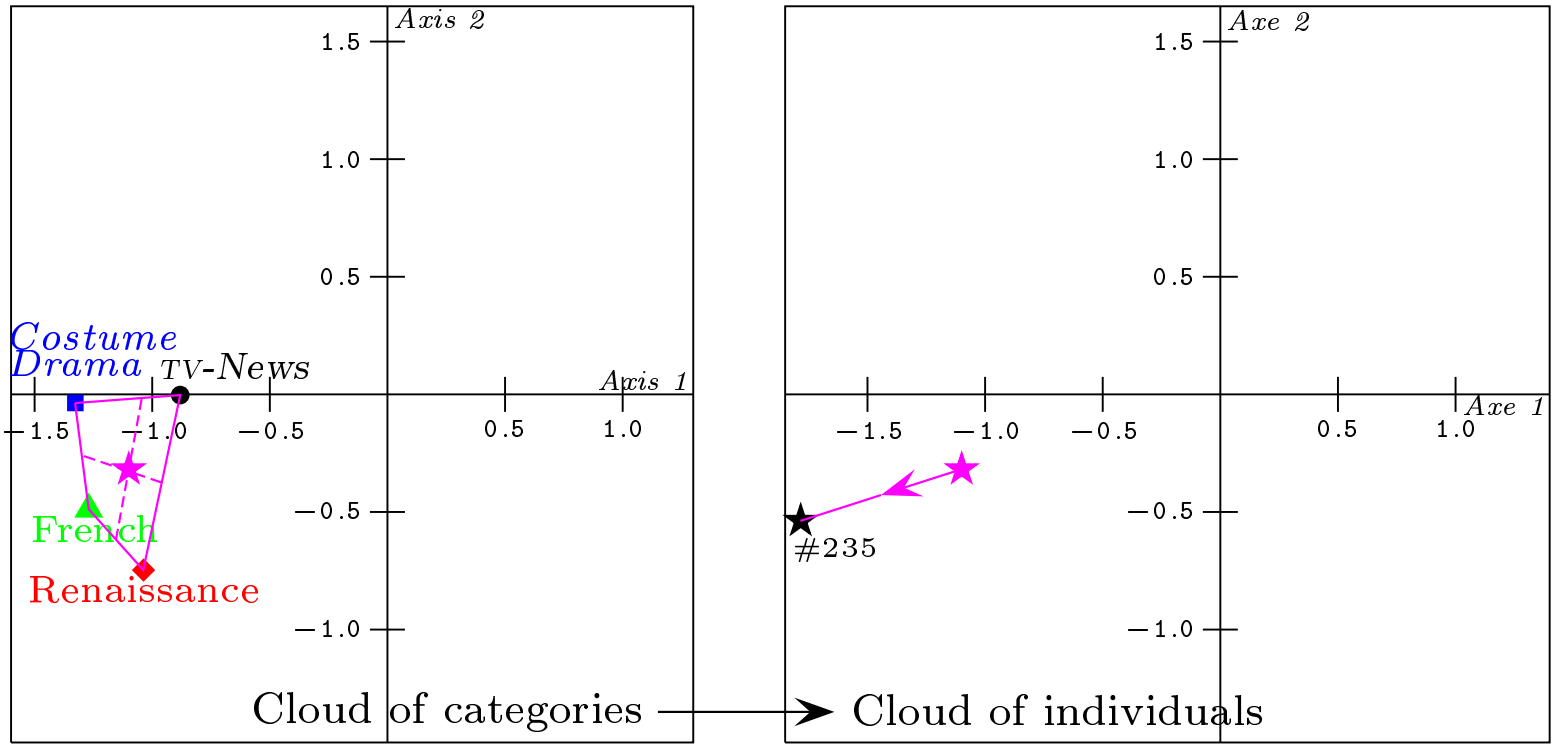
$$\cos^2 \theta_{k\ell} = \frac{(\mathrm{GM}^k_\ell)^2}{(\mathrm{GM}^k)^2} = \frac{(y^k_\ell)^2}{(\mathrm{GM}^k)^2}$$



— **Transition formulas**

- *First transition formula*:

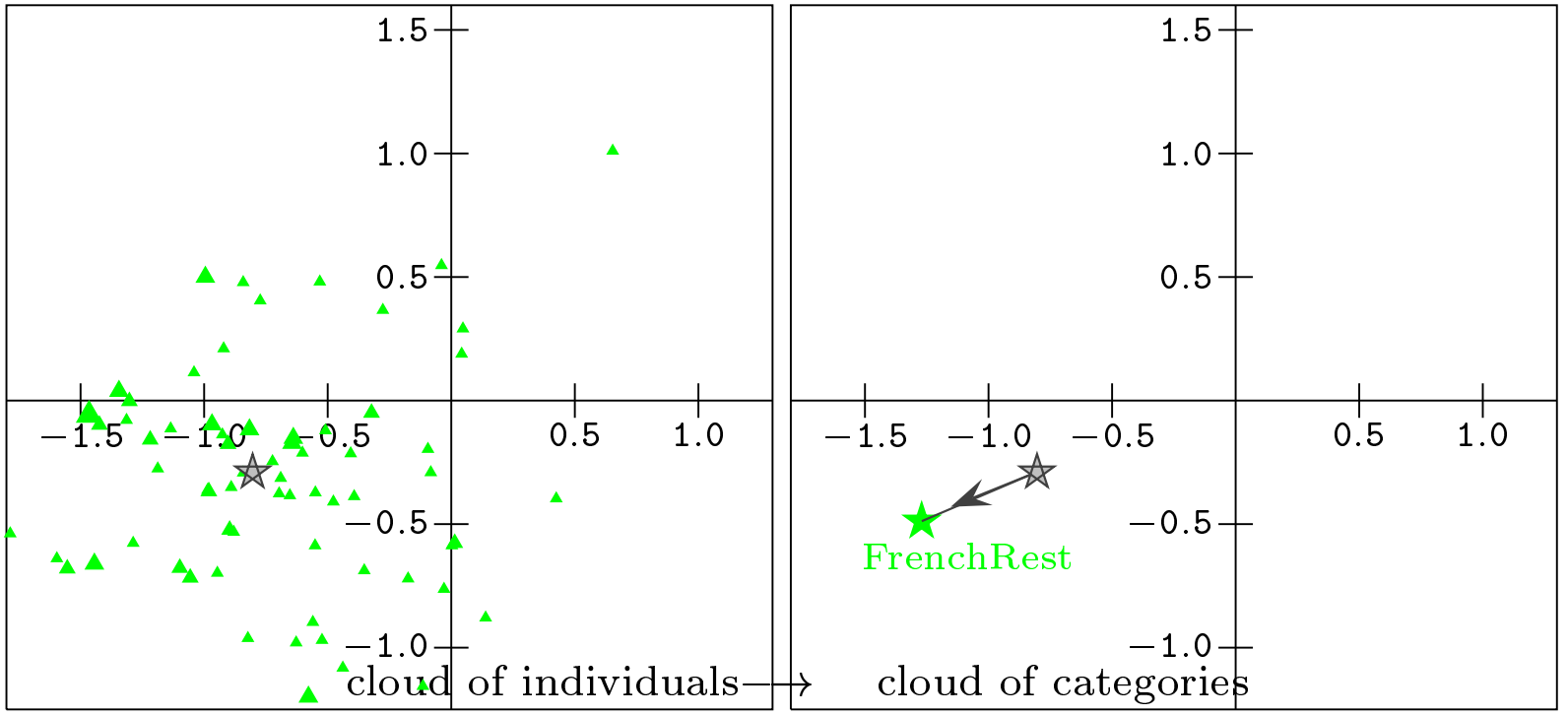
cloud of categories $\longrightarrow$ cloud of individuals: $y^i = \dfrac{1}{\sqrt{\lambda}} \sum_{k \in K_i} y^k / Q$



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

- *Second transition formula*:

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



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

— **Supplementary elements**: individuals and/or questions

— **Category mean points**

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

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

For question q, K category mean points $(\overline{\mathrm{M}}^k)_{k \in K_q}$ defines the between–q cloud.

3.6 MCA of the Taste Example

- **Data set**

The data involve :

$Q = 4$ active variables and $K = 8 + 8 + 7 + 6 = 29$ 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: $\frac{6.25}{25} = 0.25$

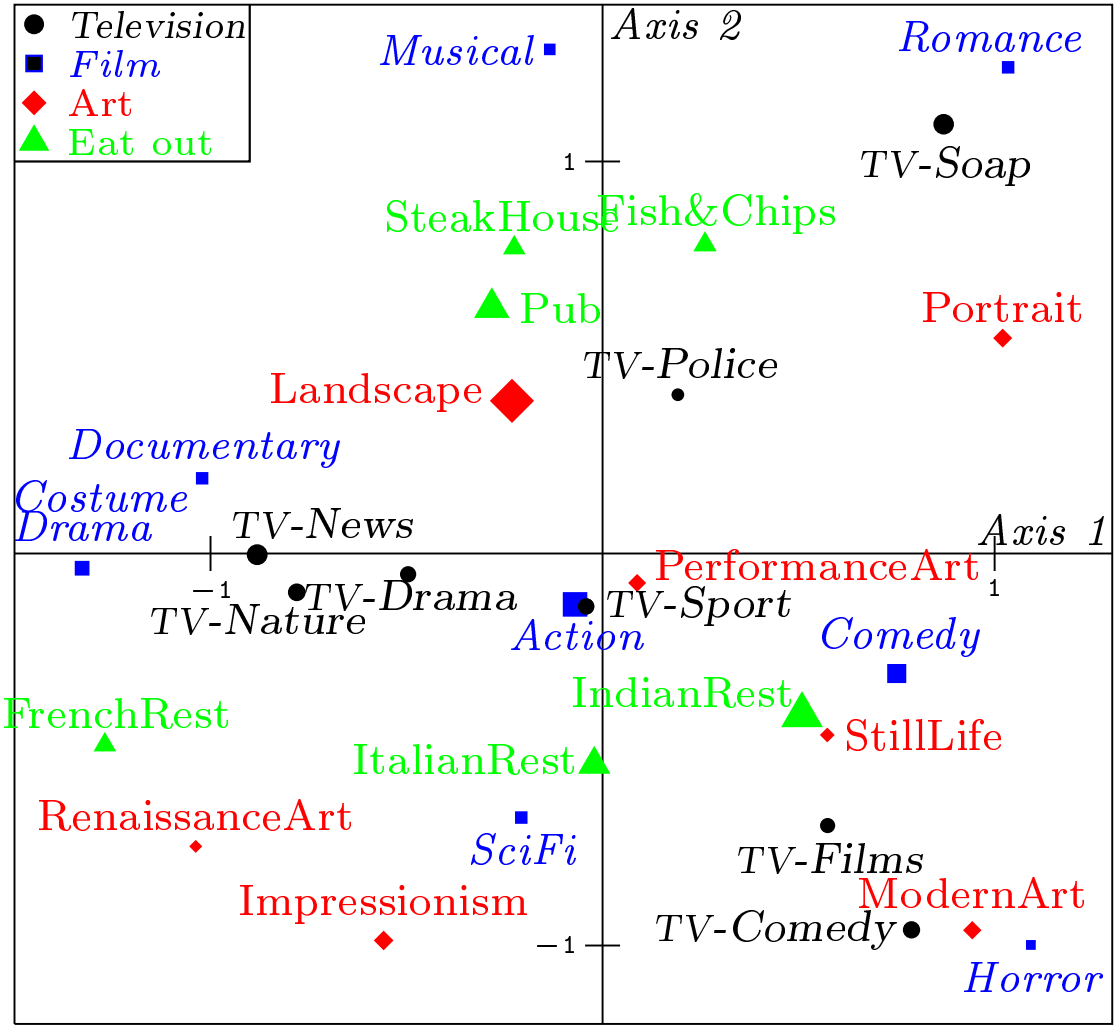
For 12 axes, the variance exceeds the average

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

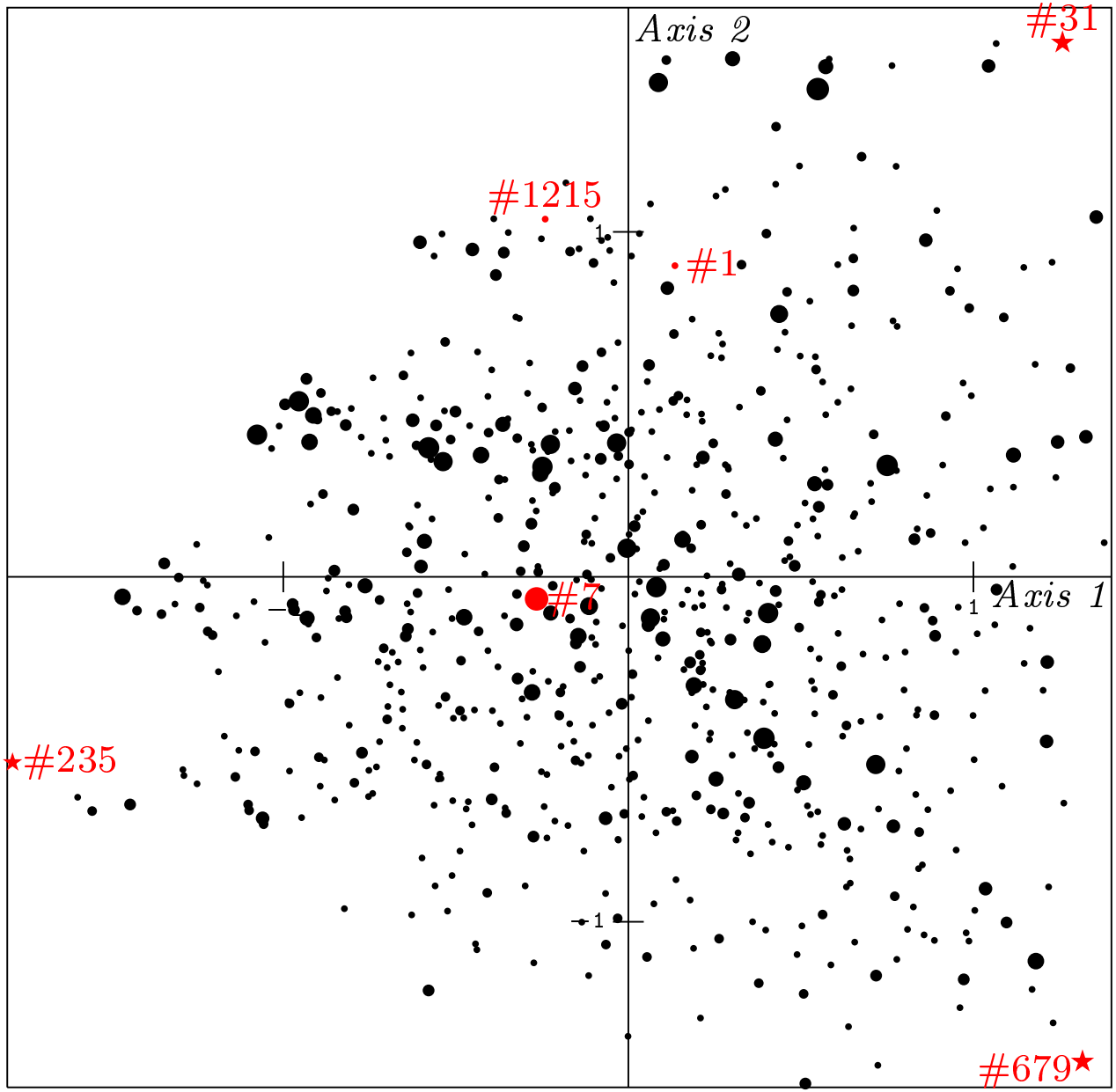
Principal coordinates and contributions of 6 individuals

	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

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 ofcategories in plane 1-2.

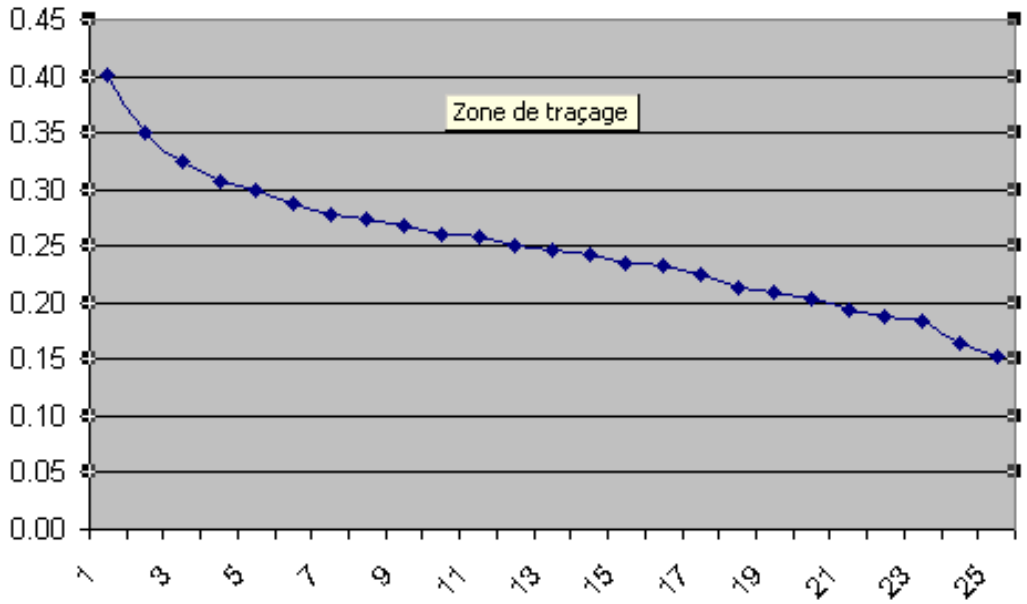


Cloud of individuals with landmark individuals mbox⋆) in plane 1-2.

• Interpretation of the analysis

How many axis to be interpreted?

$(\frac{\lambda_1-\lambda_2}{\lambda_1} = .12)$; modified rate for axis 1 = 0.48
$(\frac{\lambda_2-\lambda_3}{\lambda_2} = .07)$; modified rate for axis 2 = 0.22; The cumulated modified rate for axes 1 and 2 = 0.70 . After axis 4, variances of axes decreases 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 the first three axes = 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 = $100/29 = 3.4$

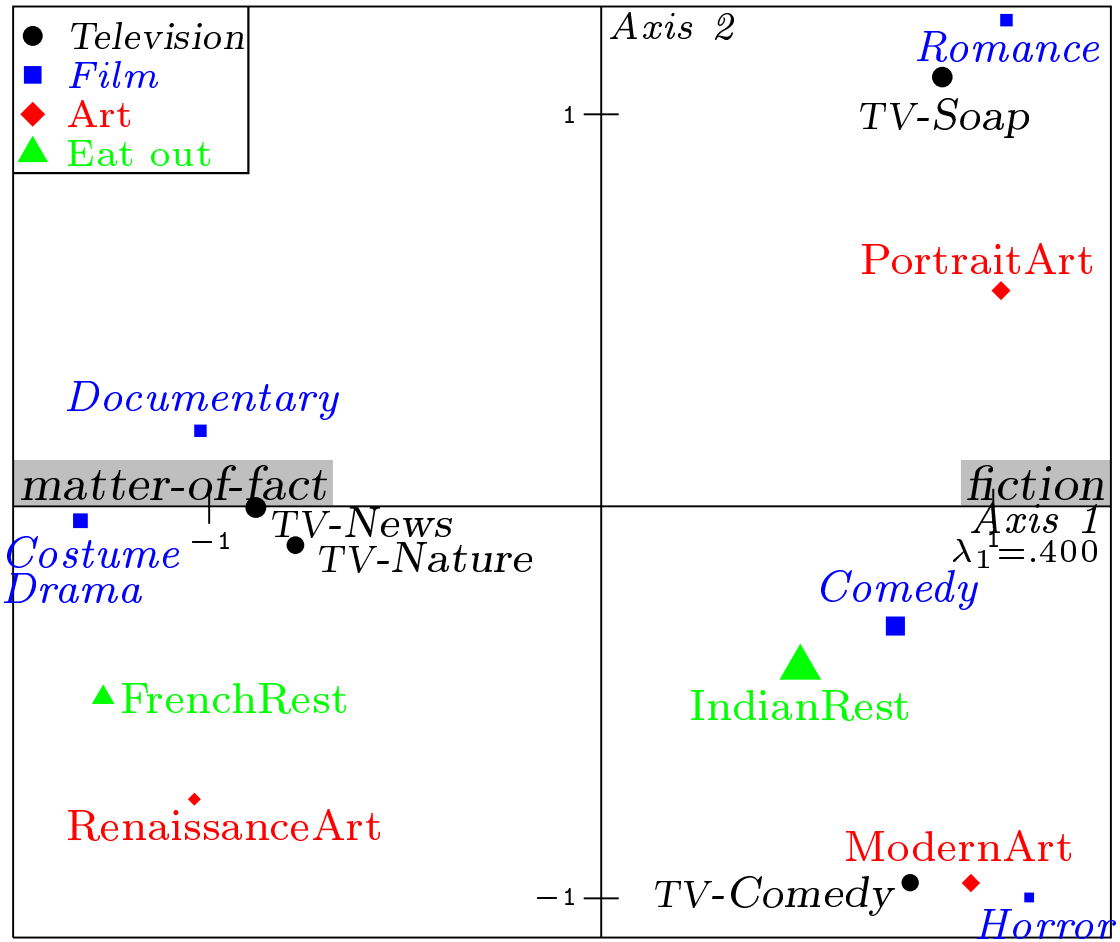
The interpretation of an axis is based on the categories which contributions to axis exceed the criterion.

Interpretation of axis 1

● *TV* (30.7%)			
	left	right	*deviation*
TV-News	8.8		
TV-Soap		8.4	*26.8*
TV-Nature	4.9		
TV-Comedy		4.9	
♦ *Art* (19.3%)			
Portrait	6.3		
Modern		5.0	*14.7*
Renaissance	3.0		

■ *Film* (34.6%)			
	left	right	*deviation*
CostumeDrama	12.7		
Comedy		6.8	
Romance		5.5	*33.1*
Documentary	5.4		
Horror		3.8	
▲ *Eat out* (15.4%)			
FrenchRest	8.2		
IndianRest		5.3	*12.9*

Total contribution: 43.0 (left) + 46.0 (right) = 89.0



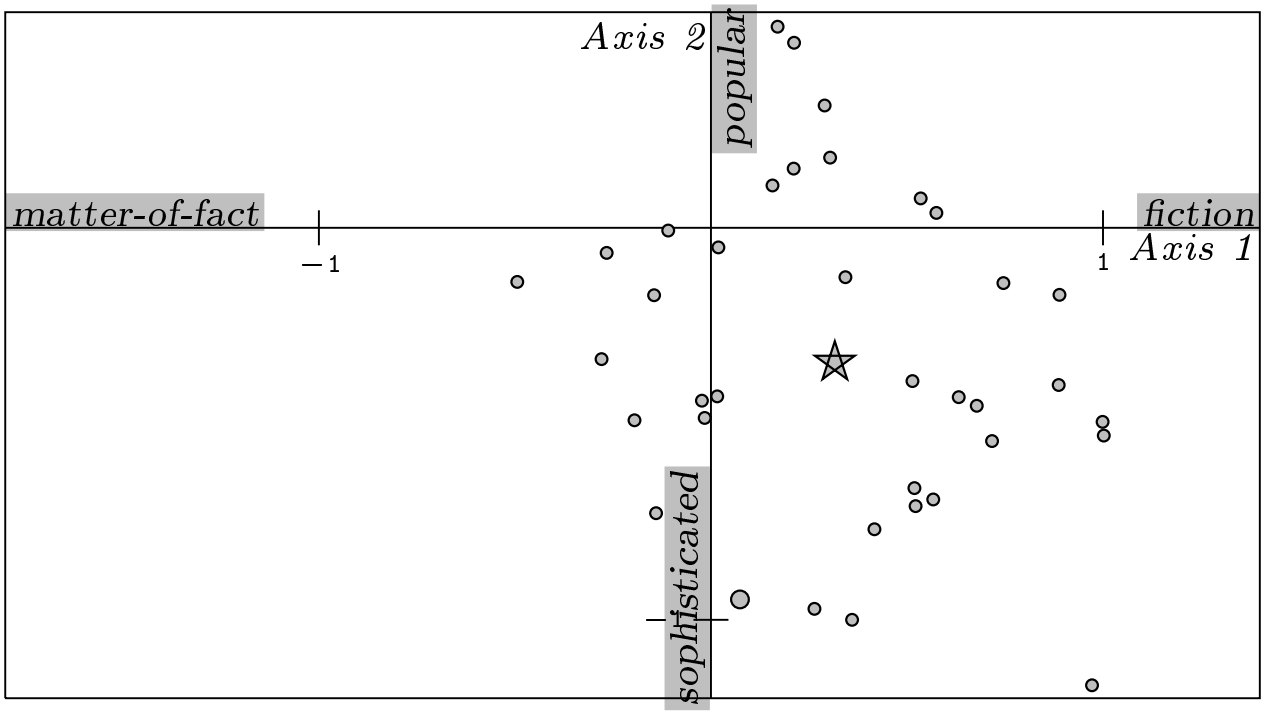
The 14 categories selected for the interpretation of Axis 1 in plane 1-2.

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 *outwardly dispositions* to *inwardly dispositions*.

• Supplementary groups of individuals



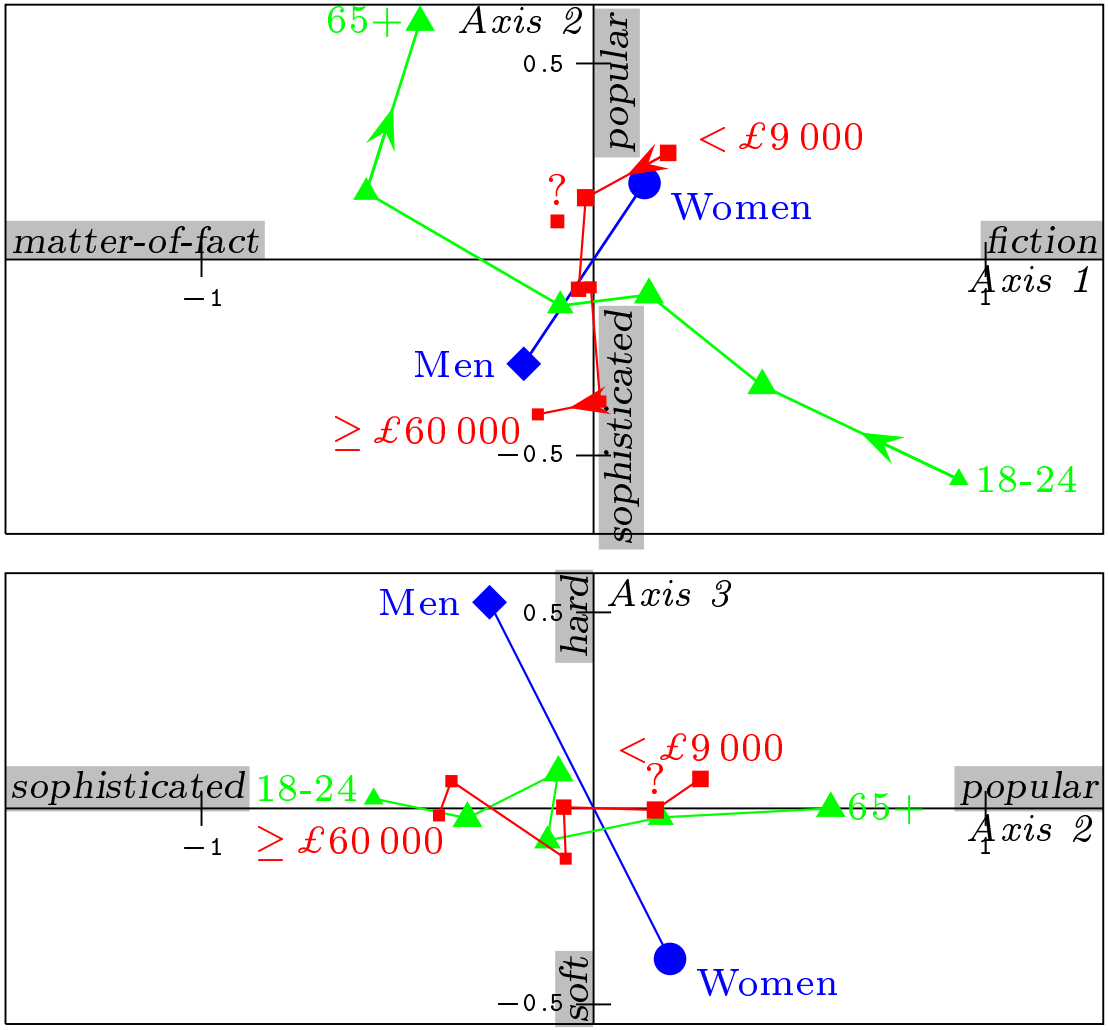
Plane 1-2. Cloud of 38 Indian immigrants with its mean point (★).

• 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).

3.7 Two variants of MCA

- *Specific MCA (SpeMCA)* consists in restricting the analysis to categories of interest.
- *Class Specific Analysis (CSA)* consists in analyzing a subcloud of individuals.

Specific MCA

The active categories are the categories of interest.

The excluded categories are:

- *Infrequent categories*
  - — remote from the center
  - — contributing to much to the variance of the question
  - — too influential for 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 category k and i' category k'

- $d_q^{2\prime} = d_q^2(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}}$ if both k and k' are active categories;
- $d_q^2(i,i') = \frac{1}{f_k}$ if k is active and k' is passive (dropping $\frac{1}{f_{k'}}$).

Geometric viewpoint: projection of the cloud onto a subspace.

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

Properties

- Number of dimensions of the cloud:

$K' - Q'$ (Q' = 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;

then modified rates are equal to $\dfrac{(\lambda - \overline{\lambda})^2}{\sum(\lambda - \overline{\lambda})^2}$ ($\sum$ 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, that is,

- 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)

3.7.1 Class Specific Analysis (CSA)

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

Determine the specific features of the class.

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