III — Introduction to Multiple Correspondence Analysis (MCA)

Adapted from "What is MCA?", paper presented by Brigitte Le Roux at *Research Methods Festival*, Oxford (July 2006).

Brigitte.LeRoux@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/l̃erb/
rouanet@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/r̃ouanet/

1 Introduction

Language of questionnaire

Basic data set: Individuals×Questions table

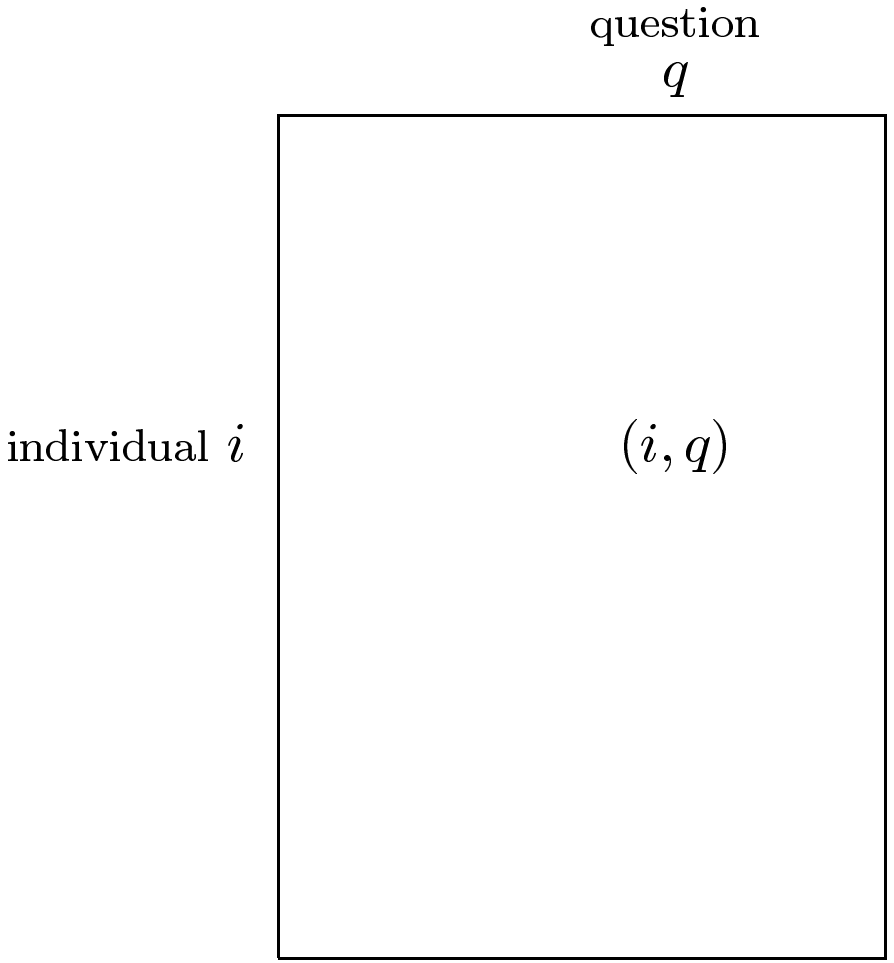
Questions are categorized variables, that is, variables with a finite number of response categories, or *modalities*.

Questionnaire in "standard format": for each question, each individual chooses *one and only one* response modality.

I: set of n individuals

Q: set of questions

Basic data table analyzed by MCA:



1.1 Historical landmarks

Guttman (1941)

Burt (1950)

Benzécri (1972-1977)

Lebart (1975)

Bourdieu & Saint-Martin 1978 (*Le patronat*).

2 Principles of MCA

$$\text{MCA} \longrightarrow \begin{cases} \text{cloud of individuals} \\ \text{cloud of modalities} \end{cases}$$

Distance between two individuals i and i' for question q

i chooses modality k; i' chooses modality $k' \neq k$:

$$d_q^2(i,i') = \frac{1}{n_k/n} + \frac{1}{n_{k'}/n}$$

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

Distance between two modalities 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 n (resp. $n_{k'}$);
$n_{kk'}$ = number of individuals who have chosen both k and k'.

Principal axes, eigenvalues and contributions

Fundamental properties:

• the principal axes of the cloud of individuals are in a one-one correspondence with those of the cloud of modalities,

• the two clouds have the same eigenvalues.

Aids to interpretation: Contributions

Overall contribution of question q: $\frac{K_q - 1}{K - Q}$

(K_q: number of modalities of question q, K overall number of modalities)

Contribution of point to axis: $\frac{p\, y^2}{\lambda}$

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

Contributions add up by grouping → contribution of a question.

3 Steps of analysis

1. Choose active individuals, active questions and encode modalities;
2. Decide how many axes to be interpreted;
3. Visualize the two clouds;
4. Interpret the axes retained in the cloud of modalities;
5. Investigate the cloud of individuals (landmark patterns, structuring factors, ellipses).

The Culture Example

Data from a 1997 survey on the cultural practices of French people conducted by O. Donnat[a] & Col at the Department of Studies and Prospective of the Ministry of Culture and Communication.

Sample: 3002 individuals aged 15 or more, representative of the French population; 125 questions

Our dataset[b]: $Q = 6$ questions pertaining to *leisure activities*, and 3 *identification variables*; a set I of 2720 individuals aged 18 or more who answered fully the six questions.

Research Questions: Are there differences about leisure practice between genders, age classes and educational levels ?

---

[a] See O. Donnat (1998), *Les pratiques culturelles des français: enquête 1997*, Paris, La documentation française.

[b] See B. Le Roux & H. Rouanet (2004), p.221-241.

Data Set

(***q*1**). As a general rule, do you prefer *Leisure activities* that you can do

*q*1	*Leisure activity*	freq.	%
*q*1*r*1	*alone*	434	16.0
*q*1*r*2	with *partner*	507	18.6
*q*1*r*3	with *friends*	1060	39.0
*q*1*r*4	with *family*	719	26.4
	Total	2720	100.0

(***q*2**). As a general rule, would you say that during your *Free time*

*q*2	*Free time*	freq.	%
*q*2*r*1	you *lack time* to do all that you would like to do	1121	41.2
*q*2*r*2	you don't lack time but you have *always something to do*	1157	42.5
*q*2*r*3	sometimes you have *nothing particular to do*	241	8.9
*q*2*r*4	often you *do nothing in particular*	201	7.4
	Total	2720	100.0

(***q3***). If you had *More time*, your first choice of activity would be

*q*3	*More time*	freq.	%
*q*3*r*1	to *rest*, not to do anything in particular	304	11.2
*q*3*r*2	to *take courses* to improve your work situation	262	9.6
*q*3*r*3	to discover or practice more *physical activities*	573	21.1
*q*3*r*4	to discover or practice more *artistic activities*		
*q*3*r*5	to develop your *general knowledge*	449	16.5
*q*3*r*6	to *take care* of your *family*	316	11.6
*q*3*r*7	to do some *home DIY* (gardening, etc.)	422	15.5
	Total	2720	100.0

(***q4***). When you *Go out* in the evening, do you usually go

*q*4	*Going out*	freq.	%
*q*4*r*1	*alone*	202	7.4
*q*4*r*2	with your *partner*	911	33.5
*q*4*r*3	with *family*, children, parents, etc.	546	20.1
*q*4*r*4*a*	with *friends*	538	19.8
*q*4*r*4*b*	with a *group* (workers' council, club, etc.)	54	1.4
*q*4*r*5	you *don't go out* in the evening.	469	17.2
	Total	2720	100.0

We have constructed two further questions $q5$ and $q6$:

Question $q5$ was built from questions about the time of watching *TV* and grouped in 5 categories.

Question $q6$ was built from questions about the number of *Books* and comic strips read during the last 12 months and grouped in 5 categories.

$q5$	*TV*	freq.	%
$q5r1$	never	257	9.4
$q5r2$	less than 10h	435	16.0
$q5r3$	10-19h	794	29.2
$q5r4$	19-30h	705	25.9
$q5r5$	over 30h	529	19.4
	Total	2720	100.0

$q6$	*Books*	freq.	%
$q6r1$	no books	603	22.2
$q6r2$	1-4 books	482	17.7
$q6r3$	5-12 books	641	23.6
$q6r4$	13-39 books	563	20.7
$q6r5$	40 books or more	431	15.8
	Total	2720	100.0

Choosing active questions and encoding modalities

- *Active questions*: 6 questions about leisure
- *Encoding modalities*: *TV* and *Books* recoded from the distribution of hours by week and # of books read the last 12 months.
- *Rare modalities, non-responses, "junk" modalities*

Rare modalities (say, of frequencies less that 5%) need to be pooled with others whenever feasible, or alternatively be put as "passive" ones (Specific MCA[a]).

For question *Go out*: we pool the rare modality "group" (1.4%) with the modality "friends".

- *Technique of Supplementary Elements*:

Supplementary *variables*; Supplementary *individuals*

[a]See B. Le Roux & H. Rouanet (2004), p.203-210.

Basic Results

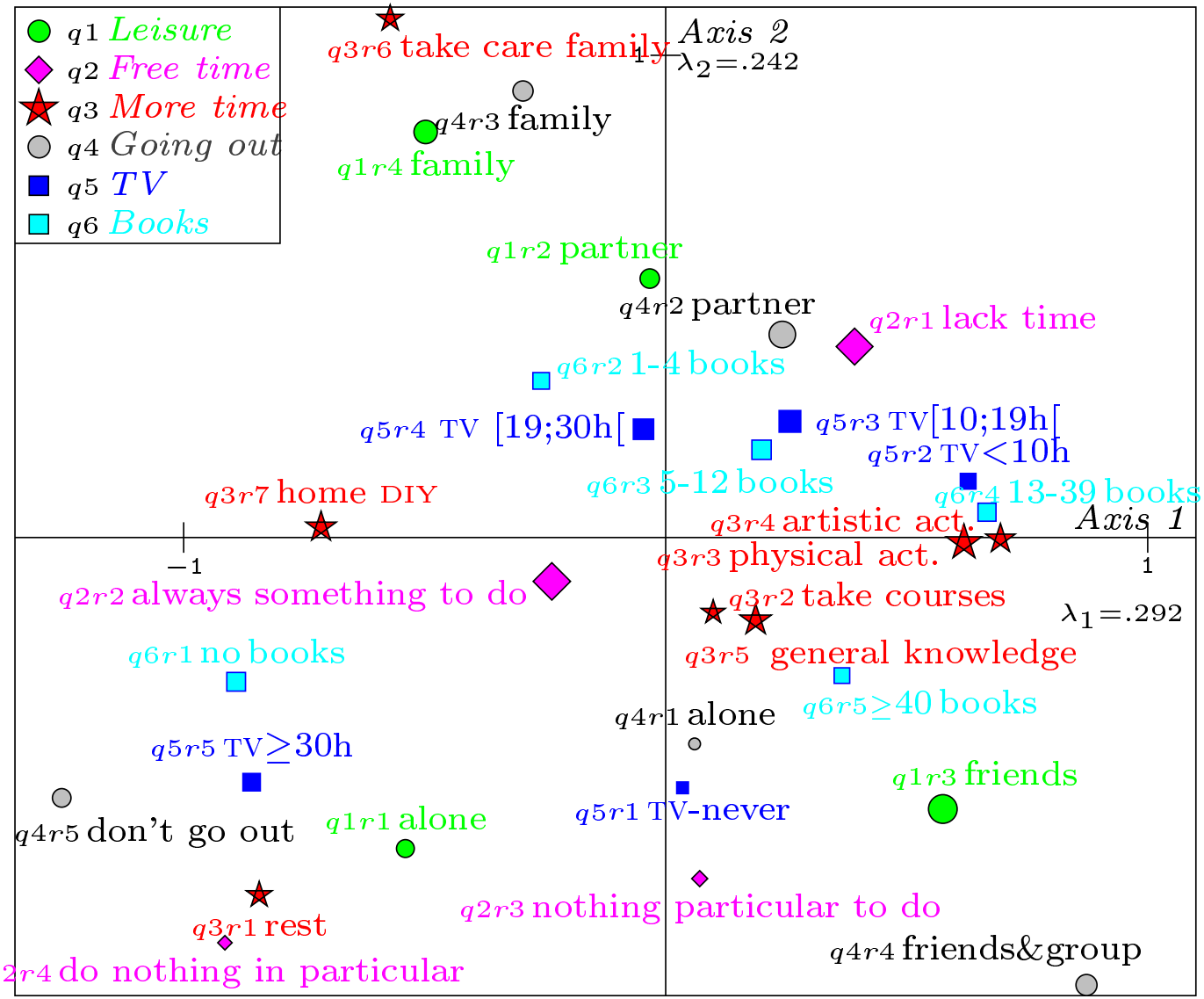
(i) the variances of axes (eigenvalues);

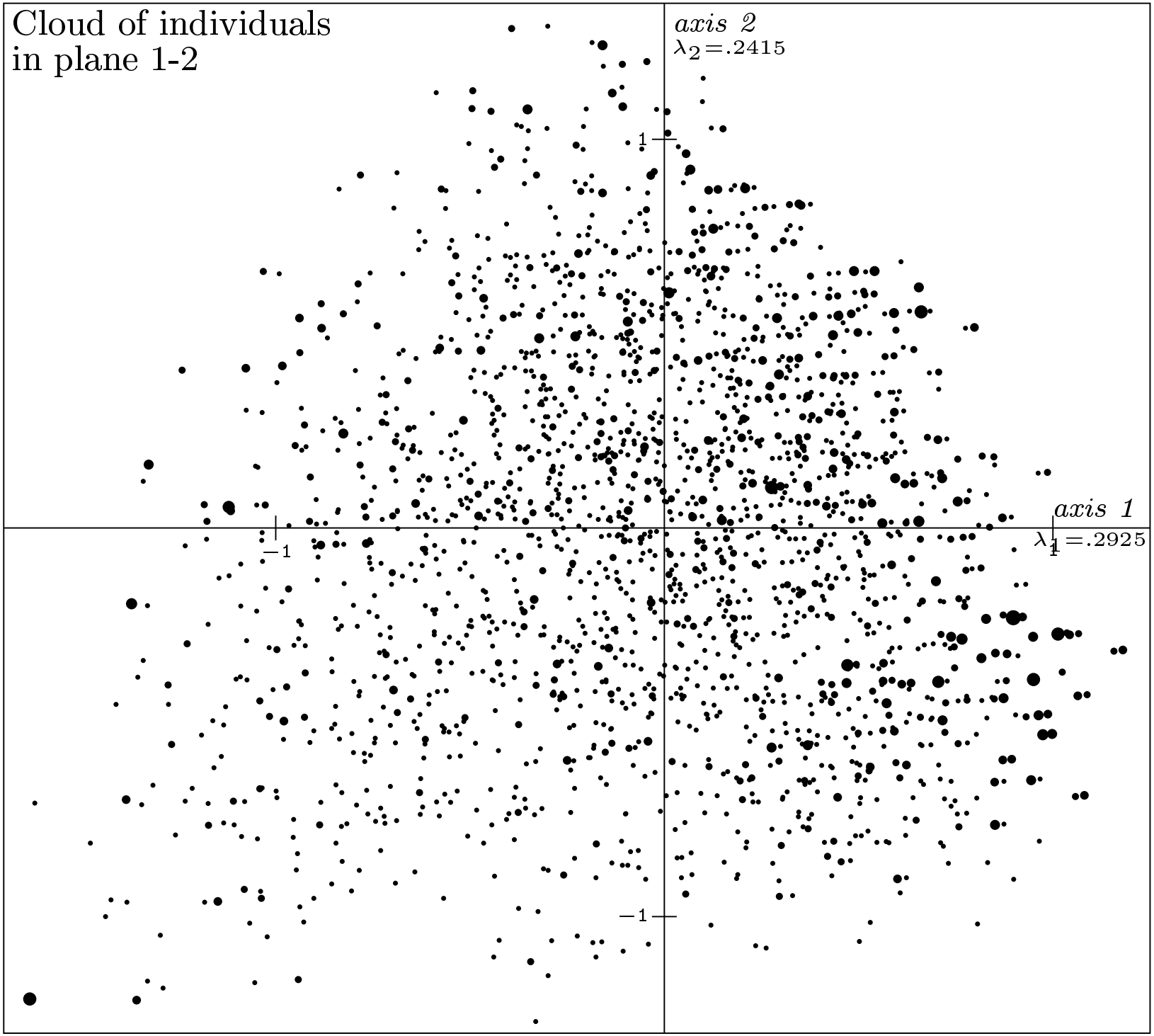
(ii) the principal coordinates of modalities (categories) and of individuals;

(iii) the contributions of modalities to axes;

(iv) the geometric representation of the two clouds (cloud of modalities and cloud of individuals).

Cloud of modalities in plane 1-2




Cloud of individuals
in plane 1-2
axis 2
λ2=.2415
1
axis 1
λ1=.2925
1
−1
−1

How many axes should be interpreted?

Eigenvalues: 10 eigenvalues exceed $\overline{\lambda} = 1/6 = .1667$:

	Eigen-values	modified rates
λ_1	.2925	.569
λ_2	.2415	.201
λ_3	.2248	.122
λ_4	.2073	.059
λ_5	.1950	
λ_6	.1832	
λ_7	.1790	
λ_8	.1758	
λ_9	.1733	
λ_{10}	.1688	

Modified rates: To assess importance of axes

Benzécri's formula, cf. Benzécri (1992), p.412, Le Roux & Rouanet (2004), p.200 &225

3 axes will be interpreted (cumulated modified rate 89%)

Interpreting axes

It is done in the cloud of modalities and based on the modalities whose contributions to axis exceed a specified threshold (e.g. average contribution) according to the *Method of contributions of points and deviations*[a], following the principle stated by Benzécri (1992, p. 405):

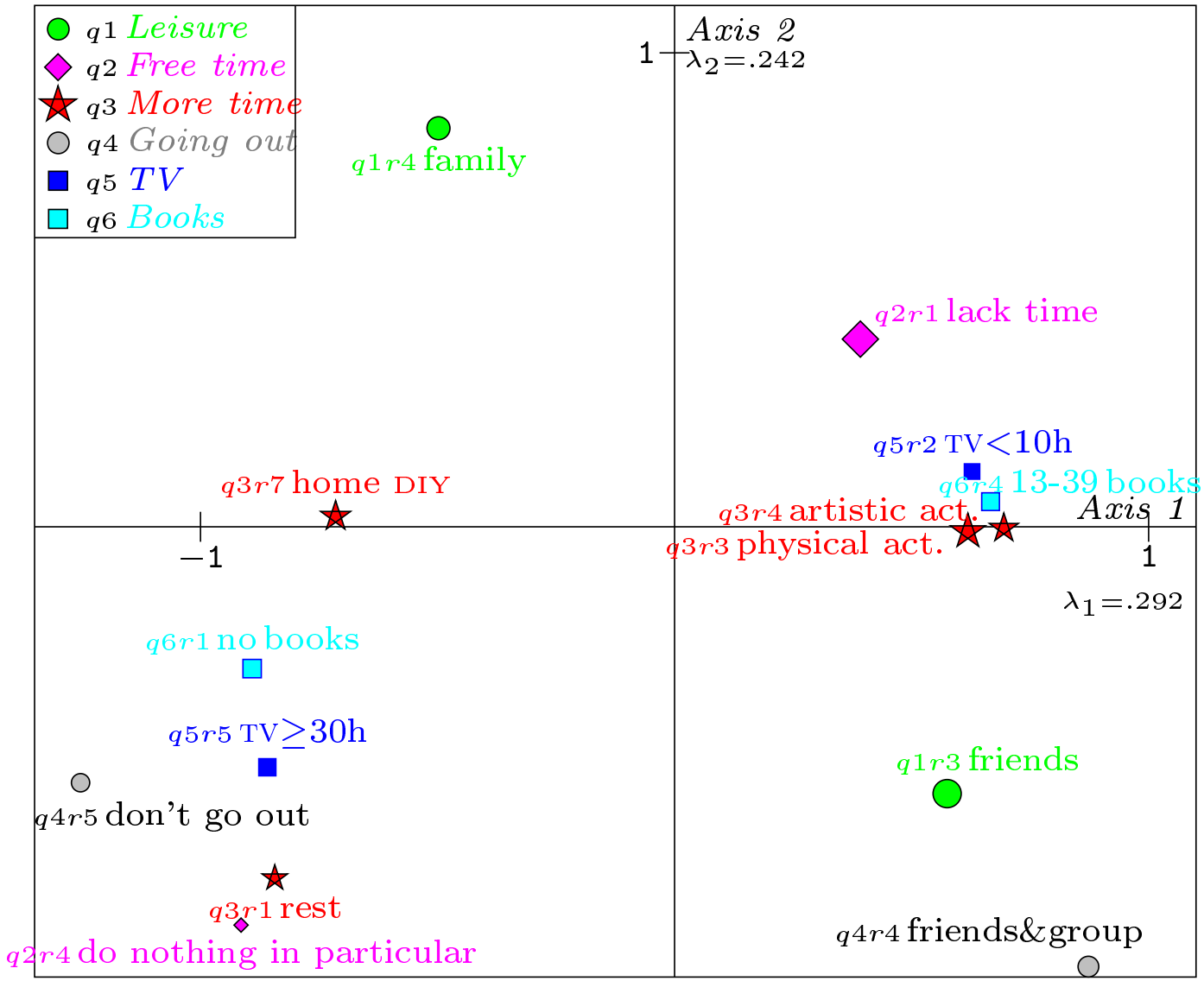
"Interpreting an axis amounts to finding out what is similar, on the one hand, between all the elements figuring on the right of the origin and, on the other hand between all that is written on the left; and expressing with conciseness and precision, the contrast (or opposition) between the two extremes."

[a]Le Roux B. & Rouanet H. (1998). Interpreting axes in Multiple Correspondence Analysis: Method of contributions of points and deviations, in *Visualization of Categorical Data*, (eds Blasius & Greenacre), 197-220, Academic Press.

3.1 Interpretation of Axis 1

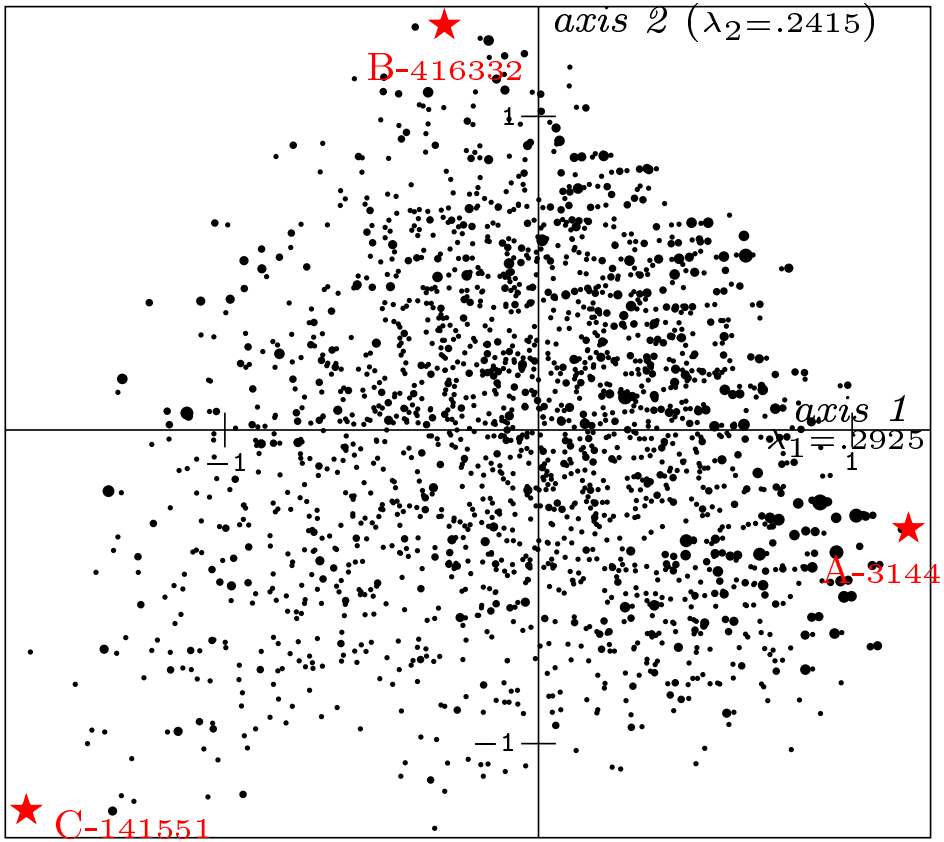
			left	right	deviation
$q4$ (27%) ● *Going out*	$r5$	don't go out	15.4		
	$r4$	friends&group		9.5	91.8
$q3$ (20%) ★ *More time*	$r3$	physical activities		4.6	
	$r7$	home DIY	4.5		
	$r1$	rest	4.5		
	$r4$	artistic activities		4.0	86.6
$q6$ (18%) ■ *Books*	$r1$	no books	10.0		
	$r4$	13-39 books		5.3	83.7
$q1$ (14%) ● *Leisure*	$r3$	friends		7.3	
	$r4$	family	3.7		75.2
$q5$ (13%) ■ *TV*	$r5$	$\geq$ 30h	8.2		
	$r2$	$<$ 10h		3.6	85.6
$q2$ (9%) ◇ *Free time*	$r1$	lack time		3.6	
	$r4$	do nothing in particular	3.5		71.6
			49.8	37.9	

Relative contribution (percentage) of each modality to axis, written either in column "left" or in column "right" according to its position on graph; relative contribution of deviation to question.

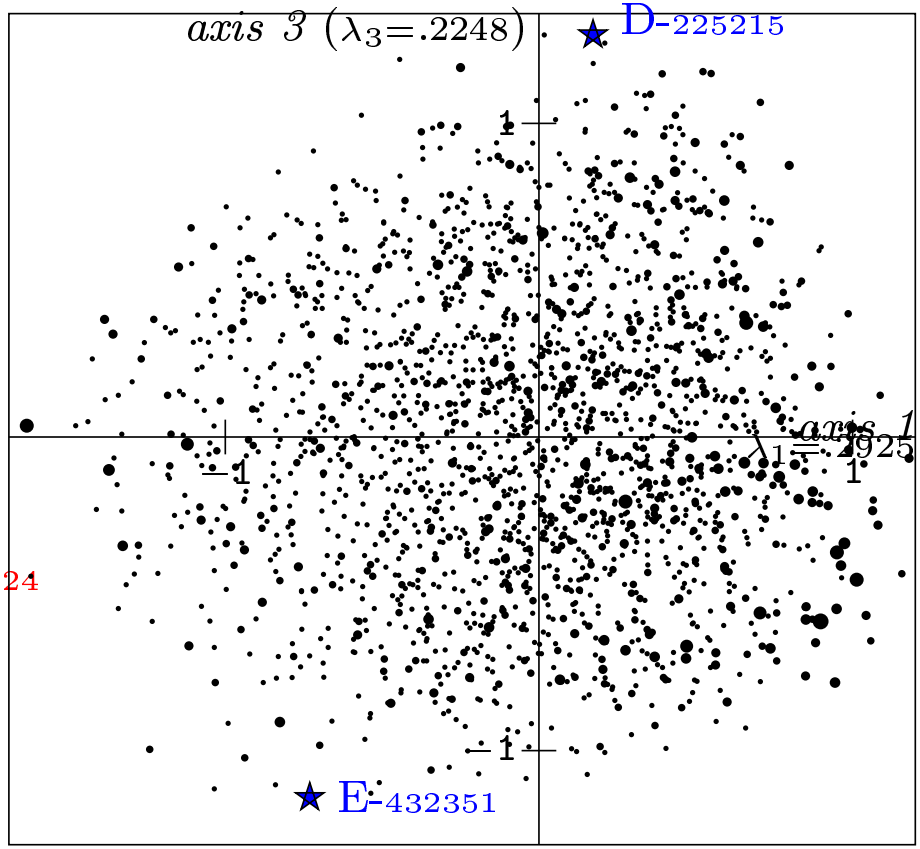


Interpretation of Axis 1: 14 modalities most contributing to axis.

Investigation of the cloud of individuals



Plane 1-2



Plane 1-3

Subclouds and modality mean points

Subcloud of the individuals having chosen one modality $\rightarrow$ its mean point is called a *modality mean-point*.

Fundamental Property:

Coordinate of the modality mean-point $= \sqrt{\lambda}\, y$

(y is the coordinate of the modality in the space of modalities)

Concentration ellipses

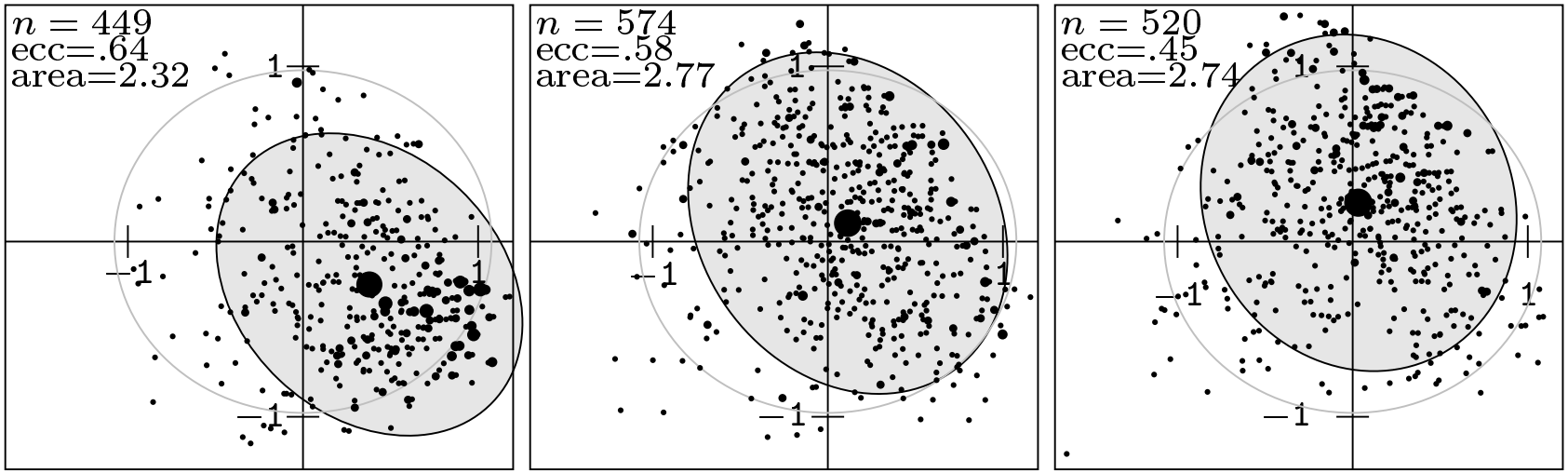
The concentration ellipse[a] of a subcloud is such that the half–axis of the ellipse is along the principal direction of the subcloud projected in the plane under study and its length is equal to $2\sqrt{\lambda'}$.

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 e.g. Age in the Culture Example.

[a] see Cramér, 1946, p. 284; Le Roux & Rouanet (2004), p.95-100


n = 449
ecc=.64
area=2.32
18-25 years class
n = 574
ecc=.58
area=2.77
26-35 years class
n = 520
ecc=.45
area=2.74
36-45 years class
n = 394
ecc=.55
area=2.93
46-55 years class
n = 317
ecc=.60
area=2.78
56-65 years class
n = 466
ecc=.71
area=2.92
>65 years class

4 Final message about MCA

MCA is a method for the *Geometric Analysis* of questionnaires

Investigate the *cloud of individuals*: individuals carry all the information!

References

— Benzécri J-P. (1977). Sur l'analyse des tableaux binaires associés à une correspondance multiple [On the analysis of binary tables associated with a multiple correspondence], *Les Cahiers de l'Analyse des Données*, 2, 55-71 (from a mimeographed note of 1972).

— Benzécri J-P. (1992). *Correspondence Analysis Handbook*, New York, Dekker.

— Escofier B. & Pagès J. (1988), *Analyses factorielles simples et multiples* [Simple and Multiple factor Analyses] (chapter 3 on mca). Paris: Dunod.

— Greenacre M. (1984). *Theory and Applications of Correspondence Analysis* (chapter 5 on mca). London: Academic Press.

— Le Roux B. (2006). What is MCA?, *Research Methods Festival*, Oxford, `www.ccsr.ac.uk/methods/festival/programme/Wita/leroux.pdf`

— Le Roux B., & Rouanet H. (2004). *Geometric Data Analysis; From Correspondence Analysis to Structured Analysis*. Dordrecht:

Kluwer (the chapter 5 contains a detailed presentation of MCA and an extensive illustration: "the Culture Example", p.221-251)

— Lebart L., Morineau A. & Warwick K.M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices* (chapter 4 on MCA). New York: Wiley.

Recent substantive studies

— Bonnet P., Le Roux B. & Lemaine G. (1996). Analyse géométrique des données: une enquête sur le racisme [Geometric Data Analysis: a survey on racism], *Mathématiques et Sciences Humaines*, 136, 5-24.
— Bourdieu P. (1999). Une révolution conservatrice dans l'édition [A conservative revolution in publishing], *Actes de la Recherche en Sciences Sociales*, Vol. 126-127, 3-28.
— Chiche J., Le Roux B., Perrineau P. & Rouanet H. (2000). L'espace politique des électeurs français à la fin des années 1990 [The political space of French electors in the late 1990s]. *Revue française de sciences politiques*, 50, 463-487.

— Hjellbrekke j., Le Roux B., Korsnes O., Lebaron F., Rosenlund L. & Rouanet H. (to appear). The Norwegian field of Power Anno 2000. *European Societies.*
— Le Roux B. (2006). Que pense votre député de la mondialisation (What does your deputy think of the globalization).
`www.telos-eu.com/2006/05/que_pense_votre_depute_de_la_m.php`
— Le Roux B. & Rouanet H. (2003). Geometric Analysis of Individual Differences in Mathematical Performance for epgy Students in the Third Grade. `www-epgy.stanford.edu/research/`.
— See chapter "Case Studies" in Le Roux & Rouanet (2004) (op. cit.)

About software

These results have been obtained using addad, ellipse and EyeLID software freely avalaible from my website `math-info.univ-paris5.fr/~lerb/` under the "Logiciels" heading.

All results presented here are obtainable from the September 2006 version of spad software distributed by spad company (`www.spad.eu`).