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/~lerb/
rouanet@math-info.univ-paris5.fr www.math-info.univ-paris5.fr/~rouanet/
1 Introduction

Language of questionnaire

Basic data set: Individuals $\times$ 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

\[
\text{MCA} \rightarrow \begin{cases} 
\text{cloud of individuals} \\
\text{cloud of modalities}
\end{cases}
\]

**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*: $\frac{py^2}{\lambda}$

($y$: coordinate of point on axis; $p$: relative weight; $\lambda$: eigenvalue)

Contributions add up by grouping $\rightarrow$ 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\textsuperscript{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\textsuperscript{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?


**Data Set**

(q1). As a general rule, do you prefer *Leisure activities* that you can do

<table>
<thead>
<tr>
<th>q1</th>
<th><em>Leisure activity</em></th>
<th>freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1r1</td>
<td>alone</td>
<td>434</td>
<td>16.0</td>
</tr>
<tr>
<td>q1r2</td>
<td>with partner</td>
<td>507</td>
<td>18.6</td>
</tr>
<tr>
<td>q1r3</td>
<td>with friends</td>
<td>1060</td>
<td>39.0</td>
</tr>
<tr>
<td>q1r4</td>
<td>with family</td>
<td>719</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>2720</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>

(q2). As a general rule, would you say that during your *Free time*

<table>
<thead>
<tr>
<th>q2</th>
<th><em>Free time</em></th>
<th>freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>q2r1</td>
<td>you <em>lack time</em> to do all that you would like to do</td>
<td>1121</td>
<td>41.2</td>
</tr>
<tr>
<td>q2r2</td>
<td>you don’t lack time but you have <em>always something to do</em></td>
<td>1157</td>
<td>42.5</td>
</tr>
<tr>
<td>q2r3</td>
<td>sometimes you have <em>nothing particular to do</em></td>
<td>241</td>
<td>8.9</td>
</tr>
<tr>
<td>q2r4</td>
<td>often you <em>do nothing in particular</em></td>
<td>201</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>2720</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>
(q3). If you had *More time*, your first choice of activity would be

<table>
<thead>
<tr>
<th>q3</th>
<th>More time</th>
<th>freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>q3r1</td>
<td>to <em>rest</em>, not to do anything in particular</td>
<td>304</td>
<td>11.2</td>
</tr>
<tr>
<td>q3r2</td>
<td>to <em>take courses</em> to improve your work situation</td>
<td>262</td>
<td>9.6</td>
</tr>
<tr>
<td>q3r3</td>
<td>to discover or practice more <em>physical activities</em></td>
<td>573</td>
<td>21.1</td>
</tr>
<tr>
<td>q3r4</td>
<td>to discover or practice more <em>artistic activities</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>q3r5</td>
<td>to develop your <em>general knowledge</em></td>
<td>449</td>
<td>16.5</td>
</tr>
<tr>
<td>q3r6</td>
<td>to <em>take care</em> of your family</td>
<td>316</td>
<td>11.6</td>
</tr>
<tr>
<td>q3r7</td>
<td>to do some <em>home DIY</em> (gardening, etc.)</td>
<td>422</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>2720</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>q4</th>
<th>Going out</th>
<th>freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>q4r1</td>
<td><em>alone</em></td>
<td>202</td>
<td>7.4</td>
</tr>
<tr>
<td>q4r2</td>
<td>with your <em>partner</em></td>
<td>911</td>
<td>33.5</td>
</tr>
<tr>
<td>q4r3</td>
<td>with <em>family</em>, children, parents, etc.</td>
<td>546</td>
<td>20.1</td>
</tr>
<tr>
<td>q4r4a</td>
<td>with <em>friends</em></td>
<td>538</td>
<td>19.8</td>
</tr>
<tr>
<td>q4r4b</td>
<td>with a <em>group</em> (workers’ council, club, etc.)</td>
<td>54</td>
<td>1.4</td>
</tr>
<tr>
<td>q4r5</td>
<td>you <em>don’t go out</em> in the evening.</td>
<td>469</td>
<td>17.2</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>2720</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>
We have constructed two further questions \( q_5 \) and \( q_6 \):

Question \( q_5 \) was built from questions about the time of watching \( TV \) and grouped in 5 categories.

Question \( q_6 \) was built from questions about the number of \( Books \) and comic strips read during the last 12 months and grouped in 5 categories.

<table>
<thead>
<tr>
<th>( q_5 )</th>
<th>( TV )</th>
<th>freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_{5r1} )</td>
<td>never</td>
<td>257</td>
<td>9.4</td>
</tr>
<tr>
<td>( q_{5r2} )</td>
<td>less than 10h</td>
<td>435</td>
<td>16.0</td>
</tr>
<tr>
<td>( q_{5r3} )</td>
<td>10-19h</td>
<td>794</td>
<td>29.2</td>
</tr>
<tr>
<td>( q_{5r4} )</td>
<td>19-30h</td>
<td>705</td>
<td>25.9</td>
</tr>
<tr>
<td>( q_{5r5} )</td>
<td>over 30h</td>
<td>529</td>
<td>19.4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2720</td>
<td>100.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( q_6 )</th>
<th>( Books )</th>
<th>freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_{6r1} )</td>
<td>no books</td>
<td>603</td>
<td>22.2</td>
</tr>
<tr>
<td>( q_{6r2} )</td>
<td>1-4 books</td>
<td>482</td>
<td>17.7</td>
</tr>
<tr>
<td>( q_{6r3} )</td>
<td>5-12 books</td>
<td>641</td>
<td>23.6</td>
</tr>
<tr>
<td>( q_{6r4} )</td>
<td>13-39 books</td>
<td>563</td>
<td>20.7</td>
</tr>
<tr>
<td>( q_{6r5} )</td>
<td>40 books or more</td>
<td>431</td>
<td>15.8</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2720</td>
<td>100.0</td>
</tr>
</tbody>
</table>
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

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

axis 1

axis 2

Leisure
Free time
More time
Going out
TV
Books

taxe family
partner
partner
lack time
1-1 books
5-12 books
13-39 books

physical act.
always something to do
no books
TV≥30h
don’t go out
rest
do nothing in particular

TV-never

friends
friends&group

take courses
general knowledge
alone
≥40 books

imp
Cloud of individuals in plane 1-2

axis 2
\[ \lambda_2 = 0.2415 \]

axis 1
\[ \lambda_1 = 0.2925 \]
How many axes should be interpreted?

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

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalues</th>
<th>modified rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>.2925</td>
<td>.569</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>.2415</td>
<td>.201</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>.2248</td>
<td>.122</td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>.2073</td>
<td>.059</td>
</tr>
<tr>
<td>$\lambda_5$</td>
<td>.1950</td>
<td></td>
</tr>
<tr>
<td>$\lambda_6$</td>
<td>.1832</td>
<td></td>
</tr>
<tr>
<td>$\lambda_7$</td>
<td>.1790</td>
<td></td>
</tr>
<tr>
<td>$\lambda_8$</td>
<td>.1758</td>
<td></td>
</tr>
<tr>
<td>$\lambda_9$</td>
<td>.1733</td>
<td></td>
</tr>
<tr>
<td>$\lambda_{10}$</td>
<td>.1688</td>
<td></td>
</tr>
</tbody>
</table>

Modified rates: To assess importance of axes


*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.”

### 3.1 Interpretation of Axis 1

<table>
<thead>
<tr>
<th>Question</th>
<th>Modality</th>
<th>Right</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>q4 (27%)</td>
<td>Going out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r5</td>
<td>don’t go out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>friends&amp;group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q3 (20%)</td>
<td>More time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>physical activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r7</td>
<td>home DIY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r1</td>
<td>rest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>artistic activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q6 (18%)</td>
<td>Books</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r1</td>
<td>no books</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>13-39 books</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q1 (14%)</td>
<td>Leisure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>friends</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>family</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q5 (13%)</td>
<td>tv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r5</td>
<td>≥ 30h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>&lt; 10h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q2 (9%)</td>
<td>Free time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r1</td>
<td>lack time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r4</td>
<td>do nothing in particular</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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=0.64
area=2.32

18-25 years class

$n = 574$
ecc=0.58
area=2.77

26-35 years class

$n = 520$
ecc=0.45
area=2.74

36-45 years class

$n = 394$
ecc=0.55
area=2.93

46-55 years class

$n = 317$
ecc=0.60
area=2.78

56-65 years class

$n = 466$
ecc=0.71
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


— 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)


**Recent substantive studies**


— **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


### About software

These results have been obtained using ADDAD, ellipse and EyelID software freely available 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).