Multiple Correspondence Analysis

François Husson

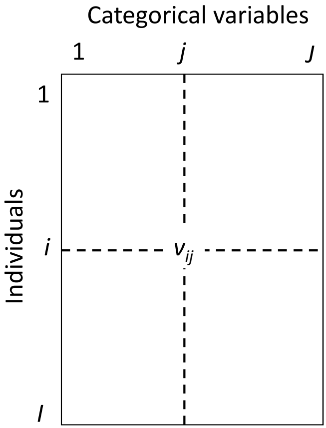
Applied Mathematics Department - AGROCAMPUS OUEST

husson@agrocampus-ouest.fr

Plan

1. Data - issues
2. Studying the individuals
3. Studying the categories
4. Interpretation aids

The data

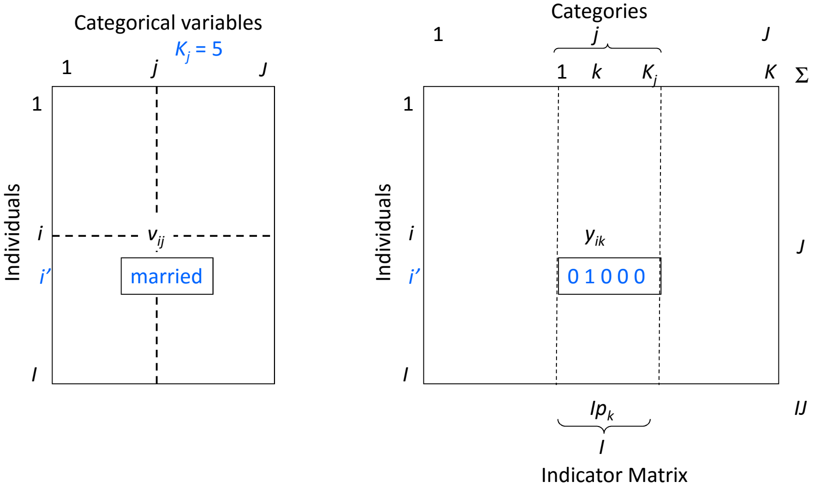


I individuals
J qualitative variables

v_{ij} : category of the j-th variable possessed by the i-th individual

Example : survey where I people reply to J multiple-choice questions

The data


Categorical variables
$K_j = 5$
1
j
J
1
Individuals
i
v_{ij}
i'
married
I
Categories
1
j
J
1
k
K_j
K
Σ
1
Individuals
i
y_{ik}
i'
0 1 0 0 0
J
I
Ip_k
IJ
I
Indicator Matrix

Goals

1. Studying the individuals

   One individual = one row of the CDT = set of categories
   Similarity of individuals – Inter-individual variability
   Principal axes of the inter-individual variability
   (in relation to the categories)

2. Studying the variables

   Links between qualitative variables
   (in relation to the categories)
   Visualization of the set of associations between categories
   Synthetic variables
   (quantitative indicators based on the qualitative variables)

$\Rightarrow$ Similar problem to PCA

Leisure activity data

- Extract from 2003 INSEE survey on identity construction, called the "history of life" survey
- 8403 individuals
- 2 sorts of variables :
  - *Which of the following leisure activities do you practice regularly* : Reading, Listening to music, Cinema, Shows, Exhibitions, Computer, Sport, Walking, Travel, Playing a musical instrument, Collecting, Voluntary work, Home improvement, Gardening, Knitting, Cooking, Fishing, Number of hours of TV per day on average
  - supplementary variables (4 questions) : sex, gender, profession, marital status

Leisure activity data

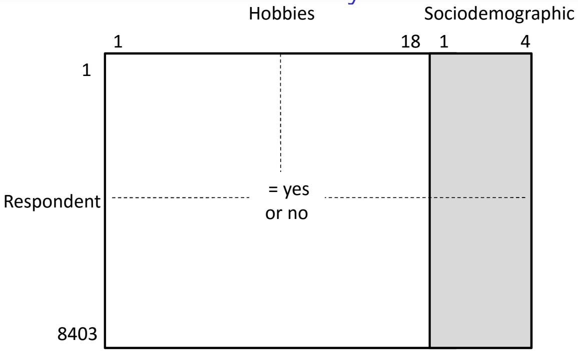
Hobbies

Hobbies		Number
Listening music		5947
Reading		5646
Walking		4175
Cooking		3686
Mechanic		3539
Travelling		3363
Cinema		3359
Gardening		3356
Computer		3158
Sport		3095
Exhibition		2595
Show		2425
Playing music		1460
Knitting		1413
Volunteering		1285
Fishing		945
Collecting		862
Number of hours watching TV	0	1017
	1	1223
	2	2156
	3	1775
	4	2232

Sociodemographic variables

Sex	Female	4616
	Male	3787
Age	[15,25]	857
	(25,35]	1302
	(35,45]	1646
	(45,55]	1837
	(55,65]	1257
	(65,75]	937
	(75,85]	482
	(85,100]	85
Marital status	Divorcee	792
	Married	4333
	Remarried	404
	Single	2140
	Widower	734
Profession	employee	2552
	foreman	735
	management	1052
	manual labourer	1161
	technician	401
	unskilled worker	792
	other	212
	No answer	1498

Leisure activity data



MCA 1 : active = leisure activity, then use supplementary data for interpretation

- 1 individual = vector of leisure activities
- Principal axes of variability of leisure vectors
- Links between these axes and the supplementary variables

MCA 2 : active = supplementary variables, leisure activities as supplementary information

MCA 3 : active = BOTH

Transforming the complete disjunctive table

An individual's weight is $\frac{1}{I}$

$$
\begin{aligned}
y_{ik} &= 1 \text{ if the } i\text{-th individual is in } k\text{-th category of the } j\text{-th variable} \\
&\quad \text{(for each } p_k\text{)} \\
&= 0 \ \text{ otherwise}
\end{aligned}
$$

$$\text{Idea :} \quad x_{ik} = y_{ik}/p_k$$

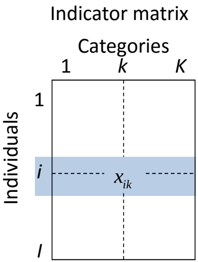
$$\frac{\sum_{i=1}^{I} x_{ik}}{I} = \frac{1}{I}\frac{\sum_{i=1}^{I} y_{ik}}{p_k} = \frac{1}{I}\frac{I \times p_k}{p_k} = 1$$

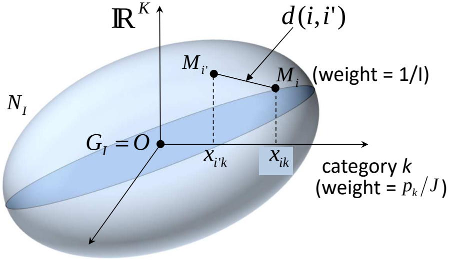
$$\text{Centering : } x_{ik} = y_{ik}/p_k - 1$$

Plan

1. Data - issues
2. Studying the individuals
3. Studying the categories
4. Interpretation aids

Point cloud of individuals

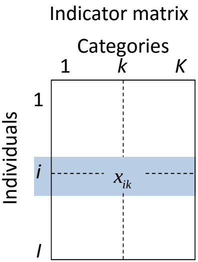


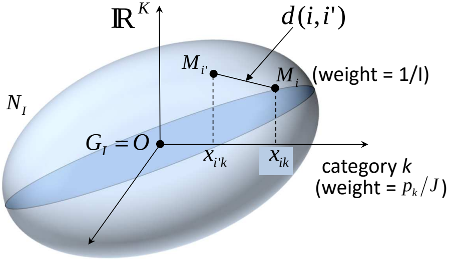


$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J} \left(x_{ik} - x_{i'k}\right)^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left(\frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k}\right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2$$

- 2 individuals with same categories : distance $= 0$
- 2 individuals with many shared categories : small distance
- 2 individuals, only 1 with a rare category : large distance to indicate this
- 2 individuals share rare category : small distance to indicate this shared specificity

Point cloud of individuals





$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{p_k}{J}\left(x_{ik} - x_{i'k}\right)^2 = \sum_{k=1}^{K} \frac{p_k}{J}\left(\frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k}\right)^2 = \frac{1}{J}\sum_{k=1}^{K}\frac{1}{p_k}(y_{ik} - y_{i'k})^2$$

$$d(i, G_I)^2 = \sum_{k=1}^{K} \frac{p_k}{J}(x_{ik})^2 = \sum_{k=1}^{K} \frac{p_k}{J}\left(\frac{y_{ik}}{p_k} - 1\right)^2 = \frac{1}{J}\sum_{k=1}^{K}\frac{y_{ij}}{p_k} - 1$$

$$\text{Inertia}(N_I) = \sum_{i=1}^{I} \underbrace{\frac{1}{I} d^2(i, O)}_{\text{inertia of } i} = \sum_{i=1}^{I}\left(\frac{1}{IJ}\sum_{k=1}^{K}\frac{y_{ik}}{p_k} - \frac{1}{I}\right) = \frac{K}{J} - 1$$

Building the point cloud of individuals

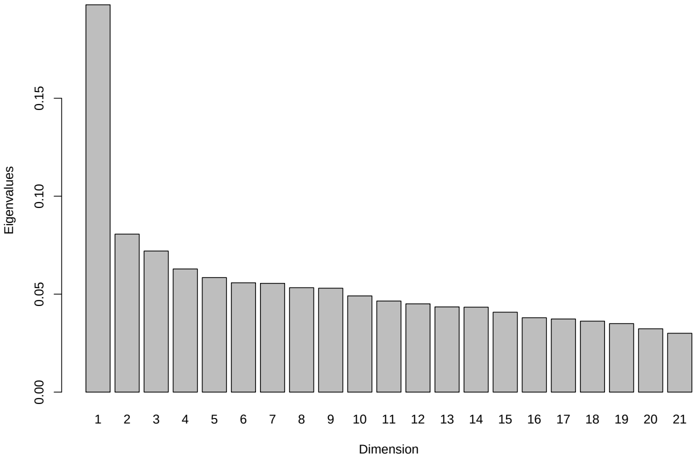
Getting factor axes, as usual, like for all factor analysis methods

Sequential construction : look for the axis maximizing the inertia and orthogonal to previous axes

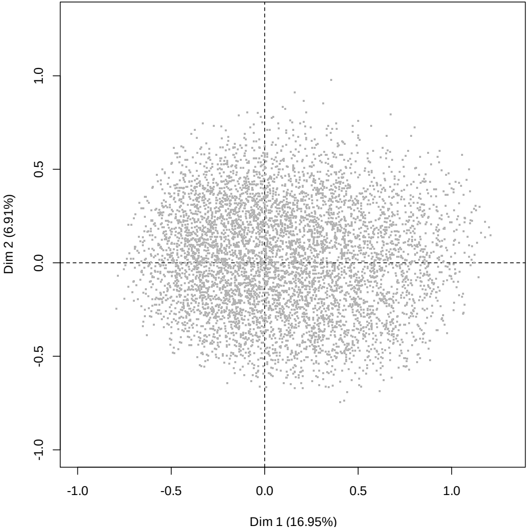
Leisure activity data

- Extract from 2003 INSEE survey on identity construction, called the "history of life" survey
- 8403 individuals
- 2 sorts of variables :
  - *Which of the following leisure activities do you practice regularly* : Reading, Listening to music, Cinema, Shows, Exhibitions, Computer, Sport, Walking, Travel, Playing a musical instrument, Collecting, Voluntary work, Home improvement, Gardening, Knitting, Cooking, Fishing, Number of hours of TV per day on average
  - supplementary variables (4 questions) : sex, gender, profession, marital status

Diagram showing the inertia

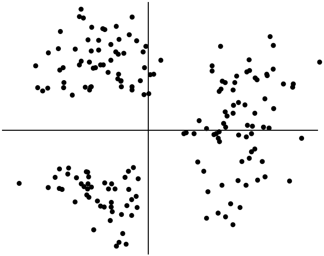


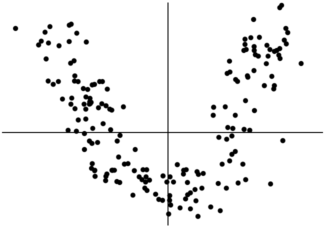
Representation of the point cloud of individuals



Representation of the point cloud of individuals

What kind of pattern might we see ?



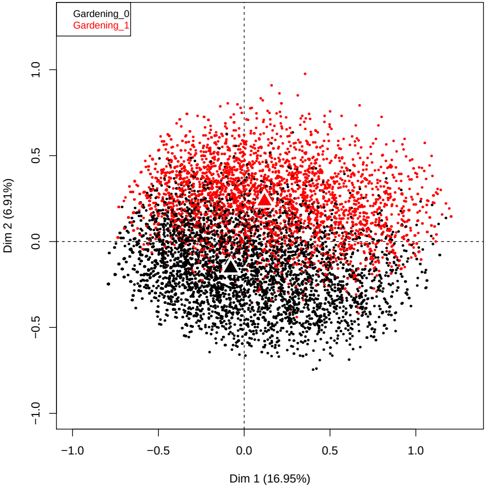


The Guttman effect

Individuals shown in terms of the gardening variable

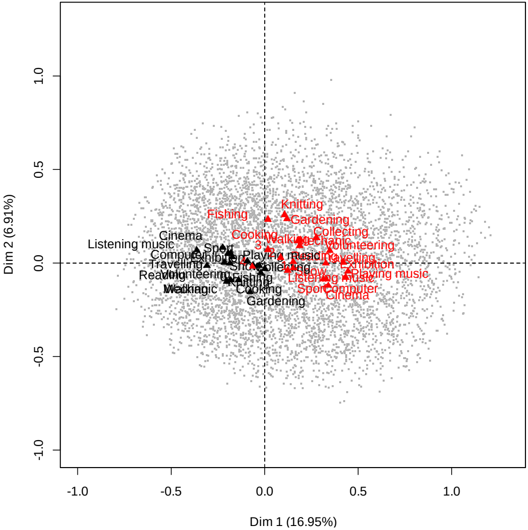
Idea : use the categories and variables to interpret the plot of the individuals

Put a category at the barycenter of the individuals in it



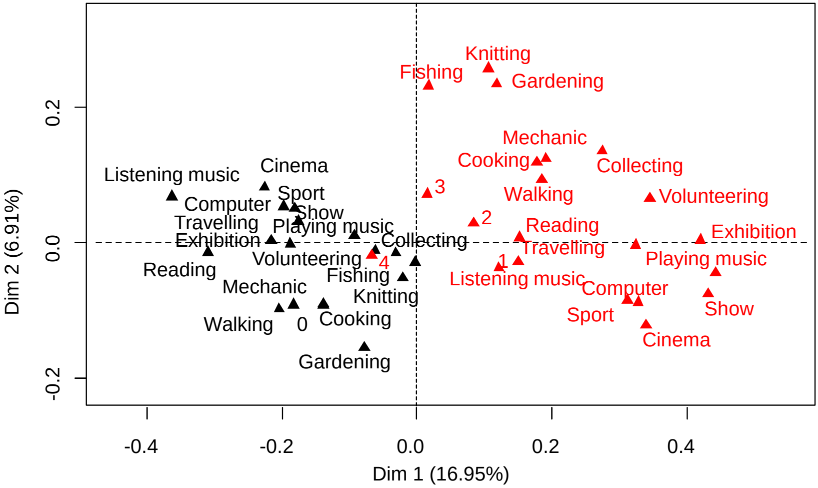
Showing the categories with the point cloud of individuals

Each category is at the barycenter of the individuals in it



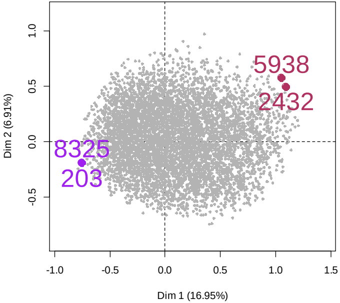
Activity not performed – activity performed

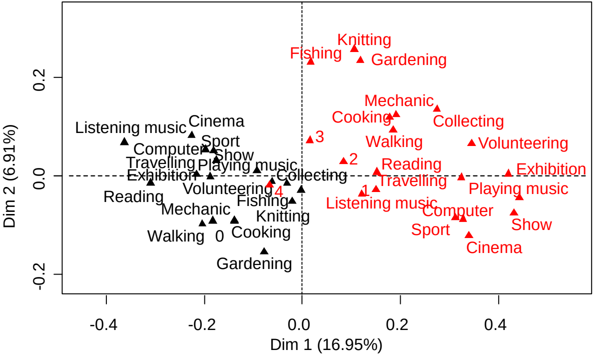
Showing the categories with the point cloud of individuals



Activity not performed – activity performed

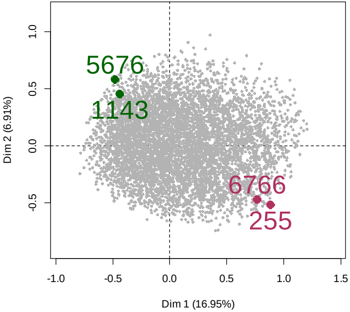
Showing the categories with the point cloud of individuals

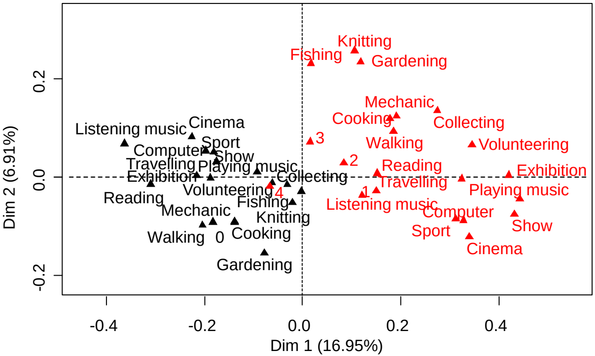




	Read	Listen music	Cinema	Show	Exhib	Comput	Sport	Walk	Travel	Play music	Collec	Volunteering	Mechanic	Garden	Knitt	Cook	Fish	TV
5938	y	y	n	y	y	y	y	y	y	y	y	y	y	y	y	y	n	3
2432	y	y	y	y	y	y	n	y	y	y	y	y	y	y	y	y	n	2
8325	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	4
203	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	4

Showing the categories with the point cloud of individuals





	Read	Listen music	Cinema	Show	Exhib	Comput	Sport	Walk	Travel	Play music	Collec	Volunteering	Mechanic	Garden	Knitt	Cook	Fish	TV
255	y	y	y	y	y	y	y	y	y	y	n	y	n	n	n	n	n	1
6766	y	y	y	y	y	y	y	y	y	y	n	n	n	n	n	y	n	0
5676	n	n	n	n	n	n	n	n	n	n	n	n	y	y	y	y	n	4
1143	y	n	n	n	n	n	n	n	n	n	n	n	y	y	y	n	n	4

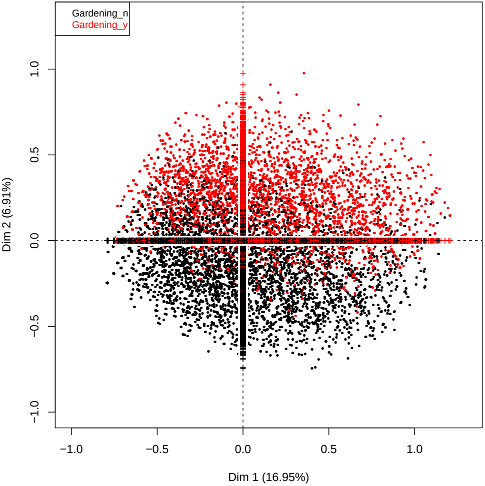
Showing the variables to help interpret the axes

Idea : look at coordinates of projected individuals on each axis, and calculate a value for the connection between these coordinates and each qualitative variable

Correlation ratio between the j-th variable and s-th component : $\eta(v_{.j}, F_s)$

$$\eta^2(F_2, Gardening) = 0.453$$

$$\eta^2(F_1, Gardening) = 0.047$$

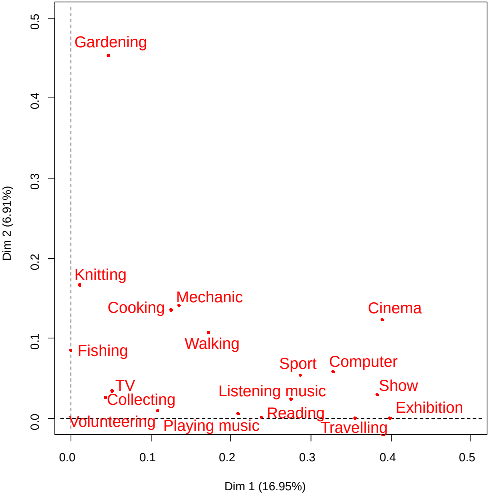


Showing the variables to help interpret the axes

Using the squared correlation ratios

The s-th axis is orthogonal to the t-th for all $t < s$, and the most related to the qualitative variables in the η^2 sense :

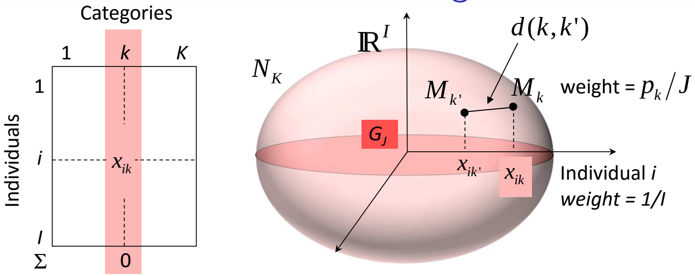
$$F_s = \max_F \sum_{j=1}^{J} \eta^2(F, v_{.j})$$



Plan

1. Data - issues
2. Studying the individuals
3. Studying the categories
4. Interpretation aids

Point cloud of categories



$$Var(k) = d^2(k,O) = \sum_{i=1}^{I} \frac{1}{I} x_{ik}^2 = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - 1 \right)^2 = \frac{1}{p_k} - 1$$

	p_k	1/2	1/5	1/10	1/101
	$d(k,O)$	1	2	3	10
(si $J = 10$)	$Inertia(k)$	0.05	0.08	0.09	0.099

$$Inertia(k) = \frac{p_k}{J} d^2(k,O) = \frac{1-p_k}{J}$$

$$d^2(k,k') = \sum_{i=1}^{I} \left(\frac{y_{ik}}{p_k} - \frac{y_{ik'}}{p_{k'}} \right)^2 = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}$$

Inertia of categories or variables

$$Inertia(k) = \frac{1 - p_k}{J}$$

$$Inertia(j) = \frac{1}{J} \sum_{k=1}^{K_j} (1 - p_k) = \frac{K_j - 1}{J}$$

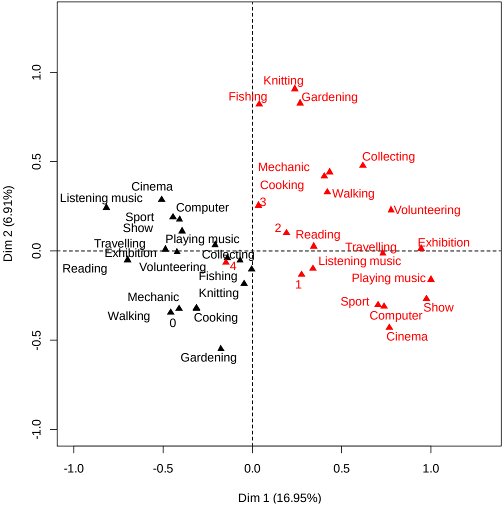
Variable	No. of categories	Inertia	No. dim. of subspace
sex	2	$1/J$	1
region	21	$20/J$	20
district	96	$95/J$	95

BUT : the inertia $\frac{K_j-1}{J}$ is spread across a $K_j - 1$ dim. subspace

$$Total\ inertia = \sum_{j=1}^{J} \frac{K_j - 1}{J} = \frac{K}{J} - 1$$

Representing the point cloud of categories

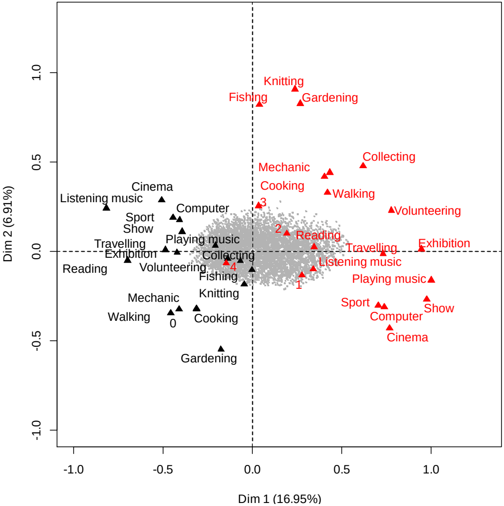
Sequential search for axes – as usual in factor analysis : each axis must maximize the inertia and be orthogonal to all previous ones



Activity not performed – activity performed

Projections of the individuals

Each individual put at barycenter of the categories they possess



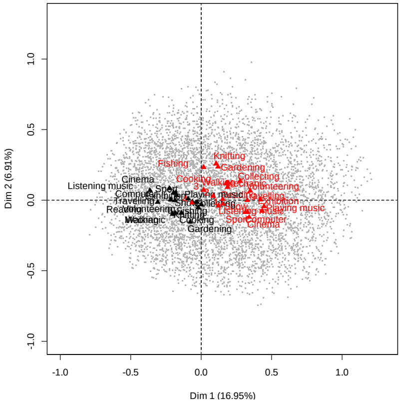
Barycentric representation – simultaneous representation

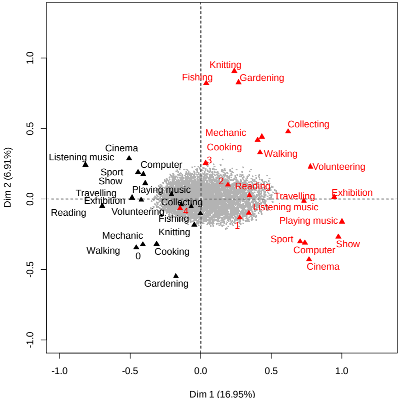
Optimal representation of individuals
Categories at the barycenter :

$$G_s(k) = \sum_{i=1}^{I} \frac{y_{ik}}{I_k} F_s(i)$$

Optimal representation of categories
Individuals at the barycenter :

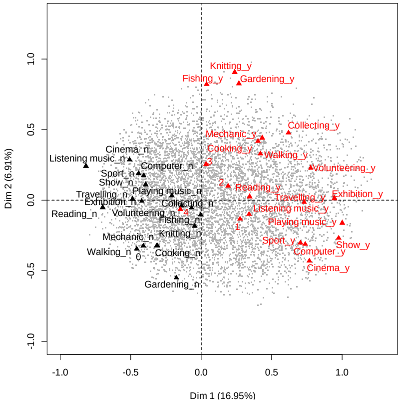
$$F_s(i) = \sum_{j=1}^{J} \frac{y_{ik}}{J} G_s(k)$$

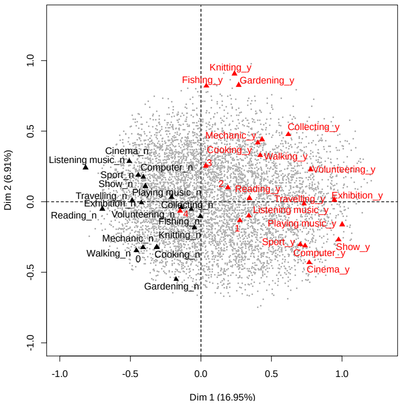




Barycentric representation – simultaneous representation

Optimal representation of individuals Categories at the pseudo-barycenter :	Optimal representation of categories Individuals at the pseudo-barycenter :
$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{y_{ik}}{I_k} F_s(i)$$	$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \frac{y_{ik}}{J} G_s(k)$$





Plan

1. Data - issues
2. Studying the individuals
3. Studying the categories
4. Interpretation aids

Inertia and percentage of inertia in MCA

$$\lambda_s = \frac{1}{J}\sum_{j=1}^{J} \eta^2(F_s, v_{.j})$$

$\Rightarrow$ λ_s is the mean of the squared correlation ratios

- Individuals live in $\mathbb{R}^{K-J}$ $\Rightarrow$ low percentages of inertia
- Maximal percentage for given axis s :

$$\begin{aligned}\frac{\lambda_s}{\sum_{t=1}^{K-J}\lambda_t}\times 100 \quad &\leq \quad \frac{1}{\frac{K-J}{J}}\times 100 \\ &\leq \quad \frac{J}{K-J}\times 100\end{aligned}$$

With $K = 100$, $J = 10$: $\lambda_s \leq 11.1$ %

- Mean of non-zero eigenvalues : $\frac{1}{K-J}\times\sum_t \lambda_t = \frac{1}{K-J}\times\left(\frac{K}{J}-1\right) = \frac{1}{J}$

  $\Rightarrow$ interpret the axes of inertia above $1/J$

Contributions and quality of representation

- Contributions and $\cos^2$ for individuals and categories

  $\Rightarrow$ distant categories don't necessarily contribute a lot (depends on their frequency)
  $\Rightarrow$ small $\cos^2$ as expected – many dimensions

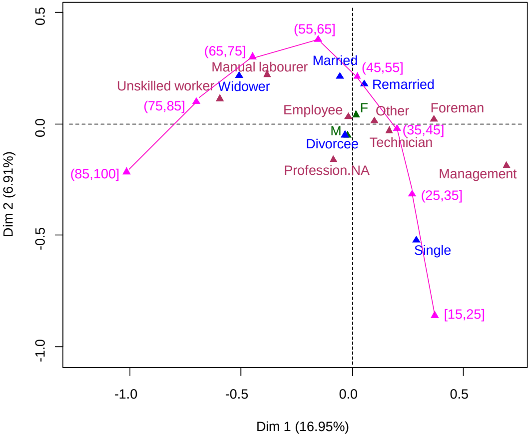
- Absolute contribution of a variable :

$$CTR(j) = \sum_{k=1}^{K_j} CTR(k) = \frac{\eta^2(F_s, v_{.j})}{J}$$

- Relative contribution : $CTR(j) = \frac{\eta^2(F_s, v_{.j})}{J\lambda_s}$

Representing supplementary elements

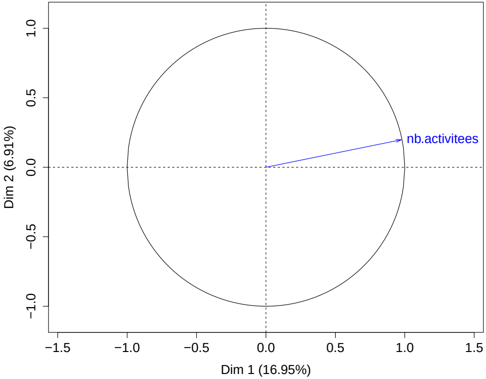
Use transition formulas to represent supplementary elements (individuals, variables, etc.)



Quantitative supplementary variables

⇒ What can we do with quantitative variables ?

- Supplementary information : project onto the axes, calculate correlation coefficients with each axis
- break up quantitative variable into categories/classes



Describing the axes

Using qualitative variables (Fisher test), using categories (Student test), using quantitative variables (correlations)

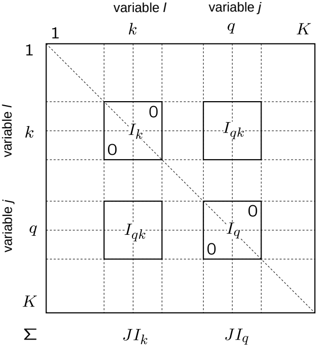
```
Quantitative variables
             correlation p.value
nb.activitees   0.9753459       0

           Categorical variables                        Categories
                      R2   p.value                  Estimate p.value
Reading            0.239  0.00e+00   Playing music_Y    0.268       0
Listening music 0.275  0.00e+00       Travelling_Y    0.270       0
Cinema             0.389  0.00e+00          Walking_Y    0.184       0
Show               0.383  0.00e+00            Sport_Y    0.247       0
Exhibition         0.399  0.00e+00         Computer_Y    0.263       0
Computer           0.327  0.00e+00       Exhibition_Y    0.304       0
Sport              0.287  0.00e+00             Show_Y    0.304       0
Walking            0.172  0.00e+00            Sport_N   -0.247       0
Travelling         0.355  0.00e+00         Computer_N   -0.263       0
Playing music      0.209  0.00e+00       Exhibition_N   -0.304       0
Mechanic           0.135 8.82e-267             Show_N   -0.304       0
Cooking            0.125 9.42e-247           Cinema_N   -0.283       0
Profession         0.128 7.20e-245 Listening music_N   -0.257       0
Volunteering       0.109 2.25e-212          Reading_N   -0.231       0
```

Different MCA strategy : Burt table

Burt table :

- Pairwise links between variables (like a correlation matrix between quantitative variables)
- Correspondence analysis on Burt table
- Gives results uniquely for categories : same representation but different eigenvalues : $\lambda_s^{Burt} = (\lambda_s^{TDC})^2$
- λ_s^{TDC} mean of squared correlation ratios

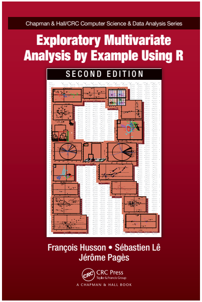


⇒ The MCA only depends on pairwise links between variables (just like PCA only depends on the correlation matrix)

Conclusion

- MCA is the best factor analysis method for tables of individuals with qualitative variables
- Eigenvalues represent the means of squared correlation ratios
- The values of these squared links are particularly important when there are lots of variables
- Return to the data by analyzing the contingency table with CA
- Convergence of CDT analysis and Burt table analysis is a strong argument in favor of the general method
- MCA can be use to pre-treat data before doing classification

Extras



Husson F., Lê S. & Pagès J. (2017)
Exploratory Multivariate Analysis by Example Using R
2nd edition, 230 p., CRC/Press.

The `FactoMineR` package for running MCA :
`http://factominer.free.fr`

Videos on Youtube :

- Youtube channel : `youtube.com/HussonFrancois`
- video playlist in English
- video playlist in French