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

\[ K_j = 5 \]

\[ i' \]

\[ j \]

\[ J \]

\[ 1 \]

\[ I \]

\[ i \]

\[ v_{ij} \]

\[ i' \]

\[ j' \]

\[ married \]

Categories

\[ j \]

\[ J \]

\[ 1 \]

\[ k \]

\[ K_j \]

\[ K \]

\[ \Sigma \]

\[ 1 \]

\[ I \]

\[ i \]

\[ y_{ik} \]

\[ 01000 \]

\[ j' \]

\[ J \]

\[ 1 \]

\[ I \]

\[ I \]

\[ I \]

\[ J \]

Indicator Matrix

\[ Ip_k \]
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)

⇒ 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

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

### Sociodemographic variables

<table>
<thead>
<tr>
<th>Sex</th>
<th>Female</th>
<th>4616</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>3787</td>
</tr>
<tr>
<td>Age</td>
<td>[15,25]</td>
<td>857</td>
</tr>
<tr>
<td></td>
<td>(25,35]</td>
<td>1302</td>
</tr>
<tr>
<td></td>
<td>(35,45]</td>
<td>1646</td>
</tr>
<tr>
<td></td>
<td>(45,55]</td>
<td>1837</td>
</tr>
<tr>
<td></td>
<td>(55,65]</td>
<td>1257</td>
</tr>
<tr>
<td></td>
<td>(65,75]</td>
<td>937</td>
</tr>
<tr>
<td></td>
<td>(75,85]</td>
<td>482</td>
</tr>
<tr>
<td></td>
<td>(85,100]</td>
<td>85</td>
</tr>
<tr>
<td>Marital status</td>
<td>Divorcee</td>
<td>792</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>4333</td>
</tr>
<tr>
<td></td>
<td>Remarried</td>
<td>404</td>
</tr>
<tr>
<td></td>
<td>Single</td>
<td>2140</td>
</tr>
<tr>
<td></td>
<td>Widower</td>
<td>734</td>
</tr>
<tr>
<td>Profession</td>
<td>employee</td>
<td>2552</td>
</tr>
<tr>
<td></td>
<td>foreman</td>
<td>735</td>
</tr>
<tr>
<td></td>
<td>management</td>
<td>1052</td>
</tr>
<tr>
<td></td>
<td>manual labourer</td>
<td>1161</td>
</tr>
<tr>
<td></td>
<td>technician</td>
<td>401</td>
</tr>
<tr>
<td></td>
<td>unskilled worker</td>
<td>792</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>No answer</td>
<td>1498</td>
</tr>
</tbody>
</table>
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 \( \frac{1}{l} \)

\( y_{ik} = 1 \) if the \( i \)-th individual is in \( k \)-th category of the \( j \)-th variable (for each \( p_k \))

\( = 0 \) otherwise

Idea: \( x_{ik} = y_{ik}/p_k \)

\[
\frac{\sum_{i=1}^{l} x_{ik}}{l} = \frac{1}{l} \sum_{i=1}^{l} y_{ik} = \frac{1}{l} l \times p_k = 1
\]

Centering: \( x_{ik} = y_{ik}/p_k - 1 \)
Plan

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

Studying the individuals

Point cloud of individuals

Studying the categories

Interpretation aids

**Inertia**

\[
\text{inertia of } i = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik} - \bar{x}_{ik})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left( \frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k} \right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} (y_{ik} - y_{i'k})^2
\]

- 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**

- **Indicators matrix:**
  - **Categories:** $k \rightarrow K$
  - **Individuals:** $i \rightarrow I$
  - **Category:** $x_{ik}$

- **Inertia ($N_I$):**
  - $\text{Inertia} = \sum_{i=1}^{I} \frac{1}{l} d^2(i, G_I) = \sum_{i=1}^{I} \left( \frac{1}{lJ} \sum_{k=1}^{K} \frac{y_{ik}}{p_k} - \frac{1}{l} \right) = \frac{K}{J} - 1$

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

- **Distance ($d(i, i')$):**
  - $d_{i,i'}^2 = \sum_{k=1}^{K} \frac{p_k}{J} (x_{ik} - x_{i'k})^2 = \sum_{k=1}^{K} \frac{p_k}{J} \left( \frac{y_{ik}}{p_k} - \frac{y_{i'k}}{p_k} \right)^2 = \frac{1}{J} \sum_{k=1}^{K} \frac{1}{p_k} \left( y_{ik} - y_{i'k} \right)^2$
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
Showing the categories with the point cloud of individuals
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 $\eta^2$ 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

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

\[
p_k = \frac{1}{2} \quad \frac{1}{5} \quad \frac{1}{10} \quad \frac{1}{101}
\]

\[
d(k, O) = \begin{pmatrix} 1 & 2 & 3 & 10 \\ si J = 10 \end{pmatrix}
\]

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

\[
d^2(k, k') = \sum_{i=1}^{l} \left( \frac{y_{ik}}{p_k} - \frac{y_{ik'}}{p_{k'}} \right)^2 = \frac{p_k + p_{k'} - 2p_{kk'}}{p_k p_{k'}}
\]
Inertia of categories or variables

\[ \text{Inertia}(k) = \frac{1 - p_k}{J} \]

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>No. of categories</th>
<th>Inertia</th>
<th>No. dim. of subspace</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>2</td>
<td>1/J</td>
<td>1</td>
</tr>
<tr>
<td>region</td>
<td>21</td>
<td>20/J</td>
<td>20</td>
</tr>
<tr>
<td>district</td>
<td>96</td>
<td>95/J</td>
<td>95</td>
</tr>
</tbody>
</table>

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

\[ \text{Total inertia} = \sum_{j=1}^{J} \frac{K_j - 1}{J} = \frac{K}{J} - 1 \]
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**:

\[
G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{y_{ik}}{I_k} F_s(i)
\]

Optimal representation of categories

Individuals at the **pseudo-barycenter**:

\[
F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \frac{y_{ik}}{J_k} G_s(k)
\]
Plan

1. Data - issues

2. Studying the individuals

3. Studying the categories

4. Interpretation aids
Inertia and percentage of inertia in MCA

\[ \lambda_s = \frac{1}{J} \sum_{j=1}^{J} \eta^2(F_s, v_j) \]

\[ \Rightarrow \lambda_s \text{ is the mean of the squared correlation ratios} \]

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

\[ \frac{\lambda_s}{\sum_{t=1}^{K-J} \lambda_t} \times 100 \leq \frac{1}{K-J} \times 100 \]
\[ \leq \frac{J}{K-J} \times 100 \]

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

- Mean of non-zero eigenvalues :

\[ \frac{1}{K-J} \times \sum_{t} \lambda_t = \frac{1}{K-J} \times \left( \frac{K}{J} - 1 \right) = \frac{1}{J} \]

\[ \Rightarrow \text{ interpret the axes of inertia above } 1/J \]
Contributions and quality of representation

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

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

- Absolute contribution of a variable :

  \[
  CTR(j) = \sum_{k=1}^{K_j} CTR(k) = \frac{\eta^2(F_s, \nu_j)}{J}
  \]

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

<table>
<thead>
<tr>
<th>correlation p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb.activitiees 0.9753459 0</td>
</tr>
</tbody>
</table>

Categorical variables

<table>
<thead>
<tr>
<th>Categories</th>
<th>R2</th>
<th>p.value</th>
<th>Estimate</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Playing music_Y</td>
<td>0.239</td>
<td>0.00e+00</td>
<td>0.268</td>
<td>0</td>
</tr>
<tr>
<td>Travelling_Y</td>
<td>0.275</td>
<td>0.00e+00</td>
<td>0.270</td>
<td>0</td>
</tr>
<tr>
<td>Walking_Y</td>
<td>0.389</td>
<td>0.00e+00</td>
<td>0.184</td>
<td>0</td>
</tr>
<tr>
<td>Sport_Y</td>
<td>0.383</td>
<td>0.00e+00</td>
<td>0.247</td>
<td>0</td>
</tr>
<tr>
<td>Computer_Y</td>
<td>0.399</td>
<td>0.00e+00</td>
<td>0.263</td>
<td>0</td>
</tr>
<tr>
<td>Exhibition_Y</td>
<td>0.327</td>
<td>0.00e+00</td>
<td>0.304</td>
<td>0</td>
</tr>
<tr>
<td>Show_Y</td>
<td>0.287</td>
<td>0.00e+00</td>
<td>0.304</td>
<td>0</td>
</tr>
<tr>
<td>Sport_N</td>
<td>0.172</td>
<td>0.00e+00</td>
<td>-0.247</td>
<td>0</td>
</tr>
<tr>
<td>Computer_N</td>
<td>0.355</td>
<td>0.00e+00</td>
<td>-0.263</td>
<td>0</td>
</tr>
<tr>
<td>Exhibition_N</td>
<td>0.209</td>
<td>0.00e+00</td>
<td>-0.304</td>
<td>0</td>
</tr>
<tr>
<td>Show_N</td>
<td>0.135</td>
<td>8.82e-267</td>
<td>-0.304</td>
<td>0</td>
</tr>
<tr>
<td>Cinema_N</td>
<td>0.125</td>
<td>9.42e-247</td>
<td>-0.283</td>
<td>0</td>
</tr>
<tr>
<td>Listening music_N</td>
<td>0.128</td>
<td>7.20e-245</td>
<td>-0.257</td>
<td>0</td>
</tr>
<tr>
<td>Reading_N</td>
<td>0.109</td>
<td>2.25e-212</td>
<td>-0.231</td>
<td>0</td>
</tr>
</tbody>
</table>
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: $\lambda_s^{Burt} = (\lambda_s^{TDC})^2$
- $\lambda_s^{TDC}$ 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
*Exploratory Multivariate Analysis by Example Using R*

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