LAB: Correspondance Analysis

2025-03-16

M1 MIDS/MFA/LOGOS
Université Paris Cité
Année 2024
Course Homepage
Moodle

Besides the usual packages (`tidyverse`, …), we shall require `FactoMineR` and related packages.

```
stopifnot(
  require(FactoMineR),
  require(factoextra),
  require(FactoInvestigate)
)
```

Correspondence Analysis

The `mortality` dataset

The goal is to investigate a possible link between age group and cause of death. We work with dataset `mortality` from package `FactoMineR`

```
data("mortality", package = "FactoMineR")
```

```
#help(mortality)
```

> A data frame with 62 rows (the different causes of death) and 18 columns. Each column corresponds to an age interval (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85-94, 95 and more) in a year. The 9 first columns correspond to data in 1979 and the 9 last columns to data in 2006. In each cell, the counts of deaths for a cause of death in an age interval (in a year) is given.

Source Centre d'épidemiologie sur les causes de décès médicales

See also EuroStat:

- Causes of death (hlth_cdeath) Reference Metadata in Single Integrated Metadata Structure (SIMS)
-

Question

Read the documentation of the `mortality` dataset. Is this a sample? an aggregated dataset?
If you consider `mortality` as an agregated dataset, can you figure out the organization of the sample `mortality` was built from?

Elementary statistics and table wrangling

Before proceeding to Correspondence Analysis (CA), let us draw some elementary plots.

Question

- Start by partially *pivoting* `mortality`, so as to obtain a tibble with columns `cause`, `year`, while keeping all columns named after age groups (tidy up the data so as to obtain a tibble in partially long format).
- Use `rowwise()` and `sum(c_cross())` so as to compute the total number of deaths per `year` and `cause` in column `total`. This allows to mimic `rowSums()` inside a pipeline. Column `grand_total` is computed using a *window* function over grouping by `cause`.

Question

Build a bar plot to display the importance of causes of deaths in France in years 1979 and 2006

Question

Compute and display the total number of deaths in France in years 1979 and 2006.

Question

Compute the marginal counts for each year (1979, 2006). Compare.

Correspondance Analysis

CA executive summary

- Start from a 2-way contingency table X with $\sum_{i,j} X_{i,j} = N$
- Normalize $P = \frac{1}{N} X$ (*correspondance matrix*)
- Let r (resp. c) be the row (resp. column) wise sums vector
- Let $D_r = \text{diag}(r)$ denote the diagonal matrix with row sums of P as coefficients
- Let $D_c = \text{diag}(c)$ denote the diagonal matrix with column sums of P as coefficients
- The *row profiles matrix* is $D_r^{-1} \times P$
- The *standardized residuals matrix* is $S = D_r^{-1/2} \times (P - rc^T) \times D_c^{-1/2}$

CA consists in computing the SVD of the standardized residuals matrix $S = U \times D \times V^T$

From the SVD, we get - $D_r^{-1/2} \times U$ *standardized coordinates of rows* - $D_c^{-1/2} \times V$ *standardized coordinates of columns* - $D_r^{-1/2} \times U \times D$ *principal coordinates of rows* - $D_c^{-1/2} \times V \times D$ *principal coordinates of columns* - Squared singular values: the principal *inertia*

When calling `svd(.)`, the argument should be

$$D_r^{1/2} \times (D_r^{-1} \times P \times D_c^{-1} - \mathbf{I} \times \mathbf{I}^T) \times D_c^{1/2}$$

CA and extended SVD

As

$$D_r^{-1} \times P \times D_c^{-1} - \mathbf{I}\mathbf{I}^T = (D_r^{-1/2} \times U) \times D \times (D_c^{-1/2} \times V)^T$$

$(D_r^{-1/2} \times U) \times D \times (D_c^{-1/2} \times V)^T$ is the *extended SVD* of

$$D_r^{-1} \times P \times D_c^{-1} - \mathbf{I}\mathbf{I}^T$$

with respect to D_r and D_c

Question

Perform CA on the two contingency tables.

You may use `FactoMineR::CA()`. It is interesting to compute the correspondence analysis in your own way, by preparing the matrix that is handled to `svd()` and returning a named list containing all relevant information.

> Do the Jedi and Sith build their own light sabers? Jedi do. It's a key part of the religion to have a kyber crystal close to you, to build the saber through the power of the force creating a blade unique and in tune with them

Question

If you did use `FactoMineR::CA()`, explain the organization of the result.

Screeplots

ℹ **Question**

Draw screeplots. Why are they useful? Comment briefly.

Row profiles analysis

ℹ **Question**

Perform row profiles analysis.
What are the classical plots? How can you build them from the output of `FactoMiner::CA`?
Build the table of row contributions (the so-called cos^2)

ℹ **Question**

Plot the result of row profile analysis using `plot.CA` from `FactoMineR`.

ℹ **Question**

Perform column profiles analysis

Symmetric plots

ℹ **Question**

Build the symmetric plots (biplots) for correspondence analysis of Mortalitity data

Mosaicplots

ℹ **Question**

Mosaic plots provide an alternative way of exploring contingency tables. They are particularly handy when handling 2-way contingency tables.
Draw mosaic plots for the two contingency tables living inside `mortality` datasets.

ℹ **Question**

Are you able to deliver an interpretation of this Correspondence Analysis?

Hierarchical clusetring of row profiles

ℹ **Question**

Build the standardized matrix for row profiles analysis. Compute the pairwise distance matrix using the χ^2 distances. Should you work centered row profiles?

ℹ

Question

Perform hierarchical clustering of row profiles with method/linkage `"single"`. Check the definition of the method. Did you know the underlying algorithm? If yes, in which context did you get acquainted with this algorithm?

Question

Choose the number of classes (provide justification).

Question

Can you explain the size of the different classes in the partition?

Atypical row profiles

Question

Row profiles that do not belong to the majority class are called *atypical*.

1. Compute the share of inertia of atypical row profiles.
2. Draw a symmetric plot (biplot) outlining the atypical row profiles.

Investigating independence/association

Question

1. Calculate the theoretical population table for `deces`. Do you possible to carry out a chi-squared test?
2. Perform a hierarchical classification of the line profiles into two classes.
3. Merge the rows of `deces` corresponding to the same class (you can use the the `tapply` function), and perform a chi-square test. chi-square test. What's the conclusion?
4. Why is it more advantageous to carry out this grouping into two classes compared to arbitrarily grouping two classes, in order to prove the dependence between these two variables?

About the "average profile"

Question

1. Represent individuals from the majority class. Do they all seem to you to correspond to an average profile?
2. Try to explain this phenomenon considering the way in which hierarchical classification uses the Single Linkage method.

Caveat

The `mortality` dataset should be taken with grain of salt. Assigning a single *cause* to every death is not a trivial task. It is even questionable: if somebody dies from some infection because she could not be cured using an available drug due to another preexisting pathology, who is the culprit?