Exploration Statistique Multidimensionnelle

Correspondence analysis

PHILIPPE BESSE & OLIVIER ROUSTANT

INSA de Toulouse
Institut de Mathématiques

Motivation

Question

Question : given two qualitative variables, how analyzing and visualizing the correspondence between their levels ?

Example - Correspondence of clustering results

	hc1	hc2	hc3	hc4	hc5	hc6
km1	17	19	0	129	46	0
km2	0	31	40	0	0	0
km3	1	94	15	4	0	0
km4	0	14	0	3	277	0
km5	135	40	0	7	0	1
km6	0	13	136	4	20	0
km7	10	34	0	0	0	99

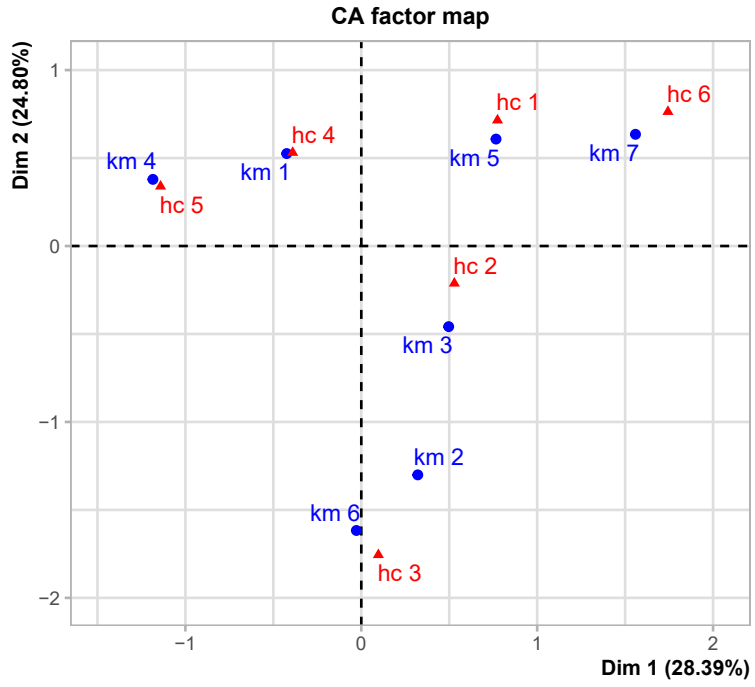


Figure – Left : contingency table between classes found by hierarchical clustering and k-means. Right : output of correspondence analysis.

Example 2 - Sociological data

	EAG	PT	PLCS	CM	EMP	AUT
DR	86	168	470	236	161	305
SCE	38	74	191	99	58	115
LET	149	312	806	493	308	624
SC	105	137	400	264	144	247
MD	98	261	1040	337	175	348
PD	12	21	45	36	22	42
IUT	62	62	79	87	62	90

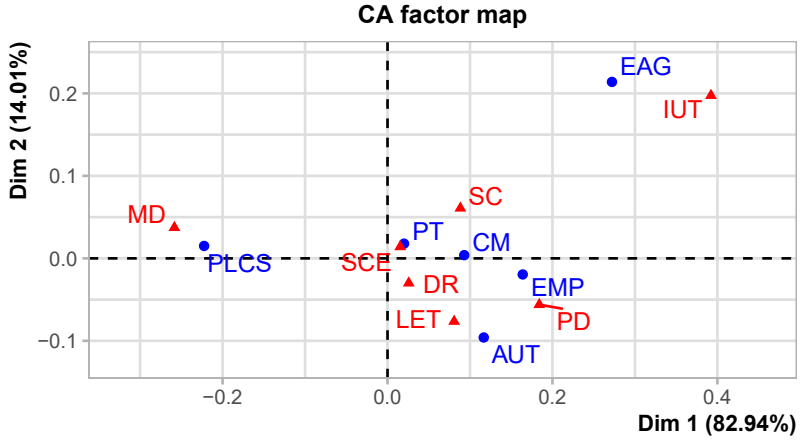


Figure – Left : contingency table between father job and children study type. Right : output of correspondence analysis.

Contingency table

- Two qualitative variables observed on n individuals
- X, with r levels $x_1, \ldots, x_\ell, \ldots, x_r$
- Y, with c levels $y_1, \ldots, y_h, \ldots, y_c$
- $\mathbf{T}$: contingency table $r \times c$

	y_1	$\cdots$	y_h	$\cdots$	y_c	sums
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sums	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Dependence

- X and Y are not linked relatively to T if and only if :

$$n_{\ell h} = \frac{n_{\ell +} n_{+h}}{n} \qquad \forall (\ell, h) \in [\![1, r]\!] \times [\![1, c]\!]$$

- Equivalent to probability independence
- χ^2 test, whose statistics

$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{\left(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n}\right)^2}{\frac{n_{\ell +} n_{+h}}{n}};$$

follows asymptotically χ^2 with $(r-1)(c-1)$ d.f.

One contingency table, two dual datasets

- A first dataset is given by the row profiles, associated to the conditional probability $P(Y = .|X = x_\ell)$:

	y_1	$\cdots$	y_h	$\cdots$	y_c	sums
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$\frac{n_{\ell 1}}{n_{\ell +}}$	$\cdots$	$\frac{n_{\ell h}}{n_{\ell +}}$	$\cdots$	$\frac{n_{\ell c}}{n_{\ell +}}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{g}_c$	$\frac{n_{+1}}{n}$	$\cdots$	$\frac{n_{+h,}}{n}$	$\cdots$	$\frac{n_{+c}}{n}$	1

- The ℓ-th row profile has weight $f_{\ell +} = \frac{n_{\ell +}}{n}$, ass. to $P(X = x_\ell)$.
- The mean row profile $\mathbf{g}_c$, asso. to $P(Y = .)$, is their centroid :

$$\sum_{\ell=1}^{r} \left(\frac{n_{\ell +}}{n}\right) \frac{n_{\ell h}}{n_{\ell +}} = \sum_{\ell=1}^{r} \frac{n_{\ell h}}{n} = \frac{n_{+h}}{n} = (\mathbf{g}_c)_h$$

One contingency table, two dual datasets

- A second dataset is given by the column profiles, associated to the conditional probability $P(X = .|Y = y_h)$:

	y_1	$\cdots$	y_h	$\cdots$	y_c	$\mathbf{g}_r$
x_1	$\frac{n_{11}}{n_{+1}}$	$\cdots$	$\frac{n_{1h}}{n_{+h}}$	$\cdots$	$\frac{n_{1c}}{n_{+c}}$	$\frac{n_{1+}}{n}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$\frac{n_{\ell 1}}{n_{+1}}$	$\cdots$	$\frac{n_{\ell h}}{n_{+h}}$	$\cdots$	$\frac{n_{\ell c}}{n_{+h}}$	$\frac{n_{\ell +}}{n}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	$\frac{n_{r1}}{n_{+1}}$	$\cdots$	$\frac{n_{rh}}{n_{+h}}$	$\cdots$	$\frac{n_{rc}}{n_{+r}}$	$\frac{n_{r+}}{n}$
sums	1	$\cdots$	1	$\cdots$	1	1

- The h-th column profile has weight $f_{+h} = \frac{n_{+h}}{n}$, ass. to $P(Y = y_h)$.
- The mean profile $\mathbf{g}_r$, asso. to $P(X = .)$, is their centroid.

Example

Question extracted from a quizz :

Question 10 We give the following contingency table corresponding to two qualitative variables 'day', 'night', with levels '-', '=', '+':

	night-	night=	night+
day-	2	0	0
day=	1	1	0
day+	1	1	3

The row profile for 'day=' is

☐ $(4, 2, 3)$ ☐ $(1/2, 1/2, 0)$ ☐ $(1/9, 1/9, 0)$ ☐ $(1, 1, 0)$

Features of Correspondence Analysis (CA)

- CA aims at visualizing the correspondences that explain the χ^2 value
- CA provides a double principal component analysis : one for row profiles + one column profiles, with the χ^2 metric

Notations

- $\mathbf{T}$: the initial contingency table

Row profiles

- $\mathbf{a}^{\ell}$: the ℓ-th row profile
- $\mathbf{A} = [\mathbf{a}^1, \ldots, \mathbf{a}^r]$: matrix of row profiles (placed in columns)
- $\mathbf{D}_r = \text{diag}(f_{1+}, \ldots, f_{r+})$, matrix of row profile weights

Column profiles

- $\mathbf{b}^{h}$: the h-th column profile
- $\mathbf{B} = [\mathbf{b}^1, \ldots, \mathbf{b}^c]$: matrix of column profiles
- $\mathbf{D}_c = \text{diag}(f_{+1}, \ldots, f_{+c})$, matrix of column profile weights

The chi2 metric for variables

- The dataset of row (resp. column) profiles in $\mathbb{R}^c$ (resp. $\mathbb{R}^r$) is endowed with the χ^2 metric : $\mathbf{D}_c^{-1}$ (resp. $\mathbf{D}_r^{-1}$).
- Distance between levels :

$$\|\mathbf{a}^\ell - \mathbf{a}^i\|^2_{\mathbf{D}_c^{-1}} = \sum_{h=1}^{c} \frac{1}{f_{+h}} (a_h^\ell - a_h^i)^2$$

- The χ^2 metric is a weighted Euclidean norm, which gives more importance to levels with small frequencies.

Example

Question extracted from a quizz (following the last one) :

Question 10 We give the following contingency table corresponding to two qualitative variables 'day', 'night', with levels '-', '=', '+':

	night-	night=	night+
day-	2	0	0
day=	1	1	0
day+	1	1	3

The row profile for 'day=' is

☐ $(4, 2, 3)$ ☐ $(1/2, 1/2, 0)$ ☐ $(1/9, 1/9, 0)$ ☐ $(1, 1, 0)$

Question 11 Continuing the example of question 10: with the usual metric of correspondence analysis (cf question 9), the square distance between the row profiles of 'day-' and 'day=' is:

☐ 2 ☐ 27/4 ☐ 27/16 ☐ 1/2

Two dataset, one inertia

The inertia of the row profiles and column profiles are both equal to $\Phi^2 = \frac{\chi^2}{n}$.

$$
\begin{aligned}
I_r &= \sum_{\ell=1}^{r} f_{\ell+} \|\mathbf{a}^\ell - \mathbf{g}_c\|^2_{\mathbf{D}_c^{-1}} \\
&= \text{... (left to exercice) ...} \\
&= \sum_{\ell=1}^{r}\sum_{h=1}^{c} \frac{1}{n_{\ell+}n_{+h}} \left(n_{\ell h} - \frac{n_{\ell+}n_{+h}}{n}\right)^2 = \frac{\chi^2}{n}
\end{aligned}
$$

Theorem

- The PCA of $\mathbf{A}'$ with metric $\mathbf{D}_c^{-1}$ (var.) and weights $\mathbf{D}_r$ (ind.) is given by the spectral decomposition of $\mathbf{AB}$.
- The PCA of $\mathbf{B}'$ with metric $\mathbf{D}_r^{-1}$ (var.) and weights $\mathbf{D}_c$ (ind.) is given by the spectral decomposition of $\mathbf{BA}$.
- The two PCA give the same inertia decomposition,

$$\Phi^2 = \lambda_1 + \cdots + \lambda_{\min(r-1,c-1)}$$

  and all the eigenvalues belong to $[0, 1]$.
- An eigenvalue of 1 corresponds to a perfect link between one level of X and one level of Y (block diagonal form of $\mathbf{T}$)

Proof steps (exercise !)

For the spectral decomposition (column profile case) :

1. The covariance matrix is :
   $\Gamma = \mathbf{B}\mathbf{D}_c\mathbf{B}' - \mathbf{g}_r\mathbf{1}_c'\mathbf{D}_c(\mathbf{g}_r 1_c')' = \mathbf{B}\mathbf{D}_c\mathbf{B}' - \mathbf{g}_r\mathbf{g}_r'$
2. Using the isometry $\|x\|_{\mathbf{D_r}^{-1}} = \|\mathbf{D_r}^{-1/2}x\|$, the matrix to diagonalize is $\mathbf{D_r}^{-1/2}\Gamma\mathbf{D_r}^{-1/2}$
3. (x, λ) eig. for $\mathbf{D_r}^{-1/2}\Gamma\mathbf{D_r}^{-1/2}$ iff $(\mathbf{D}_r^{1/2}x, \lambda)$ eig. for $\Gamma\mathbf{D}_r^{-1}$
4. Using $\mathbf{B}\mathbf{D}_c = \mathbf{D}_r\mathbf{A}' = \frac{\mathbf{T}}{n}$, we have : $\Gamma\mathbf{D}_r^{-1} = \mathbf{B}\mathbf{A} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$
5. Using the relations $\mathbf{A}\mathbf{g}_r = \mathbf{g}_c$, $\mathbf{B}\mathbf{g}_c = \mathbf{g}_r$, $\mathbf{g}_r'\mathbf{D}_r^{-1}\mathbf{g}_r = 1$, see that $\mathbf{g}_r$ is an eigenvector of $\Gamma\mathbf{D}_r^{-1}$ with eigenvalue 0.
6. The spectral decomposition of $\Gamma\mathbf{D}_r^{-1}$ is the same as $\mathbf{B}\mathbf{A}$, except for $\mathbf{g}_r$, eigenvector assoc. to 1 for $\mathbf{B}\mathbf{A}$ and 0 for $\Gamma\mathbf{D}_r^{-1}$.

.../...

Proof steps (exercise !)

For the properties of the eigenvalues :

1. The eigenvalues of $\mathbf{BA}$ are those of $\mathbf{D_r}^{-1/2}\Gamma\mathbf{D_r}^{-1/2}$, which is symmetric, positive semi-definite $\rightarrow$ they belong to $\mathbb{R}_+$
2. $\mathbf{A}', \mathbf{B}'$ are stochastic matrices (coef. ≥ 0, rows sum to 1)
3. A product of stochastic matrices is stochastic
4. The eigenvalues of a stochastic matrix are of modulus ≤ 1.

Theory says that our 2 clouds (row & column profiles) :

- have the same inertia (χ^2/n)
  $\rightarrow$ same volume
- have the same projected inertia onto the first $\min(r-1, c-1)$ PCA axis (common eigenvalues)
  $\rightarrow$ same shape

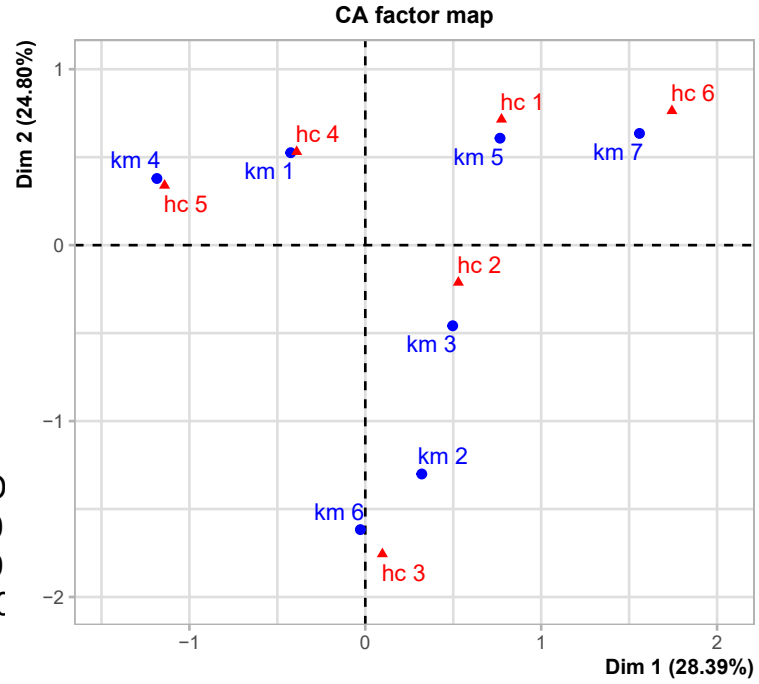
Thus, intuitively, they are similar up to some rotation.
Representing the results in the first PCA coordinates visualize correspondence between them.

Interpretation, come-back to example 1

	hc1	hc2	hc3	hc4	hc5	hc6
km1	17	19	0	129	46	0
km2	0	31	40	0	0	0
km3	1	94	15	4	0	0
km4	0	14	0	3	277	0
km5	135	40	0	7	0	1
km6	0	13	136	4	20	0
km7	10	34	0	0	0	99

Eigenvalues

	Dim.1	Dim.2	Dim.3	D
Variance	0.791	0.691	0.617	0
% of var. 28.386	24.798	22.124	15.594	9
Cumulative % of var.	28.386	53.184	75.309	9(



Interpretation is similar to PCA with the following specificities :

- Eigenvalues close to 1 indicate a strong link between levels
- Percentage of inertia corresponds to part of explained χ^2
- Each PCA plot shows the (2D projected) profiles which are close/far to each other in terms of χ^2 metric

  Ex : $\{km2, km6\}$ or $\{hc3, hc1\}$
- The biplot shows the correspondence between the levels of X and the levels of $Y \rightarrow$ explain which pairs are responsible for the high values of χ^2.

  Ex : $\{km4, hc5\}$