Exploration Statistique Multidimensionnelle

Multiple Correspondence Analysis

Olivier Roustant & Philippe Besse

INSA de Toulouse
Institut de Mathématiques

Main features

- Generalization of CA to more than 2 variables
- Representations of the correspondences between levels
- Methodology : from $p = 2$ to $p > 2$

Utilization

- Data analysis in presence of qualitative data
- MCA provides a particular PCA for qualitative data
  $\rightarrow$ can be used for clustering (on the PCA coordinates)

The variables in Velib are both quantitative & qualitative.
It is easy to convert them to qualitative variables only,
by discretizating the numerical values.

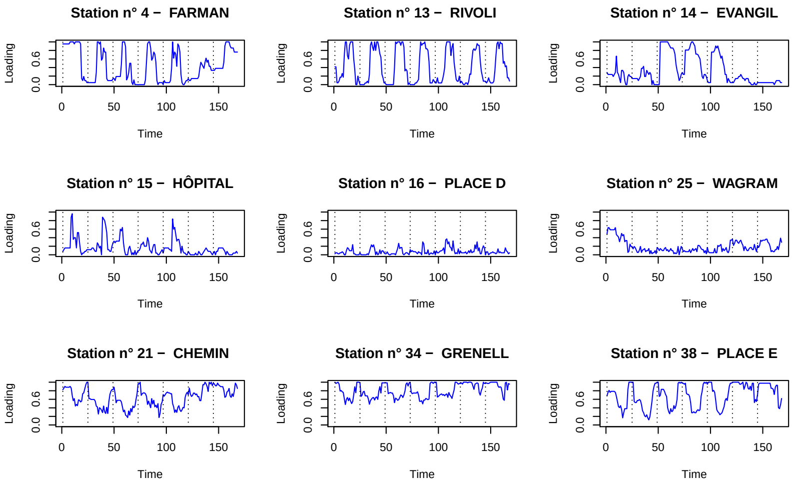


FIGURE – Selection of nine stations

Creation of two qualitative variables according to

- the daily time : day = [7h, 20h], night = [21h, 6h]
- loading values : ”-” = [0, 0.2), ”=” = [0.2, 0.5), ”+” = [0, 5, 1]

	loadDay	loadNight
FARMAN	day+	night=
RIVOLI	day+	night-
EVANGIL	day=	night=
HÔPITAL	day=	night-
PLACE D	day-	night-
WAGRAM	day-	night-
CHEMIN	day+	night+
GRENELL	day+	night+
PLACE E	day+	night+

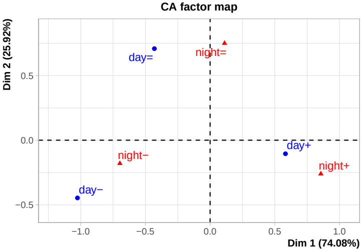
	night-	night=	night+
day-	2	0	0
day=	1	1	0
day+	1	1	3

Row profiles :

	night-	night=	night+
day-			
day=			
day+			

Column profiles :

	night-	night=	night+
day-			
day=			
day+			



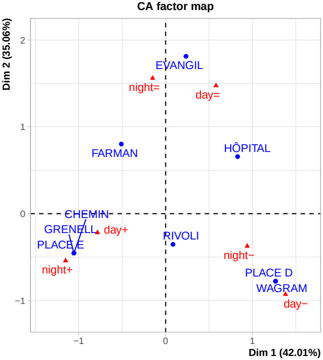
Idea 1. To apply CA on the contingency table T.

	day-	day=	day+	night-	night=	night+
FARMAN	0	0	1	0	1	0
RIVOLI	0	0	1	1	0	0
EVANGIL	0	1	0	0	1	0
HÔPITAL	0	1	0	1	0	0
PLACE D	1	0	0	1	0	0
WAGRAM	1	0	0	1	0	0
CHEMIN	0	0	1	0	0	1
GRENELL	0	0	1	0	0	1
PLACE E	0	0	1	0	0	1

Idea 2. Create a ”disjunctive table” D with dummy variables.

Q : What can you say of the table of row profiles ?

	day-	day=	day+	night-	night=	night+
FA	0	0	1	0	1	0
RI	0	0	1	1	0	0
EV	0	1	0	0	1	0
HÔ	0	1	0	1	0	0
P. D	1	0	0	1	0	0
WA	1	0	0	1	0	0
CH	0	0	1	0	0	1
GR	0	0	1	0	0	1
P. E	0	0	1	0	0	1



Idea 2. To apply CA on the disjunctive table D.

Q : Interpret the results. Compare with idea 1.

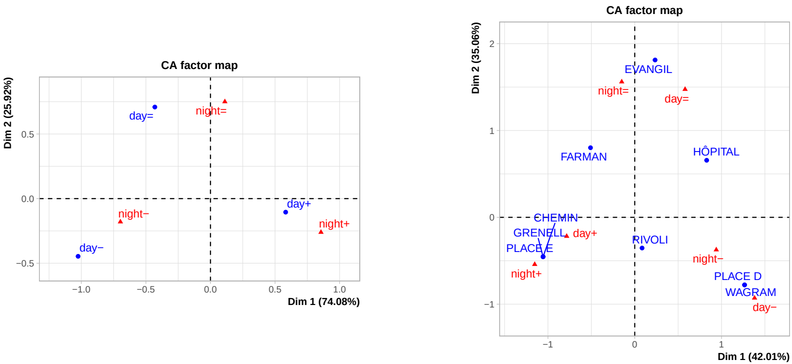


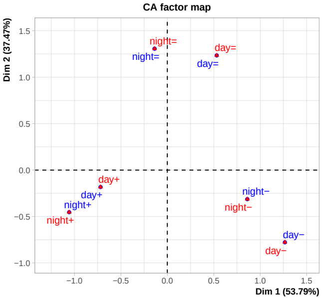
FIGURE – Comparison of the two CA, on T or on D

	day-	day=	day+	night-	night=	night+
day-	2	0	0	2	0	0
day=	0	2	0	1	1	0
day+	0	0	5	1	1	3
night-	2	1	1	4	0	0
night=	0	1	1	0	2	0
night+	0	0	3	0	0	3

Idea 3. Create the ”Burt” table, $B = D^\top D$.

Q : Interpret each block in terms of contingency. Where is T ?

	day-	day=	day+	night-	night=	night+
day-	2	0	0	2	0	0
day=	0	2	0	1	1	0
day+	0	0	5	1	1	3
night-	2	1	1	4	0	0
night=	0	1	1	0	2	0
night+	0	0	3	0	0	3



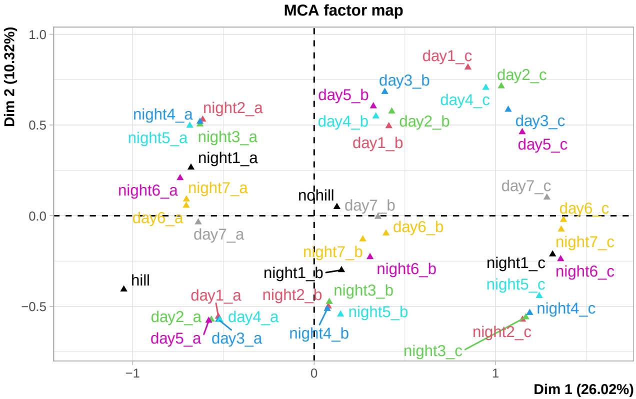
Idea 3. To apply CA on the Burt table B.

Q : Interpret the results. Compare with idea 1 & 2.

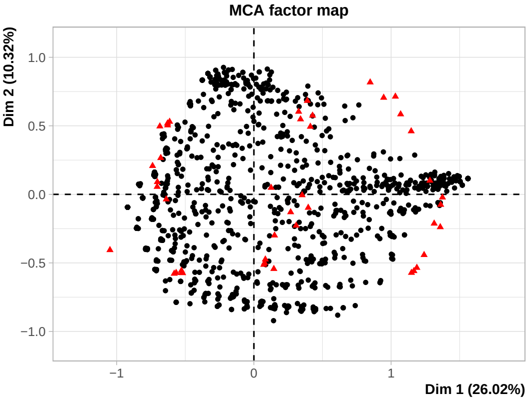
We now do a more realistic analysis, by creating 15 variables :

- hill : binary variable with levels "hill / nohill"
- dayi $(i = 1, \ldots, 7)$: 3 levels according to the mean loading in [7h, 20h] for day n°i, defined by [0, 1/3, 2/3, 1].
- nighti $(i = 1, \ldots 7)$: Same thing for the nights

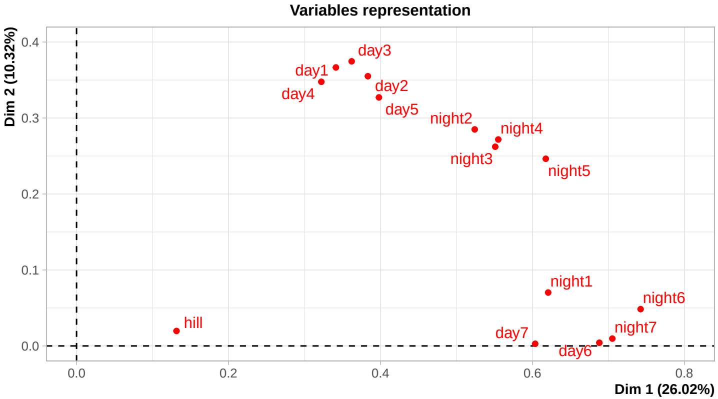
	hill	day1	day2	...	day7	night1	night2	...	night7
1	nohill	a	a	...	a	a	b	...	a
2	nohill	a	a	...	b	b	c	...	b
3	nohill	a	c	...	b	a	b	...	b
4	nohill	c	b	...	c	b	a	...	b
5	nohill	c	c	...	b	b	b	...	b
6	nohill	a	a	...	a	b	b	...	a
...									



CA on the disjunctive table.



CA on the disjunctive table, with individuals.



CA on the disjunctive table (representation of variables only).

Application to clustering : let's use the 5 first coordinates of the individuals (PCA scores of the row profiles obtained from D).

In the next slides, we compare the results obtained :

- by kMeans on the 5th first PCA principal components, on the original quantitative data (without hill)
- by kMeans on the 5th first MCA principal components, on the qualitative data (including hill)

	1	2	3	4	5	6	7
1	1	0	0	137	15	7	9
2	0	1	81	0	0	9	0
3	0	1	55	0	0	12	49
4	7	0	0	43	261	0	0
5	0	1	0	17	8	7	96
6	2	58	4	7	0	61	20
7	168	14	0	10	10	18	0

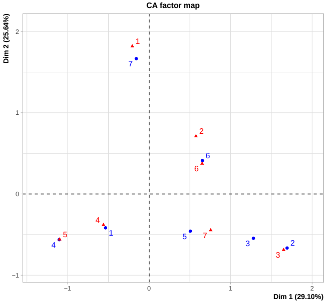
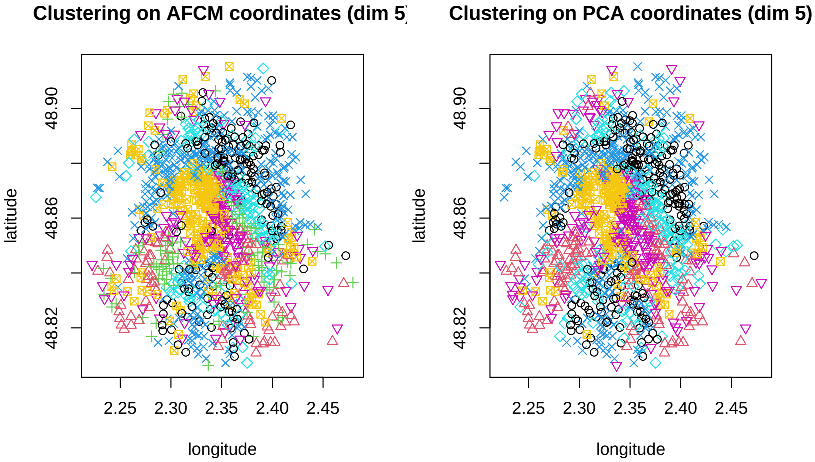
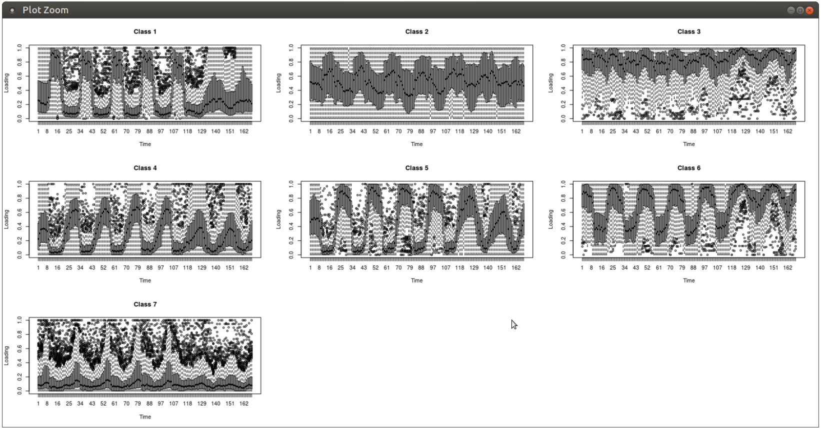


FIGURE – CA on the contingency table of the two clusterings : kMeans on the 5th first coordinates of PCA or MCA.


Clustering on AFCM coordinates (dim 5)
Clustering on PCA coordinates (dim 5)
latitude
48.82
48.86
48.90
2.25
2.30
2.35
2.40
2.45
longitude


Plot Zoom
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Loading
Time
0.0 0.2 0.4 0.6 0.8 1.0
1 8 16 25 34 43 52 61 70 79 88 97 107 118 129 140 151 162

Some conclusions from this example.

- With 2 qualitative variables, three CA give the same conclusions but the CA on D also includes individuals :)
- Thus the disjunctive table D is used in general.
- Realistic interpretation for the velib data.
- Realistic results for clustering with mixed quantitative / qualitative data

Indicator functions of levels

- X : qualitative variable with c levels.
- Indicator variable of the k^{th} level of X (size n_k) :

$$X_{(k)}(i) = \begin{cases} 1 & \text{if } X(i) = \mathcal{X}_k, \\ 0 & \text{else,} \end{cases}$$

- Matrix of indicators (or 'dummy' variables) $\mathbf{X}(n \times c)$:

$$x_i^k = X_{(k)}(i). \quad \text{with} \quad \sum_{k=1}^{c} x_i^k = 1, \forall i \;\text{ and }\; \sum_{i=1}^{n} x_i^k = n_k.$$

Disjunctive table

- $X^j; j = 1, \dots, p$, p qualitative variables
  - X^j with c_j levels : $c = \sum_{j=1}^{p} c_j$
  - $\mathbf{X}_j$ the matrix of indicators of X^j
- Disjunctive table :

$$\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p] \quad \text{with} \quad \sum_{k=1}^{c} x_i^k = p, \forall i \text{ and } \sum_{i=1}^{n} \sum_{k=1}^{c} x_i^k = np$$

Burt table $(c \times c)$

$$\mathcal{B} = \mathbf{X}'\mathbf{X}$$

$$\mathcal{B} = [\mathcal{B}_{j\ell}] \quad (j = 1, \ldots, p; \ell = 1, \ldots, p);$$

where $\mathcal{B}_{j\ell}$, $(c_j \times c_\ell)$ is the contingency table :

$$\mathcal{B}_{j\ell} = \mathbf{X}_j'\mathbf{X}_\ell$$

$$\mathcal{B}_{jj} = \text{ diag } (n_1^j, \ldots, n_{c_j}^j)$$

$\mathcal{B}$ is symmetric, with marginal totals $n_\ell^j p$ and grand total np^2

Notations for the MCA of $\mathbf{X}$

X^1 and X^2 with r and c levels.

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2] \; ; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n \; ; \\
\overline{\mathbf{D}}_c &= \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\boldsymbol{\Delta} \; ; \\
\overline{\mathbf{A}} &= \frac{1}{2n}\overline{\mathbf{T}}'\overline{\mathbf{D}}_r^{-1} = \frac{1}{2}\mathbf{X}' \; ; \\
\overline{\mathbf{B}} &= \frac{1}{2n}\overline{\mathbf{T}}\,\overline{\mathbf{D}}_c^{-1} = \frac{1}{n}\mathbf{X}\boldsymbol{\Delta}^{-1}.
\end{aligned}$$

MCA = PCA of row and column profiles.

PCA of the row profiles of $\mathbf{X}$

That PCA leads to the spectral decomposition of the $\overline{\mathbf{D}}_c^{-1}$–symmetric and p.s.d. matrix : $\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix}$

- $r+c$ eigenvalues of $\overline{\mathbf{A}}\,\overline{\mathbf{B}}$: $\mu_k = \frac{1 \pm \sqrt{\lambda_k}}{2}$ (λ_k, eigenvalue of $\mathbf{AB}$)
- $\overline{\mathbf{D}}_c^{-1}$–orthonormal eigenvectors : $\overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ with $\mathbf{U}$ (resp. $\mathbf{V}$) = $\mathbf{D}_r^{-1}$–ortho. (resp. $\mathbf{D}_c^{-1}$) eigenvectors of $\mathbf{BA}$ (resp. $\mathbf{AB}$)
- Principal components : $\overline{\mathbf{C}}_r = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r + \mathbf{X}_2\mathbf{C}_c\right]\mathbf{\Lambda}^{-1/2}$, with $\mathbf{C}_r$ and $\mathbf{C}_c$: principal components of CA

Caution

- CA provide additional non-zero eigenvalues, without statistical signification

PCA of colums profils of $\mathbf{X}$

That PCA leads to the spectral decomposition of the $\overline{\mathbf{D}}_r^{-1}$–symmetric and p.s.d. matrix :

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{2n}\left[\mathbf{X}_1\mathbf{D}_r^{-1}\mathbf{X}_1' + \mathbf{X}_2\mathbf{D}_c^{-1}\mathbf{X}_2'\right].$$

- μ_k : $r+c$ non-zero eigenvalues of $\overline{\mathbf{B}}\,\overline{\mathbf{A}}$
- $\overline{\mathbf{D}}_r^{-1}$–orthonormal eigenvectors : $\overline{\mathbf{U}} = \frac{1}{n}\overline{\mathbf{C}}_r\mathbf{M}^{-1/2}$
- Principal components : $\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{M}^{1/2}$

Notations of MCA for $\mathcal{B}$

$\mathcal{B}$ is symmetric $\Rightarrow$ row profiles = column profiles

$$\begin{aligned}
\widetilde{\mathbf{T}} &= \mathcal{B} = \begin{bmatrix} n\mathbf{D}_r & \mathbf{T} \\ \mathbf{T}' & n\mathbf{D}_c \end{bmatrix}; \\
\widetilde{\mathbf{D}}_r &= \widetilde{\mathbf{D}}_c = \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\boldsymbol{\Delta} = \overline{\mathbf{D}}_c\ ; \\
\widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.
\end{aligned}$$

PCA of the row (or column) profiles of $\mathcal{B}$

That PCA leads to the spectral decomposition of the $\widetilde{\mathbf{D}_c}^{-1}$–symmetric and p.s.d. matrix :

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2$$

- $\widetilde{\mathbf{D}_c}^{-1}$–orthonormal eigenvectors

$$\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$$

- Eigenvalues : $\nu_k = \mu_k^2$
- Principal components : $\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \boldsymbol{\Lambda}^{-1/2}\mathbf{M}$

Comparison

- The three MCA of $\mathbf{T}, \mathbf{X}, \mathcal{B}$, give homothetic representations of levels $\Rightarrow$ same interpretation
- The MCA of $\mathbf{X}$ and $\mathcal{B}$ have non-zero eigenvalues without signification
- The MCA of $\mathbf{X}$ provides a representation of individuals

Notations of MCA

- $\{X^j \,; j = 1, \dots, p\}$, p qualitative variables
- n^j_k counts for the k^{th} level of X^j
- $\mathbf{D}_j = \frac{1}{n} \text{ diag } (n^j_1, \dots, n^j_{c_j})$
- $\boldsymbol{\Delta} = \text{ diag } (\mathbf{D}_1 \dots \mathbf{D}_p)$ (squared, $c \times c$)
- $\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p]$: disjunctive table
- $\mathcal{B} = \mathbf{X}'\mathbf{X}$: Burt table

Definition

We call Multiple Correspondence Analysis (MCA) of the variables $(X^1, \ldots, X^p)$ the CA done either on $\mathbf{X}$ or on $\mathcal{B}$

Limitation

Mind that only 2nd order interactions are considered, but not the link between triplets (or more) of variables.

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X}\,; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n\,; \quad \overline{\mathbf{D}}_c = \frac{1}{p}\boldsymbol{\Delta}\,; \\
\overline{\mathbf{A}} &= \frac{1}{p}\mathbf{X}'\,; \quad \overline{\mathbf{B}} = \frac{1}{n}\mathbf{X}\boldsymbol{\Delta}^{-1}.
\end{aligned}$$

Row profiles of $\mathbf{X}$

- Diagonalize : $\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{np}\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}^{-1}$
- $m \leq c - p$ eigenvalues μ_k, in $\mathbf{M}$
- Eigenvectors : $\overline{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_p \end{bmatrix}$
- Principal components : $\overline{\mathbf{C}}_r = \sum_{j=1}^{p} \mathbf{X}_j \mathbf{D}_j^{-1} \mathbf{V}_j$
- Caution : The eigenvectors of the blocks $\mathbf{V}_j$ are not the $\mathbf{D}_j^{-1}$–orthonormal eigenvectors of a known matrix.

PCA of the column profiles of $\mathbf{X}$

- Diagonalize : $\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{np}\mathbf{X}\boldsymbol{\Delta}^{-1}\mathbf{X}' = \frac{1}{np}\sum_{j=1}^{p}\mathbf{X}_j\mathbf{D}_j^{-1}\mathbf{X}_j'$
- Eigenvectors : $\overline{\mathbf{U}} = \overline{\mathbf{B}}\overline{\mathbf{V}}\mathbf{M}^{-1/2}$
- Principal components :

$$\overline{\mathbf{C}}_c = p\boldsymbol{\Delta}^{-1}\overline{\mathbf{V}}\mathbf{M}^{1/2} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}$$

Properties of the MCA of $\mathcal{B}$

$\mathcal{B}$ is symmetric $\Rightarrow$ row profiles = column profiles.

$$\begin{aligned}
\widetilde{\mathbf{T}} &= \mathcal{B} \,; \\
\widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{p}\boldsymbol{\Delta} = \overline{\mathbf{D}}_c \,; \\
\widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{np}\mathcal{B}\boldsymbol{\Delta}^{-1} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.
\end{aligned}$$

PCA of the row (or column) profiles of $\mathcal{B}$

- Diagonalize :

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2$$

- Eigenvectors : $\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}}$
- Eigenvalues : $\nu_k = \mu_k^2$
- Principal components :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \overline{\mathbf{C}}_c\mathbf{M}^{1/2} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix} \mathbf{M}^{1/2}$$

Interpretation

- proximities and oppositions of the levels of different variables, preferring levels far from the origin
- mind to levels with small counts
- the $\%$ of explained inertia and other indicators are not easy to interpret anymore with the original data

Table de contingence complète

			Histologie			
			Inflam minime		Grande inflam	
Centre	Âge	Survie	Maligne	Bénigne	Maligne	Bénigne
Tokyo	< 50	non	9	7	4	3
		oui	26	68	25	9
	$50 - 69$	non	9	9	11	2
		oui	20	46	18	5
	> 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
	$50 - 69$	non	8	20	3	2
		oui	18	58	10	3
	> 70	non	9	18	3	0
		oui	15	26	1	1
Glamor.	< 50	non	16	7	3	0
		oui	16	20	8	1
	$50 - 69$	non	14	12	3	0
		oui	27	39	10	4
	> 70	non	3	7	3	0
		oui	12	11	4	1

We will study this example in a computer lab session :

- The direct analysis only considers second-order interactions on Survival $\Rightarrow$ not enough to see the interaction between $\{$Center, Age, Inflam.$\}$ on Survival.
- A solution is to create new variables and to redo the analysis.