Data Science
Static data analysis

Stéphane Marchand-Maillet

Master en Sciences Informatiques - Semestre d'Automne

Factorial Component Analysis (FCA)

Factorial Component Analysis is a particular application of PCA principles over categorical contingency tables:

$$N = \begin{array}{cc|c|cccc|} & & & & v_2 & & \\ \hline & & & 1 & 2 & \cdots & p \\ \hline & & 1 & & & & \\ & v_1 & 2 & \ddots & \vdots & & \\ & & \vdots & \cdots & n_{ij} & \cdots & \\ & & q & & \vdots & \ddots & \\ \hline \end{array}$$

n elements modeled along 2 symbolic (categorical) variables $v_1 \in I = \{1, \ldots q\}$ et $v_2 \in J = \{1, \ldots p\}$.

Example : exam results *vs* local regions

		Baccalauréat 1976 (série x région)							
*	A	B	C	D	E	F	G	H	Total
CHAM	924	464	567	984	132	423	736	12	4242
	22%	11%	13%	23%	3%	10%	17%	0,3%	100%
PICA	1081	490	830	1222	118	410	743	13	4907
	22%	10%	17%	25%	2%	8%	15%	0,3%	100%
HNOR	1135	587	686	904	83	629	813	13	4850
	23%	12%	14%	19%	2%	13%	17%	0,3%	100%
CENT	1482	667	1020	1535	173	629	989	26	6521
	23%	10%	16%	24%	3%	10%	15%	0,4%	100%
BNOR	1033	509	553	1063	100	433	742	13	4446
	23%	11%	12%	24%	2%	10%	17%	0,3%	100%
BOUR	1272	527	861	1116	219	769	1232	13	6009
	21%	9%	14%	19%	4%	13%	21%	0,2%	100%
NOPC	2549	1141	2164	2752	587	1660	1951	41	12845
	20%	9%	17%	21%	5%	13%	15%	0,3%	100%
LORR	1828	681	1364	1741	302	1289	1683	15	8903
	21%	8%	15%	20%	3%	14%	19%	0,2%	100%

Definitions

- ⋆ Line/column profiles

$$f^l_{ij} = \frac{n_{ij}}{n_{i,.}},\ \ f^c_{ij} = \frac{n_{ij}}{n_{.,j}}$$

with marginal sums

$n_{i,.} = \sum_j n_{ij}$, $n_{.,j} = \sum_i n_{ij}$ et $n = \sum_i n_{i,.} = \sum_j n_{.,j}$

- ⋆ Marginal profiles

$$f_{i,.} = \frac{n_{i,.}}{n} \qquad f_{.,j} = \frac{n_{.,j}}{n}$$

	A	B	C	$f_{i,.}$
a	0.61	0.10	0.29	0.41
b	0.30	0.41	0.29	0.08
		⋮		
$f_{.,j}$	0.03	0.28	0.69	1

Profile distribution

Matrix notation

$$\text{Let } D_1 = \begin{pmatrix} n_{1,\cdot} & & \\ & \ddots & \\ & & n_{q,\cdot} \end{pmatrix} \text{ and } D_2 = \begin{pmatrix} n_{\cdot,1} & & \\ & \ddots & \\ & & n_{\cdot,p} \end{pmatrix}$$

$$\text{then } L = \frac{1}{n} D_1^{-1} N \text{ and } C = \frac{1}{n} D_2^{-1} N^{\mathsf{T}}$$

are respectively the **line** and **column** profile distributions

FCA$\rightarrow$ over L and C

- ⋆ p elements from $\mathbb{R}^q$ or q elements from $\mathbb{R}^p$
- ⋆ χ^2 distance (better for frequencies)

Estimation of factorial axes

Line profile

- $\Sigma = L^{\mathsf{T}} M^{l}_{\chi^2} L$
- $M^{l}_{\chi^2} = \begin{pmatrix} 1/f_{\cdot,1} & & \\ & \ddots & \\ & & 1/f_{\cdot,q} \end{pmatrix}$

Column profile

- $\Sigma = C^{\mathsf{T}} M^{c}_{\chi^2} C$
- $M^{c}_{\chi^2} = \begin{pmatrix} 1/f_{1,\cdot} & & \\ & \ddots & \\ & & 1/f_{p,\cdot} \end{pmatrix}$

$$\Rightarrow \Sigma = U \Lambda^2 U^{\mathsf{T}}$$

Factorial axes and correlation

- ⋆ Columns of U are eigenvectors of Σ. They are the (factorial) axes of the line/column profiles.
- ⋆ Eigenvalues λ measure correlation between categories

1 e.v. (non-trivial) egal to 1 :

$$\begin{pmatrix} \boxed{\mathrm{A}} & \mathbf{0} \\ \mathbf{0} & \boxed{\mathrm{B}} \end{pmatrix}$$

all e.v. = 0 :

$$\begin{pmatrix} n_{11} & & \\ \mathbf{0} & \ddots & \mathbf{0} \\ & & n_{pp} \end{pmatrix}$$

Interpretation

FCA results into two sets of factorial axes: for line and column profiles respectively

- ⋆ Duality between the two representations : same e.v and the projection of lignes (resp. columns) along column (resp. line) axes $\mathbf{u}_k$ is identical modulo factor $\sqrt{\lambda_k}$.
- ⋆ The first e.v $\lambda_1 = 1$ is ignored
- ⋆ FCA search for a space for quantifying symbolic data and respect their correlation as much as possible

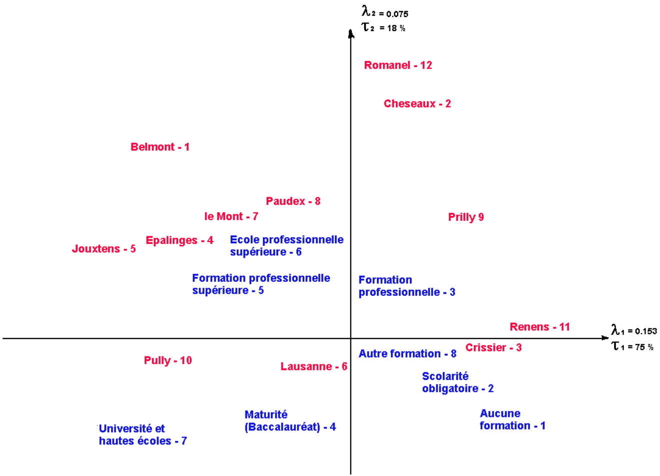
Examples

	Belmont	Cheseaux	Crissier	Epalinges	Jouxtens	Lausanne	Le Mont	Paudex	Prilly	Pully	Renens	Romanel	Margin
Aucune formation	6	26	114	36	3	2126	23	11	251	73	244	15	2928
Scolarité obligatoire	344	677	2220	1401	150	40165	994	280	3491	3670	7039	556	60987
Formation professionnelle	752	1116	1729	2253	252	39941	1486	476	4200	4721	5638	1029	63593
Maturité	163	128	249	554	51	10405	311	81	570	1465	888	126	14991
Formation professionnelle supérieure	155	135	211	497	65	5583	298	63	452	989	553	127	9128
Ecole professionnelle supérieure	62	36	90	147	24	1709	111	21	131	306	195	52	2884
Université / Haute école	196	96	169	675	110	9302	380	106	344	2010	437	84	13909
Autre	10	15	31	50	1	990	18	7	90	86	95	23	1416
Margin	1688	2229	4813	5613	656	110221	3621	1045	9529	13320	15089	2012	**169'836**

Source : François Micheloud

Visualisation

Scatterplot



Source : François Micheloud

Interpretations

- Close points from the same profile indicate similar profiles
- Close points from different profiles should be carefully analysed
- Angles between different profiles indicate facor correlation (attractive if $< 90^\circ$, repulsive if $> 90^\circ$)
- Angles between points and axes indicate their correlation

Web usage *vs* age