Cours d'Analyse des Données

Cours n°=3 : Analyse Factorielle des Correspondances

Présenté par ***Hamel Elhadj***

2021/2022

département de mathématiques
université de chlef

Introduction : Analyse Factorielle des Correspondances

- **L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples,** est une méthode exploratoire d'analyse **des tableaux de contingence**. Elle a été développée par J.-P. Benzecri durant la période 1970-1990.

- L'AFC considérée comme une ACP particulière dotée de la métrique du **(χ 2) (Khi-2)** qui ne dépend que du profil des colonnes du tableau. L'analyse permet, dans le plan des deux premiers axes factoriels, une représentation simultanée des ressemblances entre les colonnes ou les lignes du tableau et de la proximité entre lignes et colonnes.

Etudier sur N individus les **"liaisons"** entre deux variables qualitatives X et Y. Chaque variable détermine deux partitions de l'ensemble des individus selon les **modalités**.

Motivation

On suppose donnée l'observation de deux variables, X et Y , qualitatives sur n individus.
Le but de l'AFC est de résumer les dépendances entre les diverses modalités de X et Y afin de donner une vue résumée des données.

Généralement, ce type de données est représenté au travers d'une table de contingence, K:

	y_1	$\cdots$	y_j	$\cdots$	y_p	Total
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1p}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	$\ldots$	n_{kj}	$\ldots$	n_{kp}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot p}$	$n_{\cdot\cdot}$

A.F.C.

Analyse Factorielle des Correspondances

Généralisation de l'A.C.P. adaptée au traitement de **données qualitatives** se présentant sous la forme d'un **tableau de contingence**.

Principe général de l’AFC

« L’analyse factorielle traite des tableaux de nombres.

Elle ***remplace un tableau de nombres*** *difficile à analyser par* ***une série de tableaux plus simples*** *qui sont une bonne approximation de celui-ci »*

Ces tableaux sont « simples », car ***ils sont exprimables sous forme de graphiques***

Pourquoi « des correspondances » ?

variables numériques ⇒ Corrélation

variables nominales ⇒ Correspondance

Pourquoi « factorielle » ?

Il s’agit de décomposer le tableau original en une somme de tableaux/matrices qui sont chacun le **produit** de facteurs simples.

Autrement dit, on les « met en facteurs »

Méthodologie de calcul

Dans une AFC, les lignes et les colonnes jouent le même rôle

L’AFC consiste à considérer successivement
les lignes et les colonnes comme les individus d’une ACP
(les colonnes et les lignes étant successivement les variables)

AFC

=

double ACP

(sur les profils lignes
et les profils colonnes)

Le tableau de données initial

Soient X et Y deux variables **qualitatives** ayant respectivement x_n et y_m modalités.

Le tableau de contingence « T » formé à partir de ces deux variables aura autant de lignes (colonnes) que la variable X a de modalités ($n1$) et autant de colonnes (lignes) que la variable Y a de modalités ($n2$).

$$T = \begin{array}{c|ccccc|c}
 & 1 & \dots & j & \dots & n_2 & n_{i.} \\
\hline
1 & & & & & & . \\
\vdots & & & \vdots & & & \\
i & & \dots & n_{ij} & \dots & & . \\
\vdots & & & \vdots & & & \\
n_1 & & & & & & \\
\hline
n_{.j} & & & . & & & n = n_{..}
\end{array}$$

Y possède
m modalités

X possède n modalités

nombre d'individus ayant choisi simultanément les modalités x_i et y_j

Tableau de contingence

Soient *X* et *Y* deux variables qualitatives à n et s modalités respectivement décrivant un ensemble de n individus.

Définition le tableau de contingence est une matrice à n lignes et s colonnes renfermant les effectifs *nij* qui représentent le nombre d'individus ayant choisi simultanément les modalités *xi* et *yj*

$$\mathbf{T} = \begin{pmatrix} n_{11} & n_{12} & & \cdots & & n_{1s} \\ n_{21} & n_{22} & & & & \\ & & \ddots & & & \\ \vdots & & & n_{ij} & & \vdots \\ & & & & \ddots & \\ n_{r1} & & & & & n_{rs} \end{pmatrix}$$

Exemple

Considérons l'exemple où X désigne la couleur des cheveux et Y celle des yeux, de la base de HairEyeColor.

```
data(HairEyeColor)
(ex1 <- HairEyeColor[, , Sex = "Female"])

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8
```

Question: Existe-t-il une relation entre la couleur des cheveux et celle des yeux chez les femmes?

Notations

Définition (Effectifs marginaux) On définit les effectifs marginaux par:

Hair \ Eye	Brown	Blue	Hazel	Green	ni.
Black	36	9	5	2	52
Brown	66	34	29	14	143
Red	16	7	7	7	37
Blond	4	64	5	8	81
n.j	122	114	46	31	313 =n

$$n_{i\cdot} = \sum_{j=1}^{p} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}.$$

$$n = \sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{p} n_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}.$$

En lignes est présentée la variable "couleur des yeux" à n = 4 modalités (ou catégories) et en colonnes est donnée la variable "couleur des cheveux" à p= 4 modalités.

Commande sur R pour calculer les effectifs marginaux

```
# Effectifs marginaux
(ni. <- apply(ex1, 1, sum))  # ou ni.<-rowSums(ex1)

Black Brown   Red Blond
   52   143    37    81

(n.j <- apply(ex1, 2, sum))  # ou n.j<-colSums(ex1)

Brown  Blue Hazel Green
  122   114    46    31
```

Notations

Définition (Tableau des fréquences) On définit le tableau des fréquences notée « F »

$$F = f_{ij} \text{ où } f_{ij} = n_{ij}/n \text{ et } n \text{ est la somme des } n_{ij} \text{ (i.e. } n = \sum_i \sum_j n_{ij}).$$

Définition (Fréquences marginales) On définit les fréquences marginales par:

$$f_{i\cdot} = \sum_{j=1}^{p} f_{ij}, \quad f_{\cdot j} = \sum_{i=1}^{k} f_{ij}.$$

```r
# Tableau des fréquences
(fij <- ex1/sum(ex1))
```

Hair \ Eye	Brown	Blue	Hazel	Green	$f_{i\cdot}$
Black	0.115016	0.028754	0.015974	0.0063898	0.16613
Brown	0.210863	0.108626	0.092652	0.0447284	0.45687
Red	0.051118	0.022364	0.022364	0.0223642	0.11821
Blond	0.012780	0.204473	0.015974	0.0255591	0.25879
$f_{\cdot j}$	0.389776	0.364217	0.146965	0.099042	1

```r
# Fréquences marginales
(fi. <- apply(fij, 1, sum))

(f.j <- apply(fij, 2, sum))
```

Exemple

	Univ	Prepa	Autres	$n_{i.}$
Lettres	13.00	2.00	5.00	20
Economie	20.00	2.00	8.00	30
Math-Sciences	10.00	5.00	5.00	20
Tech	7.00	1.00	22.00	30
$n_{.j}$	50	10	40	100

Table de contingence - marges :

$$f_{ij} = \frac{n_{ij}}{n_{..}}, \quad f_{i.} = \frac{n_{i.}}{n_{..}} \quad et \quad f_{.j} = \frac{n_{.j}}{n_{..}}$$

	Univ	Prepa	Autres	$f_{i.}$
Lettres	13/100	2/100	5/100	20/100
Economie	20/100	2/100	8/100	30/100
Math-Sciences	10/100	5/100	5/100	20/100
Tech	7/100	1/100	22/100	30/100
$f_{.j}$	50/100	10/100	40/100	1

Tableau des fréquences : **F**

Tableau des profils-lignes et colonnes :

On appelle *tableau des profils-lignes*(*matrice des profils-lignes) ,noté* PL *où* L *, le tableau correspondant aux fréquences conditionnelles*

$$\boxed{(L)_{ij} = n_{ij}/n_{i.} = f_{ij}/f_{i.}}$$

et de même pour le *tableau des profils colonnes*, *noté* ***PC où C*** *de terme générale ,*

$$\boxed{(C)_{ij} = n_{ij}/n_{.j} = f_{ij}/f_{.j}}$$

Soit les matrices diagonales suivantes

$$\boxed{D_1 = D_r = \begin{pmatrix} f_{i.} & & \\ & \ddots & \\ & & f_{n_1.} \end{pmatrix} \quad , \quad D_2 = D_c = \begin{pmatrix} f_{.1} & & \\ & \ddots & \\ & & f_{.n_2} \end{pmatrix}}$$

'on a $L = D_1^{-1}.F$ *et* $C = D_2^{-1}.F'$. où F' *correspond à la transposée de* F.

Tableau des profils-lignes et colonnes : Exemple

	1	2	3	Eff. Marg. Lig.
A	1	0	3	4
B	0	2	0	2
C	4	0	2	6
Eff. Marg. Col	5	2	5	12

Tableau initial

	1	2	3	
A	1/4	0	3/4	1
B	0	1	0	1
C	4/6	0	2/6	1

Tableau des profils- lignes **L**

	1	2	3	
A	1/5	0	3/5	
B	0	1	0	
C	4/5	0	2/5	
	1	1	1	3

Tableau des profils-colonnes **C**

$$D_r = \begin{pmatrix} 4/12 & 0 & 0 \\ 0 & 2/12 & 0 \\ 0 & 0 & 6/12 \end{pmatrix}$$

$$D_c = \begin{pmatrix} 5/12 & 0 & 0 \\ 0 & 2/12 & 0 \\ 0 & 0 & 5/12 \end{pmatrix}$$

$$L = D_r^{-1}.F \;\; et \; C = D_c^{-1}.F'.$$

Tableau des profils-lignes et colonnes :

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Total
Bleu	20	84	17	94	215
Vert	5	29	14	16	64
Noisette	15	54	14	10	93
Marron	68	119	26	7	220
Total	108	286	71	127	592

Profils lignes

	Brun	Châtain	Roux	Blond	Total
Marron	0,31	0,54	0,12	0,3	1
Noisette	0,16	0,58	0,15	0,11	1
Vert	0,8	0,45	0,22	0,25	1
Bleu	0,9	0,39	0,8	0,44	1
Profil moyen	0,18	0,48	0,12	0,22	1

Profils colonnes

	Brun	Châtain	Roux	Blond	Profil moyen
Marron	0,63	0,42	0,37	0,6	0,37
Noisette	0,14	0,19	0,2	0,8	0,16
Vert	0,5	0,1	0,2	0,13	0,11
Bleu	0,19	0,29	0,24	0,74	0,36
Total	1	1	1	1	1

AFC et indépendance

L'analyse des correspondances (AFC) étudie les proximités entre individus décrits par deux ou plusieurs variables qualitatives ainsi que les proximités entre les modalités de ces variables.

- *Il y a* indépendance entre les deux variables considérées si : $\boxed{f_{ij} = f_{i.} \times f_{.j}}$
- deux variables sont liées si elles ne sont pas indépendantes.
- Dans le cas de l'indépendance tous les profils lignes et colonnes sont identiques.

La distance du Khi-deux « χ^2 »

la distance dite du chi deux (*écrire* χ^2 *et prononcer ki deux*) qui sert à vérifier s'il y a un lien entre les modalités prises par deux variables qualitatives (l'analogue de la corrélation pour les variables quantitatives).

Definition (Distance de χ^2) On appelle distance de chi-deux entre les variables X et Y la quantité:

$$\begin{aligned}\chi^2 &= n\varphi = n\sum_{i,j}\frac{(f_{ij} - f_{i.}\times f_{.j})^2}{f_{i.}\times f_{.j}} \\ &= n\sum_{i,j}\frac{(n_{ij} - n_{i.}\times n_{.j})^2}{n_{i.}\times n_{.j}} = n\left(\sum_{i,j}\frac{n_{ij}^2}{n_{i.}\times n_{.j}} - 1\right)\end{aligned}$$

Cette grandeur est souvent utilisée comme test d'indépendance. En effet sous l'hypothèse nulle d'indépendance, χ^2 suit la loi chi-deux à (n - 1) × (p - 1) degrés de liberté .

Utiliser la distance de χ^2

L'idée pour comparer des profils lignes ou des profils colonnes sera d'utiliser la distance du $\chi 2$

— Espace des profils lignes, $(\mathbb{R}^{n_2}, D_c^{-1})$: Soient $\ell_i, \ell_{i'}$ deux lignes de tableau L :

$$d^2_{\chi^2}(\ell_i, \ell_{i'}) = \sum_{j=1}^{n_2} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

— Espace des profils lignes, $(\mathbb{R}^{n_1}, D_r^{-1})$: Soient $C_j, C_{j'}$ deux lignes de tableau C :

$$d^2_{\chi^2}(C_j, C_{j'}) = \sum_{i=1}^{n_1} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Utiliser la distance de χ^2

Lorsque l'on effectue dans le tableau des fréquences F, la somme de deux colonnes proportionnelles (ou de deux lignes), les distances du $\chi 2$ entre profils lignes (ou colonnes) restent inchangées.

Lien avec le $\chi 2$ de contingence :

$$\frac{d^2}{n} = \sum_{i=1}^{n_1} f_{i.} d^2_{\chi^2}(L_i, \overline{L}) = \sum_{j=1}^{n_2} f_{.j} d^2_{\chi^2}(C_j, \overline{C})$$

Le coefficient $\frac{d^2}{n}$ est égal à l'inertie du nuage des profils lignes (des profils colonnes).

Le centre de gravité du nuage est le profils ligne moyen

$$\overline{L} = \frac{1}{n} (D_r F)' D_r^{-1}$$

Pourquoi la distance du χ^2 ?

- Avec la métrique du χ^2, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes.

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{q} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2$$

- La métrique du χ^2 possède la propriété d'équivalence distributionnelle : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.
- Notons qu'en revanche, il n'existe pas d'outil mesurant une "distance" entre une ligne et une colonne.

Exemple

	I	II	III	
A	2	4	2	=8
B	4	8	3	=15
C	3	3	3	
	=9			

Distance du $\chi 2$:

$$d^2(A,B) = \frac{1}{9/32}\left(\frac{2}{8}-\frac{4}{15}\right)^2 + \frac{1}{15/32}\left(\frac{4}{8}-\frac{8}{15}\right)^2 + \frac{1}{8/32}\left(\frac{2}{8}-\frac{3}{15}\right)^2$$

d(A,B) = 0,11
d(A,C) = 0,33
d(B,C) = 0,41

La distance du **$\chi 2$** montre que A et B se ressemblent beaucoup

Ceci est satisfaisant !!

A et B sont proches car ils sont similaires en termes de forme de répartition

Résumé	Points lignes	Points colonnes
matrice	$L = D_1^{-1}F$	$L = FD_2^{-1}$
nuages	$\mathbb{N}_L = L_1, ..., L_{n_1}$	$\mathbb{N}_C = C_1, ..., C_{n_2}$
poids	D_r	D_c
espace	$(\mathbb{R}^{n_2}, D_c^{-1})$	$(\mathbb{R}^{n_1}, D_r^{-1})$
point moyen	$1'_{n_1}F$	$F1'_{n_2}$
inertie	$\frac{d^2}{n}$	$\frac{d^2}{n}$

Le but de l'AFC

Les objectifs de l'analyse factorielle des correspondances (AFC) sont de

- comparer les profils-lignes entre eux,
- comparer les profils-colonnes entre eux,
- repérer les cases du tableau où les effectifs observés n_{ij} sont nettement différents des effectifs théoriques (sous l'hypothèse d'indépendance).

L'AFC est une méthode faisant apparaître les cartéristiques de la situation d'indépendance, au niveau des lignes, des colonnes, ou des cases du tableau de contingence.

Les nuages des deux profils

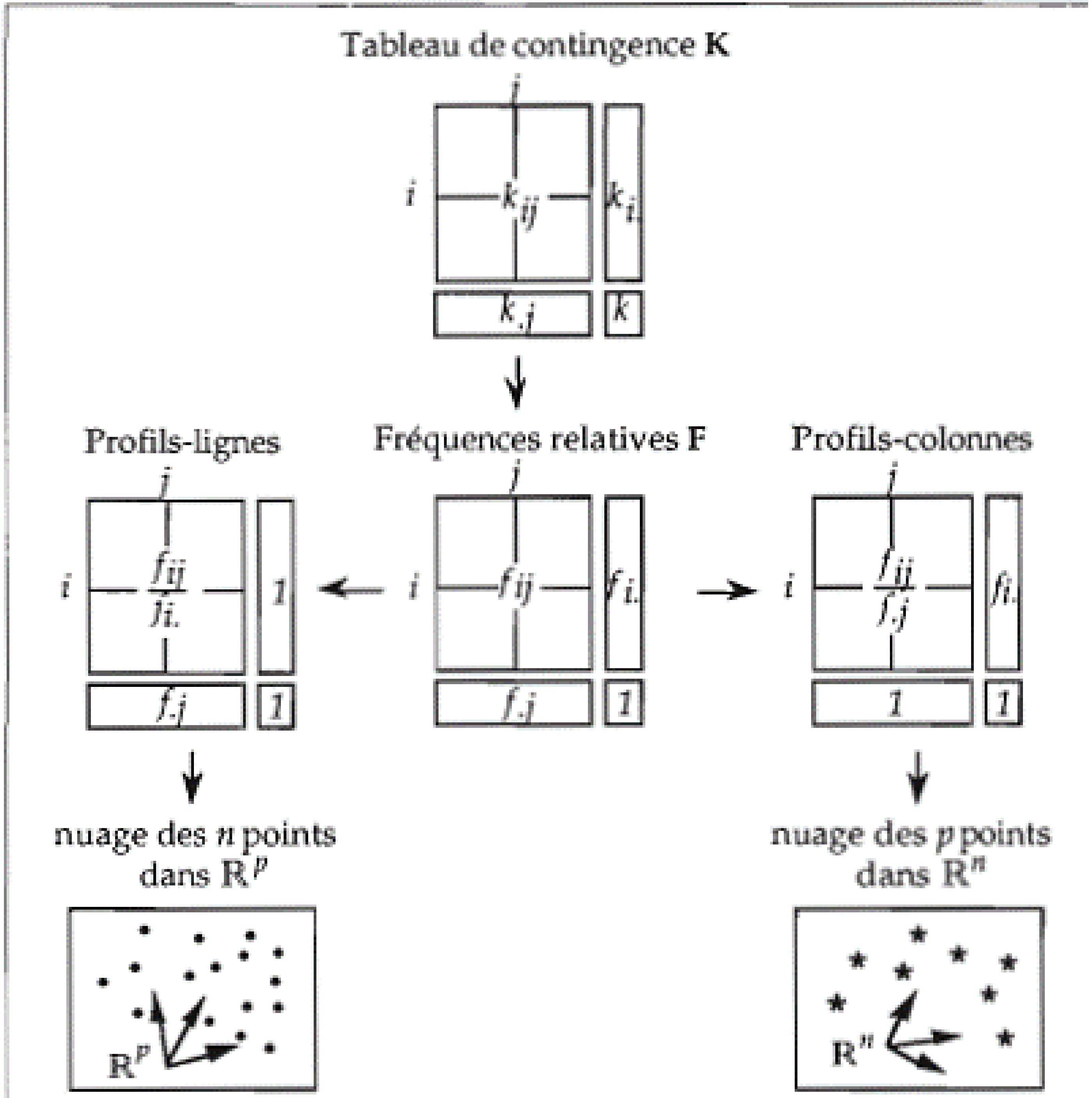
Une double ACP

L'AFC est, en fait, une double ACP:

- sur les profils lignes:
  - Tableau de données: $L = D_r F$
  - Matrice de poids: D_r^{-1}.
  - Métrique d'écart à l'indépendance: D_c
- sur les profils colonnes:
  - Tableau de données: $C = D_c F'$
  - Matrice de poids: D_c^{-1}.
  - Métrique d'écart à l'indépendance: D_r

Résumé	Points lignes	Points colonnes
matrice	$L = D_1^{-1} F$	$L = F D_2^{-1}$
nuages	$\mathbb{N}_L = L_1, ..., L_{n_1}$	$\mathbb{N}_C = C_1, ..., C_{n_2}$
poids	D_r	D_c
espace	$(\mathbb{R}^{n_2}, D_c^{-1})$	$(\mathbb{R}^{n_1}, D_r^{-1})$
point moyen	$1'_{n_1} F$	$F 1'_{n_2}$
inertie	$\frac{d^2}{n}$	$\frac{d^2}{n}$

Schéma général de l'analyse des correspondances



Nuage de n_1 points-lignes dans l'espace $\mathbb{R}^{n_2}$	Eléments de bases	Nuage de n_2 points-colonnes dans l'espace $\mathbb{R}^{n_1}$
$X = D_r^{-1}F$ n_1 coordonnées (point-ligne i) $\frac{f_{ij}}{f_{i.}}; i = 1, ..., n_1$	Analyse de tableau X	$X = D_c^{-1}F'$ n_2 coordonnées (point-colonne j) $\frac{f_{ij}}{f_{.j}}; j = 1, ..., n_2$
$M = D_c^{-1}$ $d^2(i,i') = \sum_{j=1}^{n_2} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)$	avec la metrique M	$M = D_r^{-1}$ $d^2(j,j') = \sum_{i=1}^{n_1} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)$

Axes factoriels et facteurs

Dans $\mathbb{R}^{n_2}$	Eléments de construction	Dans $\mathbb{R}^{n_1}$
$S = F'D_r^{-1}FD_c^{-1}$	Matrice à diagonaliser	$T = FD_c^{-1}F'D_r^{-1}$
$Su_\alpha = \lambda_\alpha u_\alpha$	Axe factoriel	$Sv_\alpha = \lambda_\alpha v_\alpha$
$\psi_\alpha = D_r^{-1}FD_c^{-1}u_\alpha$ $\psi_{\alpha i} = \sum_{j=1}^{n_2} \frac{f_{ij}}{f_{i.}f_{.j}} u_{\alpha j}$	Coordonnées factorielles	$\phi_\alpha = D_c^{-1}F'D_r^{-1}v_\alpha$ $\phi_{\alpha j} = \sum_{i=1}^{n_1} \frac{f_{ij}}{f_{i.}f_{.j}} v_{\alpha i}$

Les coordonnées factorielles sont centrées : $\displaystyle\sum_{i=1}^{n_1} f_{i.}\psi_{\alpha i} = \sum_{j=1}^{n_2} f_{.j}\phi_{\alpha j} = 0$

et de variance égale à λ_α : $\displaystyle\sum_{i=1}^{n_1} f_{i.}\psi_{\alpha i}^2 = \sum_{j=1}^{n_2} f_{.j}\phi_{\alpha j}^2 = \lambda_\alpha$

Relation entre les deux espaces

L'analyse générale a montré que les matrices **S** et **T** ont les mêmes valeurs propres non nulles λ_α et qu'entre le vecteur propre unitaire $u\alpha$ de S associé à $\lambda\alpha$ et le vecteur propre unitaire $v\alpha$ de T relatif à la même valeur propre, il existe les relations dites de transition :

$$\begin{cases} \psi_\alpha = D_r^{-1} F D_c^{-1} u_\alpha, \\ \phi_\alpha = D_c^{-1} F' D_r^{-1} v_\alpha \end{cases}$$

La comparaison de ces relations avec les expressions des coordonnées factorielles :

montre que celles-ci sont liées aux composantes des axes de l'autre espace par les formules :

$$\begin{cases} \psi_\alpha = \sqrt{\lambda_\alpha} D_r^{-1} v_\alpha, \\ \phi_\alpha = \sqrt{\lambda_\alpha} D_c^{-1} u_\alpha \end{cases}$$

C'est à dire explicitement

$$\begin{cases} \psi_\alpha = \dfrac{\sqrt{\lambda_\alpha}}{f_{i.}} v_{\alpha i}, \\ \phi_\alpha = \dfrac{\sqrt{\lambda_\alpha}}{f_{.j}} u_{\alpha j} \end{cases}$$

Relation entre les deux espaces

Ces Relations de transition (ou quasi-barycentriques) conduisent aux relations fondamentales existant entre les coordonnées des points lignes et des points-colonnes sur l'axe α, les relations quasi-barycentriques :

La matrice de terme général $\frac{f_{ij}}{f_{i.}}$ permettant de calculer les coordonnées d'un point *i* à partir de tous les points *j* n'est autre que le tableau des profils-lignes.

$$\begin{cases} \psi_{\alpha i} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{j=1}^{n_2} \dfrac{f_{ij}}{f_{i.}} \phi_{\alpha j}, \\ \phi_{\alpha j} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{i=1}^{n_1} \dfrac{f_{ij}}{f_{.j}} \psi_{\alpha i} \end{cases}$$

Remarque : ***Toutes les valeurs propres sont nécessairement inférieures ou égales à 1.*** **<= 1 .**

Règles d'interprétation

Les nuages de points-lignes et de points-colonnes vont être représentés dans les plans de projection formés par les premiers axes factoriels pris deux à deux.

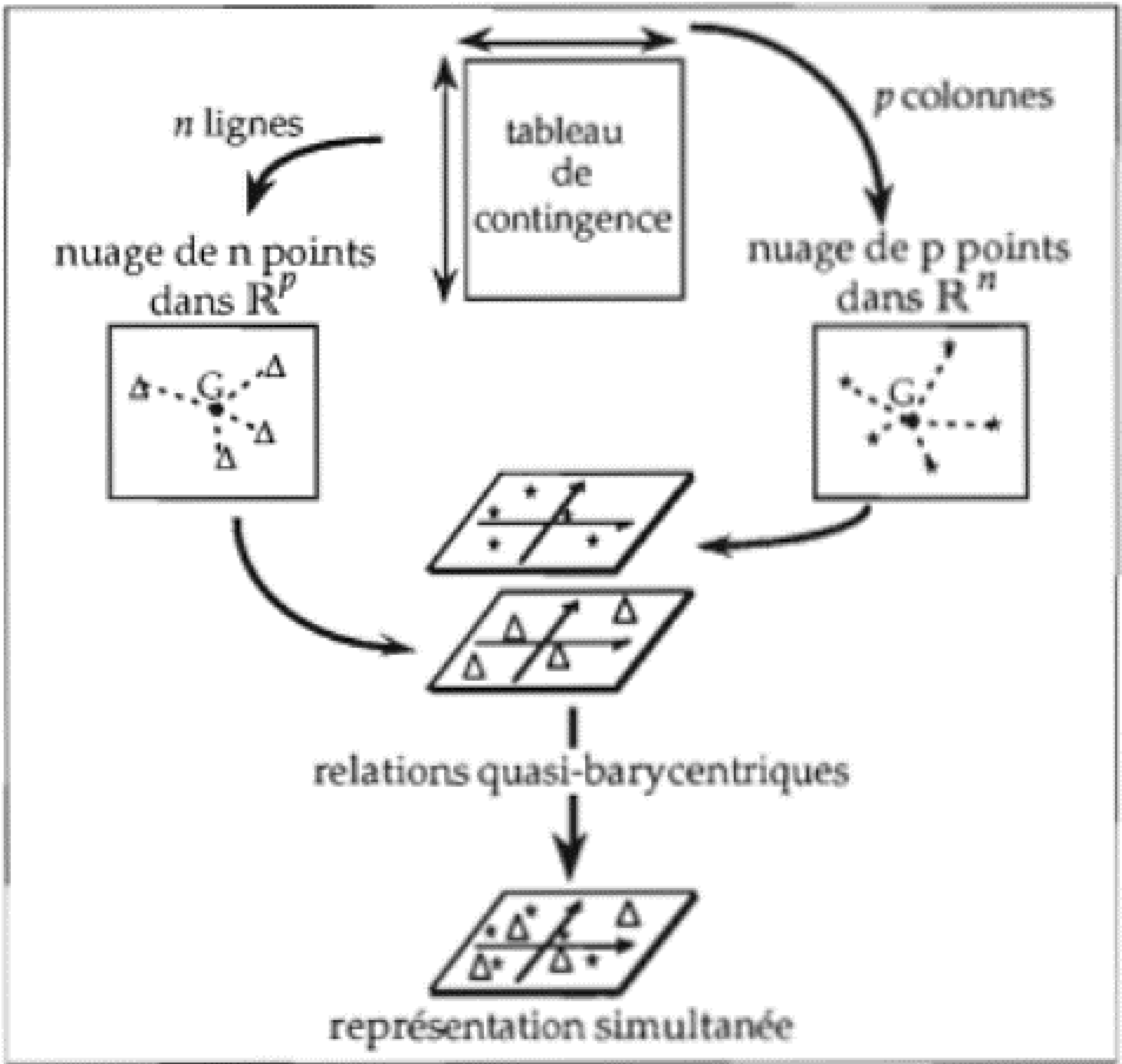
La lecture des graphiques nécessite cependant des règles d'interprétation .

1. **Inertie et test d'indépendance** : L'inertie totale ***I*** du nuage de points par rapport au centre de gravité s'écrit par définition :

$$I = \sum_{i=1}^{n_1} f_{i.} d^2(i,G) = \sum_{j=1}^{n_2} f_{.j} d^2(j,G) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.} f_{.j}} \right)$$

L'inertie s'exprime également par : $I = \sum_{\alpha=1}^{n_2-1} \lambda_\alpha$ **La somme des valeurs propres non triviales d'une analyse des correspondances a donc une interprétation statistique simple.**

Schéma de la
représentation simultanée



Règles d'interprétation

contributions et cosinus carrés

Deux séries de coefficients apportent une information supplémentaire par rapport aux coordonnées factorielles :

— ***les contributions***, parfois appelées **contributions absolues**, qui expriment la part prise par une modalité de la variable dans l'inertie (ou variance) "expliquée" par un facteur ;

— ***les cosinus carrés***, parfois appelés ***contributions relatives ou qualité de représentation***, qui expriment la part prise par un facteur dans la dispersion d'une modalité de la variable.

– **Contributions** mesure la part de l'élément i dans la variance prise en compte sur l'axe α .

$$Ctr_\alpha(i) = \frac{f_{i.}\psi^2_{\alpha i}}{\lambda_\alpha}.$$

De la même façon on définit la contribution de l'élément j à l'axe α par :

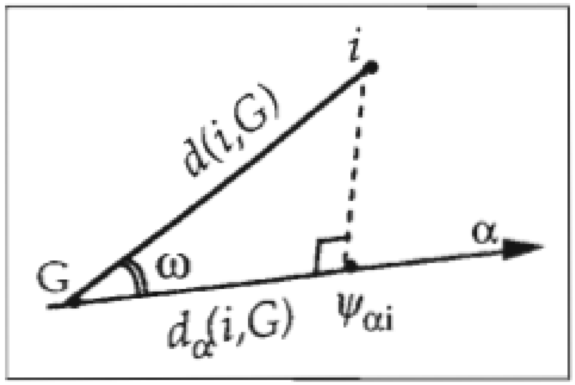
$$Ctr_\alpha(j) = \frac{f_{.j}\phi^2_{\alpha j}}{\lambda_\alpha}.$$

$$\sum_{i=1}^{n_1} Ctr_\alpha(i) = 1 \qquad \sum_{j=1}^{n_2} Ctr_\alpha(j) = 1$$

Cosinus carrés

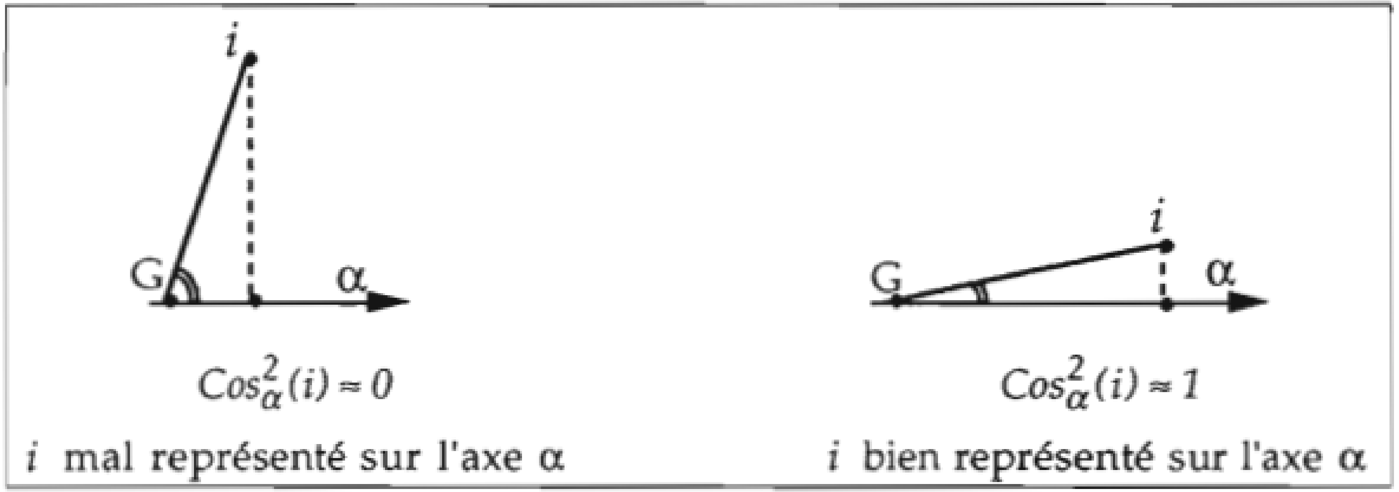
On cherche à apprécier si un point est bien représenté sur un sous-espace factoriel.
La "qualité" de la représentation du point i sur l'axe α peut être évaluée par le *cosinus* de l'angle entre l'axe et le vecteur joignant le centre de gravité du nuage au point i :

La "qualité" de la représentation du point *i* sur l'axe α peut être évaluée par le **cosinus de l'angle** entre l'axe et le vecteur joignant le centre de gravité du nuage au point *i* :



Projection du point *i* sur l'axe α

$$Cos^2_\alpha(i) = \frac{d^2_\alpha(i,G)}{d^2(i,G)} = \frac{\psi^2_{\alpha i}}{d^2(i,G)}.$$



Qualité de représentation d'un point *i* sur l'axe α

Exemple numérique

$$F = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} \\ \frac{1}{12} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{12} \end{pmatrix}$$

1/6	1/6	1/3
1/12	1/4	1/3
1/4	1/12	1/3
1/2	1/2	1

$$L = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}$$

Distance χ^2

$i=1;\ i'=2$ $$d^2(1,2)=\frac{1}{2}\left(\frac{1}{2}-\frac{1}{4}\right)^2+\frac{1}{2}\left(\frac{1}{2}-\frac{3}{4}\right)^2=\frac{1}{2}\left(\frac{1}{4}\right)^2+\frac{1}{2}\left(\frac{-1}{4}\right)^2=\frac{1}{16}$$

$i=1;\ i'=3$ $$d^2(1,3)=\frac{1}{2}\left(\frac{1}{2}-\frac{3}{4}\right)^2+\frac{1}{2}\left(\frac{1}{2}-\frac{1}{4}\right)^2=\frac{1}{2}\left(\frac{-1}{4}\right)^2+\frac{1}{2}\left(\frac{1}{4}\right)^2=\frac{1}{16}$$

$i=2;\ i'=3$ $$d^2(2,3)=\frac{1}{2}\left(\frac{1}{4}-\frac{3}{4}\right)^2+\frac{1}{2}\left(\frac{3}{4}-\frac{1}{4}\right)^2=\frac{1}{2}\left(\frac{-1}{2}\right)^2+\frac{1}{2}\left(\frac{1}{2}\right)^2=\frac{1}{4}$$

Exemple numérique

Le tableau suivant représente le type d'´études poursuivies (université, classes préparatoires, autres) en fonction du parcours suivi au lycée (Lettres, Economie, Maths-Sciences, Technique).

	Univ	Prepa	Autres
Lettres	13.00	2.00	5.00
Economie	20.00	2.00	8.00
Math-Sciences	10.00	5.00	5.00
Tech	7.00	1.00	22.00

$$F = \begin{array}{ccc|c} \left(\begin{matrix} \frac{13}{100} & \frac{2}{100} & \frac{5}{100} \\ \frac{2}{100} & \frac{2}{100} & \frac{8}{100} \\ \frac{10}{100} & \frac{5}{100} & \frac{5}{100} \\ \frac{7}{100} & \frac{1}{100} & \frac{22}{100} \end{matrix}\right) & \begin{matrix} \frac{1}{5} \\ \frac{3}{10} \\ \frac{1}{5} \\ \frac{3}{10} \end{matrix} \\ \begin{matrix} \frac{5}{10} & \frac{1}{10} & \frac{2}{5} \end{matrix} & 1 \end{array}$$

$$L = \begin{pmatrix} \frac{13}{20} & \frac{1}{10} & \frac{1}{4} \\ \frac{2}{3} & \frac{1}{15} & \frac{4}{15} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{7}{30} & \frac{1}{30} & \frac{11}{15} \end{pmatrix}$$

$$d^2(L_1, L_2) = 2\left(\frac{13}{20} - \frac{2}{3}\right)^2 + 10\left(\frac{1}{10} - \frac{1}{15}\right)^2 + \frac{5}{2}\left(\frac{1}{4} - \frac{4}{15}\right)^2 = \frac{89}{7200}$$

Exemple numérique

```r
tab <- matrix(c(13, 2, 5, 20, 2, 8, 10, 5, 5, 7, 1, 22), ncol = 3, byrow = TRUE)
colnames(tab) <- c("Univ", "Prepa", "Autres")
rownames(tab) <- c("Lettres", "Economie", "Math-Sciences", "Tech")
# F: tableau des fréquences
(F <- tab/sum(tab))
```

```
##                Univ Prepa Autres
## Lettres        0.13  0.02   0.05
## Economie       0.20  0.02   0.08
## Math-Sciences  0.10  0.05   0.05
## Tech           0.07  0.01   0.22
```

	Univ	Prepa	Autres
Lettres	13.00	2.00	5.00
Economie	20.00	2.00	8.00
Math-Sciences	10.00	5.00	5.00
Tech	7.00	1.00	22.00

1) Calculer la matrice X des profils-ligne et la matrice Y des profils-colonne.

```r
n <- sum(tab)
ni. <- rowSums(tab)
n.j <- colSums(tab)
D1 <- diag(1/ni.)
(X <- D1 %*% tab)
```

```
##        Univ   Prepa Autres
## [1,] 0.6500 0.10000 0.2500
## [2,] 0.6667 0.06667 0.2667
## [3,] 0.5000 0.25000 0.2500
## [4,] 0.2333 0.03333 0.7333
```

$L = D_1^{-1}F$

```r
fi. <- ni./n
diag(1/fi.) %*% F
```

```
##          Univ   Prepa Autres
## [1,] 0.6500 0.10000 0.2500
## [2,] 0.6667 0.06667 0.2667
## [3,] 0.5000 0.25000 0.2500
## [4,] 0.2333 0.03333 0.7333
```

$\mathbf{C} = \mathbf{D_2^{-1}}.\mathbf{F'}.$

```r
D2 <- diag(1/n.j)
(L <- tab %*% D2)
```

```
##                [,1] [,2]  [,3]
## Lettres        0.26  0.2 0.125
## Economie       0.40  0.2 0.200
## Math-Sciences  0.20  0.5 0.125
## Tech           0.14  0.1 0.550
```

2) Déterminer les axes principaux ainsi que les coordonnées principales (appliquée sur les profiles lignes).

$$\boxed{S = F'D_r^{-1}FD_c^{-1}}$$

```r
(SS <- t(tab) %*% D1 %*% tab %*% D2)
```

```
##           [,1]   [,2]    [,3]
## Univ    0.5683 0.5367 0.40542
## Prepa   0.1073 0.1617 0.07542
## Autres  0.3243 0.3017 0.51917
```

```r
ei.li <- eigen(SS)
(lamda.l <- ei.li$values)
```

```
## [1] 1.00000 0.19981 0.04936
```

```r
(ap.li <- ei.li$vectors)
```

```
##          [,1]    [,2]    [,3]
## [1,] -0.7715 -0.6063  0.7523
## [2,] -0.1543 -0.1704 -0.6509
## [3,] -0.6172  0.7767 -0.1014
```

$$\psi_\alpha = D_r^{-1}FD_c^{-1}u_\alpha,$$

```r
(coo.li <- n * D1 %*% tab %*% D2 %*% ap.li)
```

```
##        [,1]    [,2]     [,3]
## [1,] -1.543 -0.4732  0.26374
## [2,] -1.543 -0.4042  0.50156
## [3,] -1.543 -0.5469 -0.93835
## [4,] -1.543  1.0843 -0.05182
```

3) Idem pour les profiles colonnes.

```
(QQ <- tab %*% D2 %*% t(tab) %*% D1)

##                  [,1]   [,2]   [,3]   [,4]
## Lettres        0.2202 0.2200 0.2112 0.1590
## Economie       0.3300 0.3333 0.3000 0.2467
## Math-Sciences  0.2112 0.2000 0.2563 0.1550
## Tech           0.2385 0.2467 0.2325 0.4393
```

```
ei.co <- eigen(QQ)
(ap.co <- ei.co$vectors)

##         [,1]    [,2]    [,3]     [,4]
## [1,] -0.3922 -0.2516 -0.2138 -0.75388
## [2,] -0.5883 -0.3224 -0.6098  0.64302
## [3,] -0.3922 -0.2908  0.7606  0.13304
## [4,] -0.5883  0.8649  0.0630 -0.02217
```

```
(lamda.c <- ei.co$values)

## [1]  1.000e+00  1.998e-01  4.936e-02 -5.930e-17

(coo.co <- n * D2 %*% t(tab) %*% D1 %*% ap.co)

##        [,1]    [,2]     [,3]       [,4]
## [1,] -1.961 -0.6443 -0.30100 -6.099e-17
## [2,] -1.961 -0.9054  1.30212 -1.805e-15
## [3,] -1.961  1.0317  0.05072 -8.674e-17
```

Exemple numérique