Analyse factorielle des correspondances -

Correspondance Analysis

J.P Benzécri (1973),
L'analyse des données. Tome 1: La taxinomie. Tome 2: L'analyse des correspondances. Dunod, Paris.1973)

Majeure Science des données : 2020
M. Batton-Hubert

L'approche

1. Les données
2. Indépendances entre les variables
3. Tableau de contingence et fréquence
4. Liaison entre deux variables qualitatives
5. Profils *ligne* et *colonne* - profils *moyens*
6. Réduction de dimension
7. Représentations simultanées *ligne et colonne*

1. Les données

On considère deux **variables qualitatives** observées simultanément sur n individus affectés de poids identiques 1/n. On suppose que la première variable, notée X, possède I modalités notées $x_1,\ldots,x_I$, et que la seconde, notée Y, possède J modalités notées $y_1,\ldots,y_J$.

- La table de contingence (des correspondances) associée à ces observations, de dimension $L \times J$:

tableau dit de contingence

=

tableau des effectifs croisés

	1		*j*			J
1						
i			x_{ij}			
I						

Où x_{ij} le nombre d'individus appartenant à la modalité i de la première variable et à j la modalité de la deuxième variable

Distribution des n individus dans les $L \times J$ cases du tableau de correspondance T

	X	Y
1		
l	*i*	*j*
K		

	1		*j*			J
1						
i			x_{ij}			
I						

Données d'enquête

- Réponse simultanée à des questions d'enquête : *quel est le profil de la famille idéale ?: Activité pour une mère de famille (au foyer, mi-temps, plein-temps var j)/ famille idéale (deux conjoints qui travaillent , seul le conjoint a un métier, le métier du conjoint plus absorbant que celui de l'épouse : var i)*

-Bacheliers : score établi pour 2015, par discipline i et par région j :

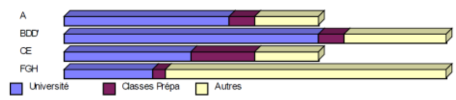
-Que deviennent les bacheliers : discipline /type études

- Abondance d'une espèce animale i, dans un milieu j : recensement

Un exemple : AFC - pour faire quoi ?

Question : Que deviennent les bacheliers ?
Recensement par discipline /type études

	destination			
	université	classes prépa	autres	total
A	13	2	5	20
BDD'	20	2	8	30
CE	10	5	5	20
FGH	7	1	22	30
total	50	10	40	100



Question :
S'il n'y avait pas de relation entre préférentielle entre un choix d'études supérieures et le bac obtenu qu'aurait-on ?

en d'autre terme s'il y avait indépendance entre les deux choix consécutifs, qu'aurait-on dans la matrice T des correspondances

→ reconstruire les données en cas d'indépendance (matrice T_0)

Matrice T à la matrice T0

tableau T

			20
			30
			20
			30
50	10	40	

→

Produit matriciel : trouver la valeur des effectifs x_{0ij}
À partir des marges
Marge_ligne (profil empirique de Y)
Marge_colonne (profil empirique de X)

La dernière ligne correspondant à la loi empirique pour Y : **P(Y = yj) = f**$_{.j}$

$$x_{0ij} = marge\_ligne(.j) \times marge\_colonne(i.)$$

Soit :

$$x_{0ij} = 50 \times 0.20 = 10$$

$$T0 = \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix}$$

Matrice de distribution des correspondances si les deux variables qualitatives sont indépendantes

Ecart entre les données de la table de contingence T et la table de contingence sous hyp. D'indépendance T0 : *R = T - T0*

$$T = \begin{bmatrix} 13 & 2 & 5 \\ 20 & 2 & 8 \\ 10 & 5 & 5 \\ 7 & 1 & 22 \end{bmatrix} \qquad T0 = \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix}$$

$$R = \begin{bmatrix} 13 & 2 & 5 \\ 20 & 2 & 8 \\ 10 & 5 & 5 \\ 7 & 1 & 22 \end{bmatrix} - \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix}$$

$$R = \begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ \dots \end{bmatrix}$$

Si R matrice de zéros : indépendance

2. Indépendance entre deux variables (a)

- Modèle d'indépendance

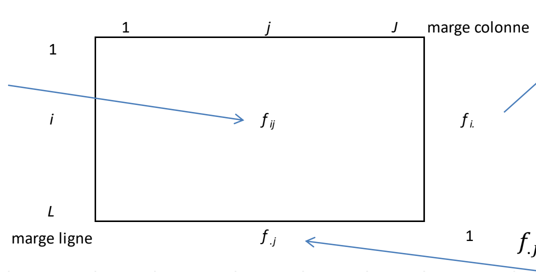
Evènements indépendants si :

P(A et B) = P(A) × P(B)

Marge colonne ou ligne = probabilité marginale

1) Passer de données du tableau de contingence à des probabilités par calcul de fréquence et marge (col. ou lig.) ***table de fréquence***

$$f_{ij} = \frac{x_{ij}}{k}$$



$$f_{i.} = \sum_{j=1}^{J} f_{ij}$$

$$f_{.j} = \sum_{i=1}^{L} f_{ij}$$

2. Indépendance entre deux variables (b)

- Modèle d'indépendance

Evènements indépendants si :

P(A et B) = P(A) × P(B)

2) Variables qualitatives indépendantes :

Probabilité conjointe = produit des probabilités marginales, soit :

$$\forall\, i, \forall\, j, f_{ij} = f_{i.} \times f_{.j}$$

$$\frac{f_{ij}}{f_{i.}} = f_{.j}$$

$$\frac{f_{ij}}{f_{.j}} = f_{i.}$$

Indépendance si on a :

→ Probabilité conditionnelle = probabilité marginale

3. Liaison entre deux variables qualitatives

- Ecart (distance) entre les données observées (f_{ij}) et le modèle d'indépendance qui est ($f_{i.}\ f_{j.}$)

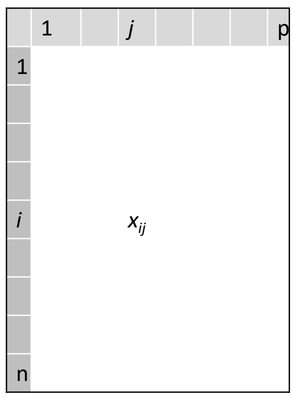
$$\chi^2{}_{obs} = \sum_{i=1}^{L}\sum_{j=1}^{J}\frac{(eff.obs - eff.théo)^2}{eff\ théo} = \sum\nolimits_{i=1}^{L}\sum\nolimits_{j=1}^{J}\frac{\left(n\, f_{ij} - n\, f_{i.}\ f_{.j}\right)^2}{n\, f_{i.}\ f_{.j}}$$

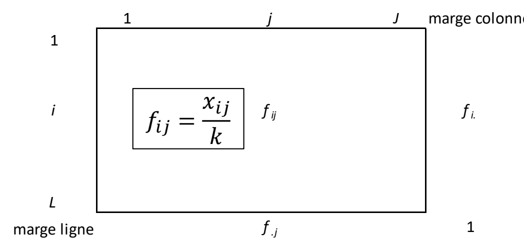
$$\chi^2{}_{obs} = \sum_{i=1}^{L}\sum_{j=1}^{J}\frac{(probabilité\ obs - probabilité\ théo)^2}{probabilité\ théo} = n\,\Phi^2$$

Φ^2 Est écart entre probabilité théorique et observée = intensité de liaison (nature de la liaison entre deux var.)

4. Du tableau de contingence à la table des fréquences

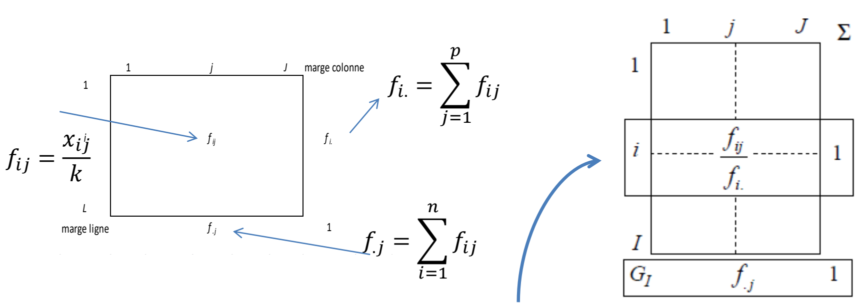
- Du tableau de correspondance à la table des fréquences





4.1 Profils-lignes et profils-colonnes : profil- ligne et profil moyen

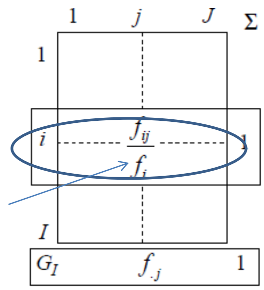
- Passage à la répartition des pourcentages à l'intérieur d'une ligne ou d'une colonne :tableau des profils-lignes et profils-colonnes

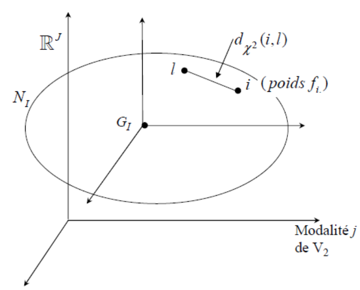


Profil ligne *i* = distribution conditionnelle : répartition de la var Y par rapport à la modalité i de la var *X* : somme des probabilités conditionnelles d'avoir *Y= j* sachant que *X=i* *(somme de ces probabilités = 1 en ligne)*
Profil moyen ligne G : répartition sur l'ensemble de la population de la variable Y avec $f_{.j} = f_{.j}/k$

4.2 Profil-ligne et profil moyen

- Profil ligne : répartition de la variable Y en fonction de la modalité i de la var X
- La proximité entre les points lignes (**espace R^p**) donc une distance entre les points lignes (profils lignes)

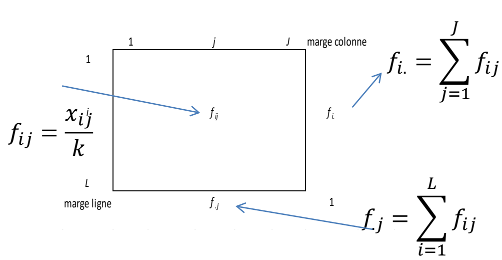


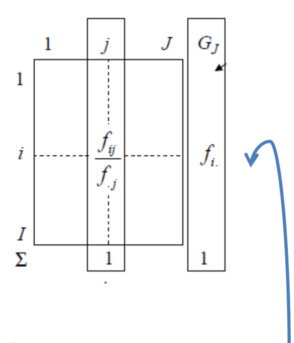


- Un point i est affecté de la masse $f_{i.}$: fréquence relative de i modalité de la var X
- Puisque $\sum_{j=1}^{p} f_{\frac{ij}{fi.}} = 1$ les *n* points sont situés dans un espace de dim *p-1*
- Le centre de gravité de ce nuage = moyenne des profils lignes affectées de leur masses correspondantes $f_{.j}$ *: sa jème* composante vaut $f_{.j} = \sum_{i=1}^{n} fi.f_{\frac{ij}{f.j}}$
- *Appelée fréquence marginale des colonnes*

5.1 Profils-lignes et profils-colonnes : profil-colonne et profil moyen

- Passage à la répartition des pourcentages à l'intérieur d'une ligne ou d'une colonne :tableau des profils-lignes et profils-colonnes



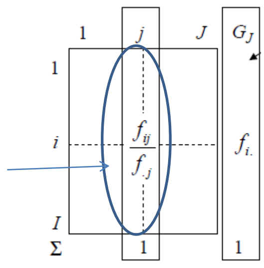


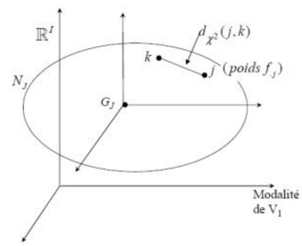
Profil colonne j = distribution conditionnelle : répartition de la var X par rapport à la modalité j de la var *Y* : somme des probabilités conditionnelles d'avoir *X= i* sachant que *Y=j = 1*

Profil moyen de la variable X : en colonne G$_J$

5.2. Profil-colonne et profil moyen

- Nuage des p profils colonnes dans R^n





- Chaque point j est affecté de la masse $f_{.j}$ dans un
- espace de dim (n-1) le centre de gravité du nuage des profils colonnes est le profil moyen de la var X; sa i ème composante est $f_{i.}$
- C'est la fréquence marginale de lignes

6. Distance entre deux points profils lignes

- Traduits les différences d'effectifs sur les deux modalité de la var X
- Distance euclidienne usuelle entre **deux profils lignes** : ressemblance ou différance entre les 2 modalités i et i' de la var X sans tenir compte des effectifs :
- $d^2(i,i') = \sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$
- Nécessité de prendre en compte les masses des colonnes

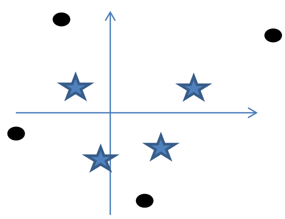
$distance\ du\ \chi 2 \quad : \quad d^2(i,i') = \sum_{j=1}^{p}\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2$

- Pour **profils colonne** : la distante devient
- $distance\ du\ \chi 2 \quad : \quad d^2(j,j') = \sum_{i=1}^{n}\frac{1}{f_{i.}}\left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2$

Distance pondérée du Chi2 propriétés remarquables + rôle de symétrie entre les lignes et les colonnes : équivalence distributionnelle (agrégation de modalités ayant le même profil sans rien changer ni aux distances entre les modalistes des deux variables x et Y (points lignes confondus) et les relations de transitions ou quais-barycentrique

7. Visualisation des nuages

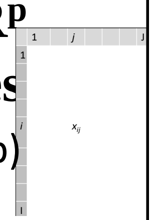
- Nuage de *n* profils lignes dans un espace à deux dimensions : réduction de R^p a R^q (2 axes factoriels)



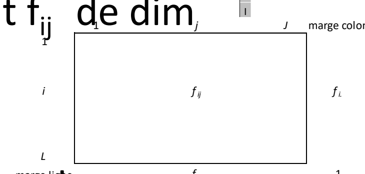
- Représentation des deux nuages dans un espace réduit

  Analyse du nuage de points pondérés dans un espace muni de la métrique du chi2 $\rightarrow$ réduction de dimension

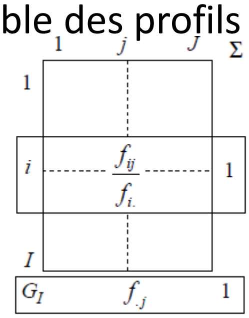
8. Réduction de dimension dans R^p et R^n : transformation des données

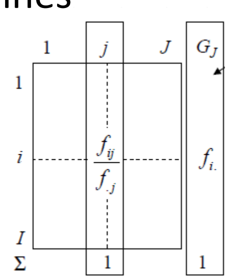


- Tableau de contingence X avec x_{ij} de dim(n,p)
- Fréquence relatives **F** d'élément f_{ij} de dim (n,p)



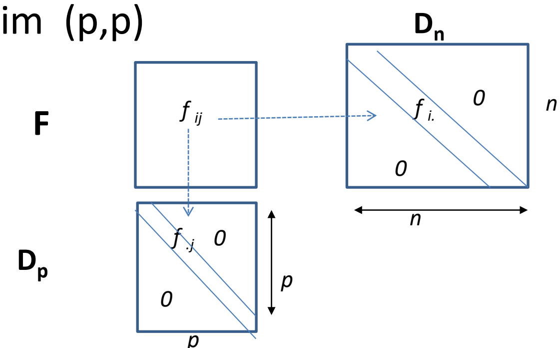
- Table des profils lignes et profils colonnes





9. Matrices de transformation

- F Matrice des fréquences relatives (n,p)
- D_n Matrice diagonale des marges en ligne $f_{i.}$ de dim (n,n)
- D_p matrice diagonale des marges en colonne $f_{.j}$ de dim (p,p)

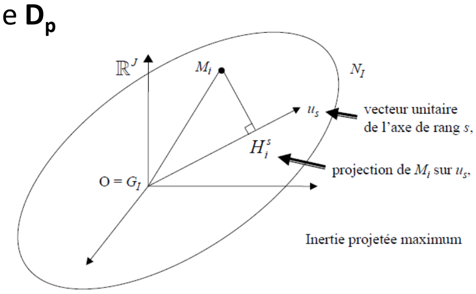


Profils lignes : D_n^{-1} F

Profils colonnes : D_p^{-1} F'

10. Réduction de dimension : Critère à maximiser

- Proximités entre profils : analyse par rapport l'origine mais possible à partir des centres de gravité
- Dans espace R^p avec analyse par rapport à l'origine : réduction de dimension
- Critère de projection orthogonale selon un axe où inertie maximale (variance) passant par O et engendré par un vecteur unitaire u et de métrique $\mathbf{D_p}$



11. Matrice à diagonaliser

- Maximiser la somme pondérée des carrés des projections sur l'axe de vecteur unitaire u soit

$$Max\left\{\sum_i f_{i.}\, d^2(i,O)\right\}$$

Soit maximiser la somme des distances au carré de Omi

Soit rendre max la quantité

$\mathbf{u' D^{-1}_p F' D^{-1}_n F D^{-1}_p u}$

Avec la contrainte $\boldsymbol{u'} \mathbf{D^{-1}_p u = 1}$

Alors u et vecteur propre la matrice

$\mathbf{S = F' D^{-1}_n F D^{-1}_p}$ de terme générale : $s_{jj\prime} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij\prime}}{f_{i.} f_{.j\prime}}$

12. Axes factoriels et coordonnées factorielles

Dans Rp		Dans Rn
$\mathbf{S = F' D^{-1}_n F D^{-1}_p}$	Matrice à diagonaliser	$\mathbf{T = F D^{-1}_p F' D^{-1}_n}$
$\mathbf{Su_\alpha = \lambda_\alpha u_\alpha}$	Axe factoriel	$\mathbf{Tv_\alpha = \lambda_\alpha v_\alpha}$
$\psi_\alpha = \mathbf{D^{-1}_n F D^{-1}_p u_\alpha}$	Coordonnées factorielles	$\varphi_\alpha = \mathbf{D^{-1}_p F' D^{-1}_n v_\alpha}$
$\psi_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.} f_{.j}} u_{\alpha i}$		$\varphi_{\alpha i} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{i.} f_{.j}} v_{\alpha i}$

Analyse factorielle sur les deux *espaces (Pagès 2008)*

Inerties (= valeurs propres). Très particulières en AFC.

Données : reconnaissance de trois saveurs (sucré, acide, amer)
Pour chaque saveur, on a demandé à dix personnes de reconnaître
la saveur d'une solution qui leur était présentée.

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	8	2
Amer	0	4	6

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	9	1
Amer	0	3	7

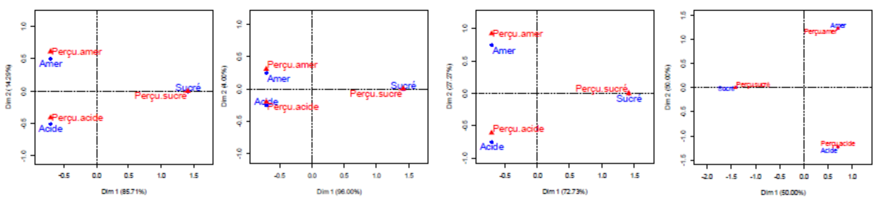
	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	10	0
Amer	0	0	10

AFC	V. Propre	%
Axe 1	1	85,714
Axe 2	0,167	14,286
Somme	1,167	100

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100

AFC	V. Propre	%
Axe 1	1	72,727
Axe 2	0,375	27,273
Somme	1,375	100

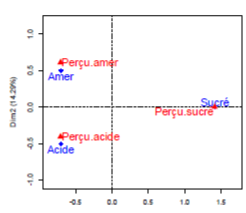
AFC	V. Propre	%
Axe 1	1	50
Axe 2	1	50
Somme	2	100



Qualité de représentation des points

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	8	2
Amer	0	4	6

AFC	V. Propre	%
Axe 1	1	85,714
Axe 2	0,167	14,286
Somme	1,167	100

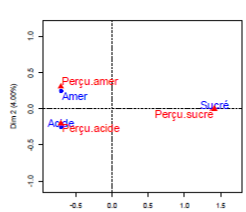


Qualité de représentation (cosinus²)

	Axe 1	Axe 2
Sucré	1.000	0.000
Acide	0.667	0.333
Amer	0.667	0.333
Perçu.sucré	1.000	0.000
Perçu.acide	0.750	0.250
Perçu.amer	0.571	0.429

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100



Qualité de représentation (cosinus²)

	Axe 1	Axe 2
Sucré	1.000	0.000
Acide	0.889	0.111
Amer	0.889	0.111
Perçu.sucré	1.000	0.000
Perçu.acide	0.923	0.077
Perçu.amer	0.842	0.158

In Pages 2015

Aides à l'interprétation (communes aux méthodes factorielles)

Contribution d'un point i à l'inertie d'un axe s

Indicateur brut : inertie projetée du point

$$f_{i.}\left(OH_i^s\right)^2$$

Indicateur relatif (en %) : inertie projetée du point/inertie totale de l'axe

$$\frac{f_{i.}\left(OH_i^s\right)^2}{\lambda_s} \quad \times 100$$

Pourcentage d'inertie

On peut additionner les contributions de plusieurs éléments.

Elles indiquent dans quelle mesure on peut considérer qu'un axe est dû à un élément ou à quelques éléments.

Les contributions réalisent un compromis opérationnel entre distance à l'origine et poids.

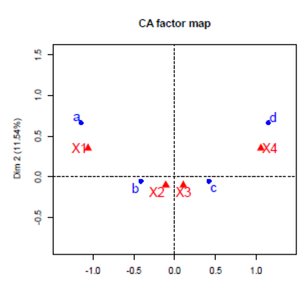
Dans le cas de grands tableaux, elles sont utiles pour sélectionner un sous-ensemble d'éléments pour commencer l'interprétation (conjointement à la qualité de représentation).

(Pagès 2008)

Contributions : exemple

	X1	X2	X3	X4
a	1	1	0	0
b	5	10	10	0
c	0	10	10	5
d	0	0	1	1

	Inertie	%
Axe 1	0.258	83.501
Axe 2	0.036	11.538
Axe 3	0.015	4.96



	Axe1	Axe2
a	18.879	46.296
b	31.121	3.704
c	31.121	3.704
d	18.879	46.296
Σ	100	100

(Pagès 2008)