Republique Algerienne Démocratique & Populaire.

Ministère De L'Enseignement Supérieur Et de La Recherche Scientifique

Université Hassiba Benbouali - Chlef

Faculté des Sciences Exactes & Informatique

Département de Mathématiques

Module : Analyse de données

Chap02 : Analyse Factorielle des Correspondances

Cours déstiné aux étudiants de Master Mathématiques
(en cours de rédaction)

Présenté par Dr :HAMEL Elhadj [1]

2020/2021

1. e-mail : hamel_2@yahoo.fr

Table des matières

Chapitre 1

Analyse Factorielle des Correspondances

1.1 Introduction

Nous avons vu au Chapitre précédent comment d'écrire des données numériques de façon multi-dimensionnelle et on a expliquer la méthode ACP. L'objet de ce chapitre concerne des variables non plus quantitatives mais qualitatives. Nous nous limiterons dans ce chapitre au cas de deux variables, que l'on notera X_1 et X_2, ayant chacune un nombre respectif de modalités égal à n_1 et n_2. avec $n = n_{..} = n_1 + n_2 = n$ la taille de la population étudiée.

1.2 Les objectifs

Les objectifs sont les mêmes que ceux de l'ACP dans le sens où l'AFC cherche donc à obtenir une typologie des lignes et une typologie des colonnes, puis de relier ces deux typologies. Il faut donc faire ressortir un bilan des ressemblances entre lignes (respectivement colonnes) en répondant aux questions du type :

- Quels sont les lignes (respectivement colonnes) qui se ressemblent ?
- Quelles sont celles qui sont différentes ?
- Existe-t-il des groupes homogènes de lignes (respectivement colonnes) ?
- Est-il possible de mettre en évidence une typologie des lignes (respectivement des colonnes) ?

La notion de ressemblance entre deux lignes ou deux colonnes diffère cependant de l'ACP. En effet, deux lignes (respectivement deux colonnes) sont proches si elles s'associent de la même façon à l'ensemble des colonnes (respectivement des lignes), i.e. elles s'associent trop ou trop peu par rapport à l'indépendance. Il faut donc chercher les lignes (respectivement colonnes) dont la répartition s'écarte le plus de l'ensemble de la population, celles qui se ressemblent entre elles et celles qui s'opposent. Afin de relier la typologie des lignes avec l'ensemble des colonnes, chaque groupe de lignes est caractérisé par les colonnes auxquelles ce groupe s'associe peu ou

fortement. Par symétrie, chaque groupe de colonnes est caractérisé par les lignes auxquelles ce groupe s'associe peu ou fortement. Ainsi nous pouvons décomposer la liaison entre deux variables en une somme de tendances simples et interprétable et mesurer leur importance respective.

1.3 Quelques définitions

Définition 1.3.1. Table de contingence - marges *:*
Un tableau de contingence est un tableau d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus.
Les données obtenues lors de l'étude d'une population à l'aide de deux variables qualitatives sont représentées sous la forme d'un tableau à n_1 lignes et n_2 colonnes (ou inversement), chaque ligne correspondant à une modalité d'une des deux variables et chaque colonne à une modalité de l'autre variable. Chaque case (i,j) de ce tableau comporte le nombre d'individus possédant la modalité i pour la première variable et j pour la seconde :

$$T = (n_{ij})_{1 \leq i \leq n_1, 1 \leq j \leq n_2}$$

Les n_{ij} sont appelés ***les effectifs*** *dits* **conjoints.**
On note que les tables de contingence sont le résultat de ce qu'appellent les praticiens des enquêtes, les tables croisés.
La somme des effectifs des lignes et des colonnes, notées respectivement $n_{.j}$ et $n_{i.}$ sont appelées les marges en colonnes et les marges en lignes. Elles sont représentées dans le tableau ci-dessus dans "Total". Nous avons

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

$T =$

	1	$\dots$	j	$\dots$	n_2	$n_{i.}$
1						.
$\vdots$			$\vdots$			
i		$\dots$	x_{ij}	$\dots$		.
$\vdots$			$\vdots$			
n_1						
$n_{.j}$			.			$n = n_{..}$

Exemple 1.3.1. *Le tableau suivant représente le type d'études poursuivies (université, classes préparatoires, autres) en fonction du parcours suivi au lycée (Lettres, Économie, Maths-Sciences, Technique).*

	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*

Définition 1.3.2. Tableau des fréquences *:*

c'est le tableau des fréquences relatives qui est considéré ou bien les fréquences conjointes sont données par :

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

les fréquences conjointes et fréquences marginales.

	1 $\ldots$ j $\ldots$ n_2	$f_{i.}$
1		.
$\vdots$	$\vdots$	
i	$\ldots$ $f_{ij} = \frac{n_{ij}}{n_{..}}$ $\ldots$	.
$\vdots$	$\vdots$	
n_1		
$f_{.j}$	.	1

Nous avons ainsi :

$$\sum_i f_{i.} = \sum_j f_{.j} = \sum_j \sum_j f_{ij} = 1$$

Exemple 1.3.2. *Le tableau de l'exemple précédent.*

	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*

Définition 1.3.3. 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*

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

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

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

On notera $L_i(\ell_i)$ *le i-ème profil ligne (la i-ème ligne de de tableau PL)*

$$\ell_i = (n_{i1}/n_{i.} \ldots\ldots n_{in_2}/n_{i.}) = (f_{i1}/f_{i.} \ldots\ldots f_{in_2}/f_{i.})$$

et C_j le j-ème profil colonne (la j-ème colonne de tableau PC)

$$C_j = (n_{1j}/n_{.j} \ldots\ldots n_{n_1 j}/n_{.j}) = (f_{1j}/f_{.j} \ldots\ldots f_{n_1 j}/f_{.j})$$

Exemple 1.3.3. Ecriture maricielle 1 *En notant $F = T/n$, la matrice des fréquences conjointes et $D_1 = D_r$ et $D_2 = D_c$ les matrices diagonales suivantes de taille respective $n_1 \times n_1$ et $n_2 \times n_2$.*

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

montrer que l'on a $L = D_1^{-1}.F$ et $C = D_2^{-1}.F'$.
où rappelons-le, F' correspond à la transposée de F.

1.4 AFC et indépendance

Nous avons vu que l'AFC considère un tableau de contingence ou de fréquence pour étudier les liaisons entre les deux variables à l'initiative du tableau. Nous ne pouvons plus définir les liaisons par les coefficients de corrélation comme pour l'ACP.

Définition 1.4.1. *Il y a indépendance entre les deux variables considérées si :*

$$f_{ij} = f_{i.} \times f_{.j} \quad \forall 1 \leq i \leq n_1, \quad \forall 1 \leq j \leq n_2;$$

Définition 1.4.2. *Nous disons qu'il y a liaison entre ces deux variables, ou que ces deux variables sont liées si elles ne sont pas indépendantes.*

Ainsi nous pouvons dire que :

- Si f_{ij} est supérieur au produit des marges, les modalités i et j s'associent plus que sous l'hypothèse d'indépendance. Nous dirons que les deux modalités i et j " s'attirent ".
- Si f_{ij} est inférieur au produit des marges, les modalités i et j s'associent moins que sous l'hypothèse d'indépendance. Nous dirons qu'il y a " répulsion " entre les deux modalités i et j.

Remarque 1.4.1. *Dans le cas de l'indépendance tous les profils lignes et colonnes sont identiques.*

1.5 la distance du Khi-deux

On est en présence de deux nuages de points pesants. Le nuage des (n_1) profils lignes, chaque profil ligne i est muni du poids $f_{i.}$. Le nuage n_2 des profils colonnes, chaque profil colonne j est muni de son poids $f_{.j}$. Se pose maintenant la question de mesurer des distances Euclidiennes entre profils. Le type de distance utilisé est la distance du du Khi-deux χ^2 qui est une distance Euclidienne.

— 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 :

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

La raison essentielle pour choisir la distance du χ^2 est qu'elle vérifie la propriété suivante :

equivalence distributionelle :

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

Lien avec le χ^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 $\dfrac{d^2}{n}$ est égal à l'inertie du nuage des profils lignes (des profils colonnes).

Démonstration.

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

□

On termine en rappelant que $\overline{L}$ est la marge colonne de F.

Remarque 1.5.1. *D'autres distances pourraient être employées, cependant la distance euclidienne usuelle entre les points-lignes ou entre les points-colonnes exprimés à partir du tableau de fréquence ne traduit que les différences d'effectifs entre deux modalités. La distance euclidienne entre les profils-lignes ou entre les profils-colonnes permet de bien modéliser les ressemblances entre deux modalités.*

Cependant cette distance favorise les colonnes ayant une masse $f_{.j}$ important, C'est pour cette raison que la distance retenue dans les profils lignes (respectivement les profils colonnes) l'écart entre les profils est pondéré par l'inverse de la masse de la colonne (respectivement de la ligne). Cette distance est nommée distance du χ^2 car elle proportionnelle à la statistique du χ^2 de Karl Pearson. De plus cette distance du χ^2 possède une propriété fondamentale nommée l'équivalence distributionelle. *Cette propriété permet d'associer deux modalités d'une même variable qui possède des profils identiques en une modalité unique affectée de la somme de leurs masses, sans modifier ni les distances entre les modalités de cette variable, ni les distances entre les modalités de l'autre variable. Ainsi, si deux colonnes proportionnelles d'un tableau sont regroupées, les distances entre profils-lignes sont inchangées, et réciproquement. Ceci permet de regrouper des*

modalités voisines pour ainsi réduire le nombre de modalités et donc la complexité de l'interprétation en garantissant une certaine invariance des résultats.

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	$\frac{d^2}{n}$	$\frac{d^2}{n}$

1.6 Les nuages des deux profils

1.6.1 Le nuage des profils-lignes

Lorsque nous nous intéressons aux modalités de la première variable, il faut considérer les données comme une juxtaposition de profils-lignes. Ainsi chaque profil-ligne i peut être représenté comme un point de l'espace $(\mathbb{R}^{n_2}, D_c^{-1})$ dont chacune des n_2 dimensions représente une modalité de la seconde variable.L'utilisation de la distance entre
deux profils est celle χ^2 , elle revient à affecter le poids $\frac{1}{f_{.j}}$ à la j ème dimension de $\mathbb{R}^{n_2}$. Du fait que la somme de chaque profil-ligne est égale à 1, le nuage N_{n-1} appartient à un hyperplan, noté H_I. Pour l'AFC les poids affectés à chaque point du nuage sont imposés et ne sont pas identiques. Le point i a pour poids la fréquence marginale $f_{i.}$. Ce poids est naturel puisqu'il est proportionnel à l'effectif de la classe d'individus qu'il représente. La coordonnée du point i sur l'axe j est donnée par $\frac{f_{ij}}{f_{i.}}$.
Le nuage N_I des profils-lignes dans $\mathbb{R}^{n_2}$. Le point i a pour coordonnée sur l'axe j : $\frac{f_{ij}}{f_{i.}}$; son poids est $f_{i.}$; la distance entre deux profils est la distance du χ^2 , Le barycentre G_I du nuage N_I a pour coordonnée sur l'axe j la fréquence marginale $f_{.j}$; le nuage N_I appartient à un hyperplan noté H_I..

Le barycentre des points de N_I munis de ces poids, noté G_I, est la moyenne pondérée de tous les points sur tous les axes δ_j. La coordonnée de G_I sur l'axe δ_j est donc donnée par :

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

Le barycentre s'interprète comme un profil-moyen. Dans l'étude des lignes, il sert de référence pour étudier dans quelle mesure et de quelle façon une classe d'individus diffère de l'ensemble de la population. Ceci se fait par l'étude de l'écart entre le profil de cette classe et le profil moyen. *Ainsi l'étude de la dispersion du nuage autour de son barycentre équivaut à l'étude de l'écart entre profils et marge ou encore à l'étude de la liaison entre les deux variables.*

1.6.2 Le nuage des profils-colonnes

La construction du nuage des profils-colonnes est identique à celle du nuage des profils lignes du fait de la symétrie entre les lignes et les colonnes en AFC. Ainsi, lorsque nous nous intéressons

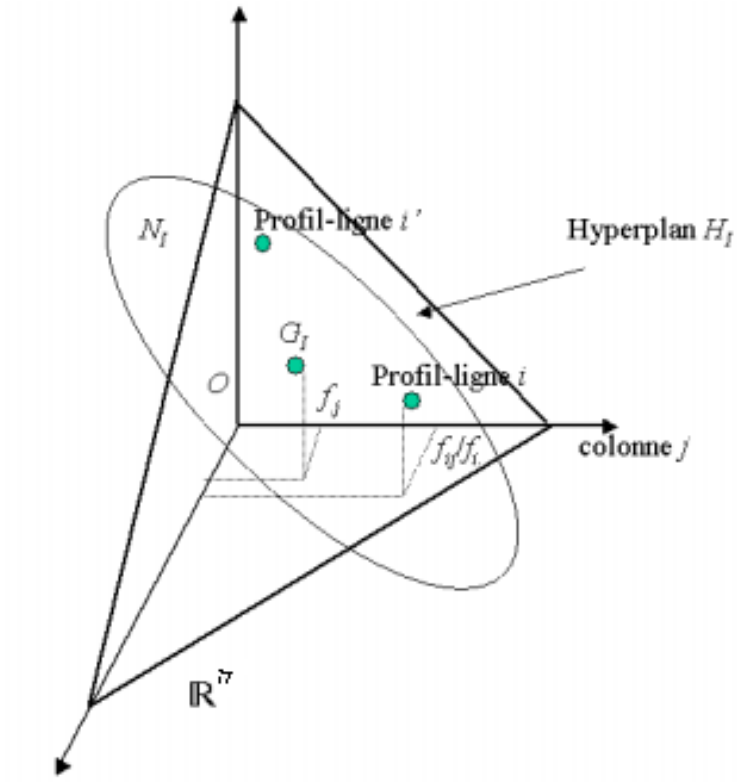

FIGURE 1.1 – Le nuage N_I des profils-lignes dans $\mathbb{R}^{n_2}$

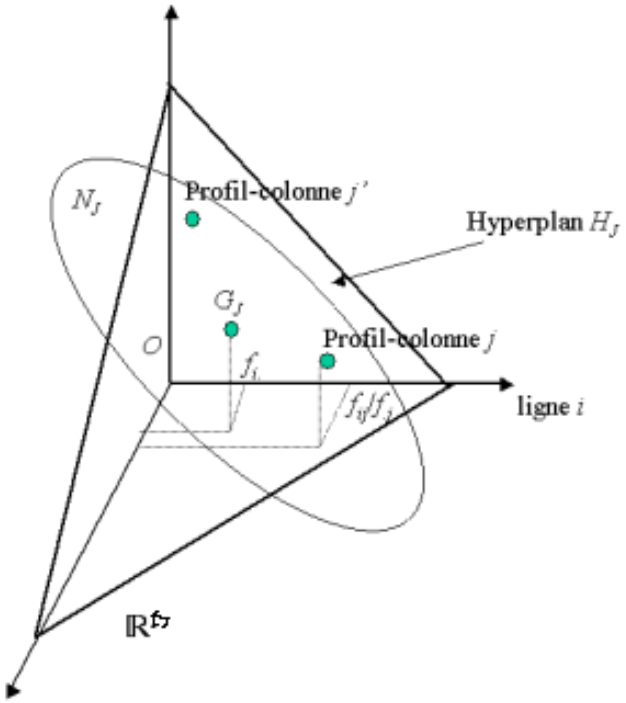

FIGURE 1.2 – Le nuage N_J des profils-colonnes dans l'espace $\mathbb{R}^{n_1}$

aux modalités de la seconde variable, il faut considérer les données comme une juxtaposition de profils-colonnes. Chaque profil-colonne j peut être représenté comme un point de l'espace $\mathbb{R}^{n_1}$ dont chacune des n_1 dimensions représente une modalité de la première variable (). Le point i a pour coordonnée sur l'axe i la proportion $\frac{f_{ij}}{f_{.j}}$, et le poids qui lui est associé est $f_{.j}$. Le nuage N_J appartient à un hyperplan noté H_J. De plus le barycentre des points de N_J munis de leur poids a pour coordonnée sur l'axe δ_i :

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

Ce barycentre s'interprète également comme un profil moyen et joue le même rôle pour l'étude de la liaison entre les deux variables.

1.7 Schéma général de l'analyse des correspondances

L'analyse des correspondances revient à effectuer l'analyse générale d'un nuage de points pondérés dans un espace muni de la métrique du χ^2. On fera donc référence à l'analyse générale avec des métriques et des critères quelconques.

1.7.1 Schéma général de l'AFC

Contrairement à l'analyse en composantes principales, le tableau de données subit deux transformations, l'une en profils-lignes, l'autre en profils-colonnes, à partir desquelles vont être construits

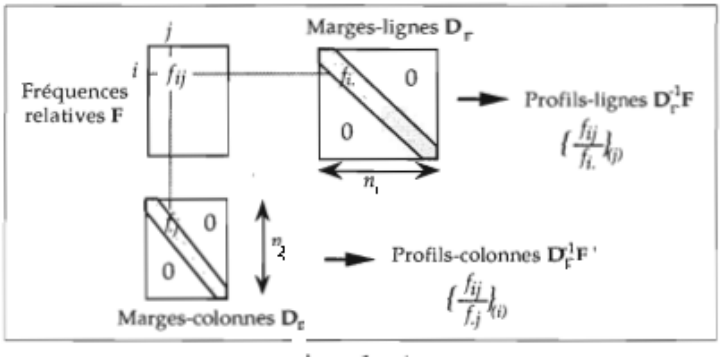

FIGURE 1.3 – Fréquences, marges, profils

les nuages de points dans $\mathbb{R}^{n_1}$ et dans $\mathbb{R}^{n_2}$ (figure **??**1.6.2).
Pour faire le lien avec l'analyse générale , nous adopterons des notations matricielles .
Les transformations opérées sur le tableau des données peuvent s'écrire à partir des trois matrices F, D_r et D_c qui définissent les éléments de base de l'analyse.
F d'ordre (n_1, n_2) désigne le tableau des fréquences relatives ; D_r d'ordre (n_1,n_1) est la matrice diagonale dont les éléments diagonaux sont les marges en lignes $f_{i.}$; D_c est la matrice diagonale d'ordre (n_2,n_2) des marges en colonnes $f_{.j}$.
Nous résumons les principaux résultats de l'AFC dans un schéma général
Les deux nuages de points (dans l'espace des colonnes et dans l'espace des lignes) sont construits de manière analogue. Nous récapitulons ici les éléments de base de l'analyse qui vont permettre la construction des facteurs.

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

1.7.2 Critère à maximiser et matrice à diagonaliser

Nous voulons représenter graphiquement les proximités entre profils.
Nous nous plaçons donc, dans les deux espaces, aux centres de gravité des nuages. Cependant, et c'est là une des particularités de l'analyse des correspondances, il est équivalent de procéder à l'analyse par rapport à l'origine ou par rapport aux centres de gravité, à condition de négliger dans le premier cas l'axe factoriel qui joint l'origine au centre de gravité (Cet axe est associé à la valeur propre égale à l, appelée valeur propre triviale) .
Plaçons-nous dans l'espace des colonnes $\mathbb{R}^{n_2}$ et cherchons l'axe d'inertie maximum du nuage des points-lignes passant par l'origine $\circ$ et engendré par un vecteur-unitaire u pour la métrique

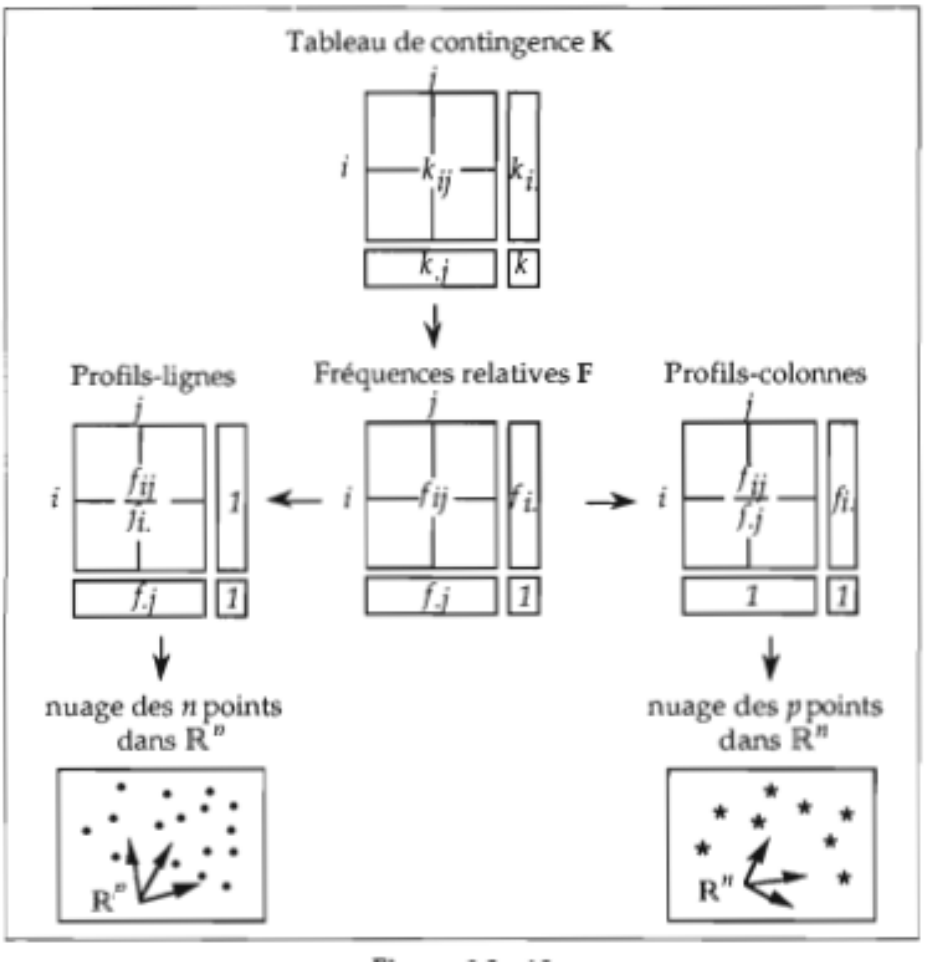

FIGURE 1.4 – Schéma général de l'afc

$M = D_c^{-1}$. Ceci nous amène à maximiser la somme pondérée des carrés des projections sur l'axe c'est-à-dire :

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

et à rendre maximale la quantité :

$$u' D_c^{-1} F' D_r^{-1} F D_c^{-1} u$$

avec la contrainte :

$$u' D_c^{-1} u = 1$$

u est vecteur propre de la matrice :

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

associé à la plus grande valeur propre λ. différente de **1** .
La matrice à diagonaliser est donc la matrice **S** de terme général :

$$s_{jj'} = \sum_i \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}}$$

De la même façon, on doit rendre maximum dans $\mathbb{R}^{n_1}$, la quantité :

$$v' D_r^{-1} F D_c^{-1} F' D_r^{-1} v$$

avec la contrainte :

$$v' D_r^{-1} v = 1$$

v est vecteur propre de la matrice :

$$\boxed{T = F D_c^{-1} F' D_r^{-1}}$$

1.7.3 Axes factoriels et facteurs

Nous supposons ici que $n_2 < n_1$ correspond à la plus petite dimension du tableau de données. Après avoir écarté la valeur propre triviale égale à 1 et le vecteur propre associé, nous retenons, de la diagonalisation de la matrice, les $n_2 - l$ valeurs propres non nulles et les vecteurs propres associés. Nous obtenons ainsi au plus $n_2 - l$ axes factoriels.

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

$$\sum_{i=1}^{n_1} f_{i.}\psi_{\alpha i} = \sum_{j=1}^{n_2} f_{.j}\phi_{\alpha j} = 0$$

et de variance égale à λ_α :

$$\sum_{i=1}^{n_1} f_{i.}\psi_{\alpha i}^2 = \sum_{j=1}^{n_2} f_{.j}\phi_{\alpha j}^2 = \lambda_\alpha$$

1.7.4 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_α de **S** associé à λ_α et le vecteur propre unitaire v_α de **T** relatif à la même valeur propre, il existe les relations dites de transition :

$$\begin{cases} v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F D_c^{-1} u_\alpha, \\ u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F' D_r^{-1} v_\alpha, \end{cases}$$

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

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

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

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

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 :

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

La matrice de terme général $\dfrac{f_{ij}}{f_{i.}}$ 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.

Remarque 1.7.1. *Toutes les valeurs propres sont nécessairement inférieures ou égales à **1**. En effet puisque :*

$$\sqrt{\lambda_\alpha}\psi_{\alpha i} = \sum_{j=1}^{n_2} \frac{f_{ij}}{f_{i.}}\phi_{\alpha j}$$

on a :

$$\min_j\{\phi_{\alpha j}\} \leq \sqrt{\lambda_\alpha}\psi_{\alpha i} \leq \max_j\{\phi_{\alpha j}\}$$

d'où :

$$\max_i\{\sqrt{\lambda_\alpha}\psi_{\alpha i}\} \leq \max_j\{\phi_{\alpha j}\}$$

De la même manière, on a :

$$\max_j\{\sqrt{\lambda_\alpha}\phi_{\alpha j}\} \leq \max_i\{\psi_{\alpha i}\}$$

comme $\lambda_\alpha \geq 0$ *:*

$$\max_j\{\sqrt{\lambda_\alpha}\phi_{\alpha j}\} \leq \max_j\{\phi_{\alpha j}\}$$

et finalement :

$$\lambda_\alpha \leq 1.$$

1.7.5 Représentation simultanée

Les relations de transition où quasi-barycentriques justifient la représentation simultanée des lignes et des colonnes. La figure ci-dessous illustre schématiquement le processus de l'analyse des correspondances.

1.8 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, en particulier pour apprécier les proximités, identifier les éléments responsables de la formation des facteurs et ceux qui en sont des caractéristiques. Ces règles s'appuient sur le bilan de l'opération de réduction que constitue la séquence des valeurs propres et des pourcentages d'inertie, ainsi que sur un ensemble de coefficients classiques : les contributions absolues et les cosinus carrés ;

1.8.1 Inertie et test d'indépendance

L'inertie totale I du nuage de points par rapport au centre de gravité s'écrit par définition :

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

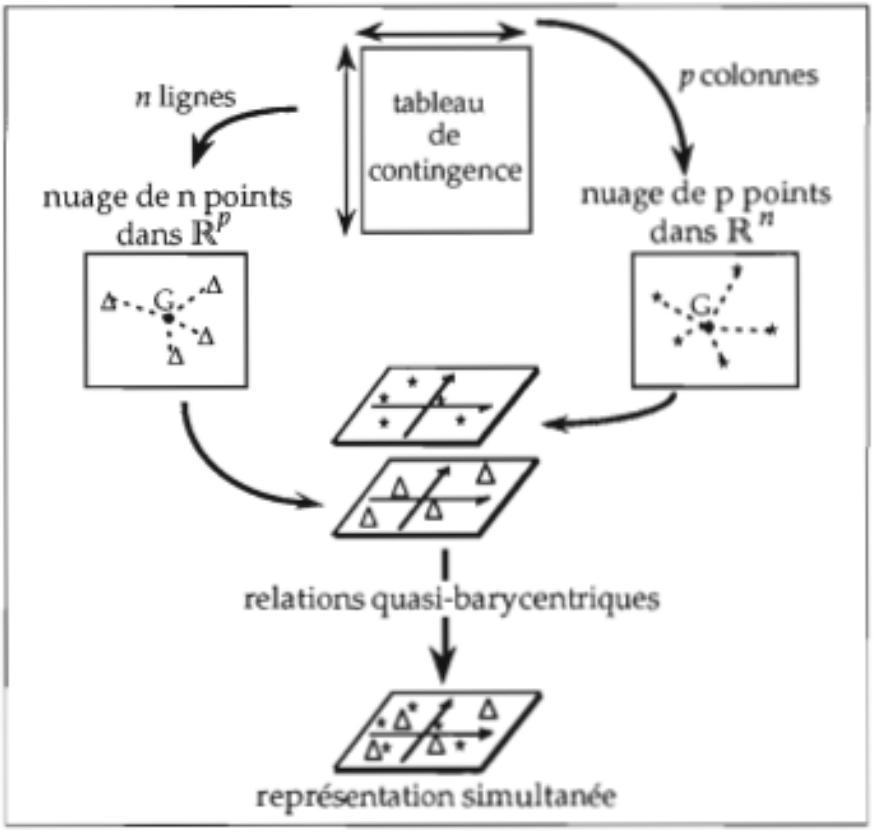

FIGURE 1.5 – Schéma de la représentation simultanée

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. On pourra rejeter l'hypothèse nulle d'indépendance des variables en lignes et en colonnes si la valeur observée χ^2 dépasse la valeur χ_0^2 qui a une probabilité d'être dépassée inférieure à un seuil fixé au préalable.

1.8.2 contributions et cosinus

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

L'origine étant prise au centre de gravité, les coordonnées factorielles sont centrées et la variance vaut λ_α . Ainsi le quotient :

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

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

Ce quotient est appelé contribution de l'élément i à l'axe α et permet de savoir dans quelle proportion un point i contribue à l'inertie λ_α du nuage projeté sur l'axe α.
On notera que pour tout axe a :

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

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

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

avec la relation :

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

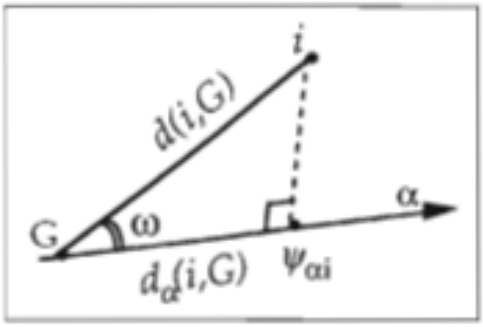

FIGURE 1.6 – Projection du point i sur l'axe α

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 :

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

Cette quantité, appelée *cosinus carré*, représente la part de la distance au centre prise en compte dans la direction α. On l'appelle aussi *la contribution relative* du facteur à la position du point i. Plus le cosinus carré est proche de 1, plus la position du point observé en projection est proche de la position réelle du point dans l'espace .
Notons que pour tout i :

$$\sum_\alpha Cos^2_\alpha(i) = 1$$

La même procédure pour les points colonnes On mesure la contribution relative du facteur α à la position du point j par le cosinus carré de j :

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

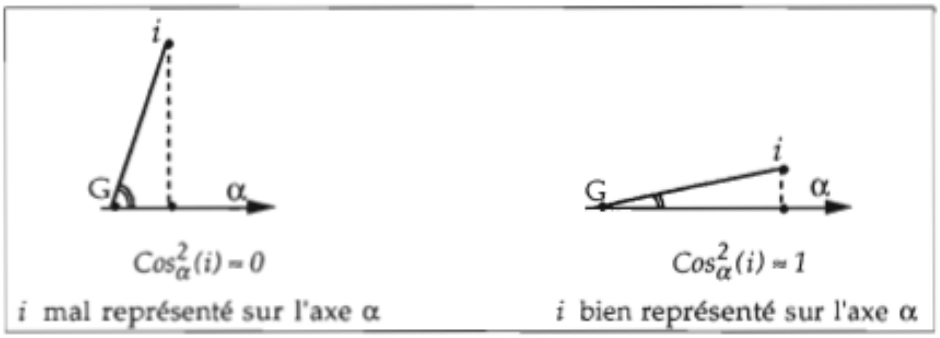

FIGURE 1.7 – Qualité de représentation d'un point i sur l'axe α

et l'on a également pour tout j :

$$\sum_{\alpha} Cos^2_\alpha(j) = 1$$

1.9 Exemple

1.10 Exercice