République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abou Bekr Belkaid Tlemcen

Faculté des sciences
Département de Mathématiques

Mémoire de Master

présentée par

Taleb Farah

Soutenue le : 22 juin 2023

Théme :

L'analyse factorielle des correspondances et applications sur R

Soutenue devant le jury composé de :

Mr. A.Allam	M.C.A, Université de Tlemcen	Président
Mme. S.Boukhiar	M.C.B, Université de Tlemcen	Examinatrice
Mme. W.Benyelles	M.C.B, Université de Tlemcen	Encadrant

Année universitaire : 2022 - 2023

Remerciements

Je remercie Dieu Tout Puissant Miséricordieux qui grâce a sa Clémence m'a donné la volonté nécessaire pour clôturer ce modeste travail après de longues années d'arrêt.

Je tiens à remercier et à exprimer ma plus grande gratitude

***Mr A.ALLAM**, Maître de Conférence à l'université de TLEMCEN de l'honneur qu'il me fait de présider le jury de mon mémoire.*

*Je tiens à remercier sincèrement **Mme S.BOUKHIAR**, Maître de Conférence à l'université de TLEMCEN d'avoir acceptée d'être membres du jury et d'être l'examinatrice de mon travail.*

Je tiens a remercier en particulier mon Encadrante

***Mme W.BENYELLES**, Maître de Conférence à l'université de TLEMCEN pour ses précieux conseils, sa disponibilité et son aide permanente durant toute la période de la conception de ce travail.*

Ma gratitude va vers le Département de Mathématiques de l'université de TLEMCEN, et à ces enseignants d'avoir enrichis mes connaissances de leurs savoirs.

Ainsi que Toutes les Personnes qui m'ont soutenu et aidé dans la réalisation de ce modeste travail.

Dédicaces

Ce n'est pas parce que c'est difficile que nous n'osons pas, c'est parce que nous n'osons pas que c'est difficile.

"Sénéque - Phylosophe Romain"

A mon cher père qui m'a soutenu tout le long de ma jeune vie et dont le soutien infaillible m'a permis d'être ce que je suis.

A ma mère qui m'a comprise et encouragé tout le long de ma jeune vie.

A mes frères **"Ghouti"** et **"Ridha"** qui m'ont toujours encouragé.

A **"Ines"**, **"Sabri"**, **"Mouna"** et **"Ramzi"** et à toute ma famille.

A mes Grands Parents qui auront été fiers de mon parcours.

Table des matières

Liste des tableaux

Table des figures

Introduction

L'analyse des données est un sous domaine des statistiques qui se préoccupe de la description des données conjointes et multivariées [1, 3, 4, 8]. On cherche par ces méthodes à décrire les liens pouvant exister entre les différentes variables et d'en tirer les informations statistiques qui servent à détailler les principales informations contenues dans ces données.

Selon la nature des variables du tableau de données (qualitatives, quantitatives, $\cdots$ etc) on distingue plusieurs méthodes : l'analyse en composante principale (ACP), l'analyse factorielle des correspondances (AFC) et l'analyse factorielle discriminante (AFD).

Dans ce travail nous aborderons l'analyse factorielle des correspondances [2, 7] (correspondence analysis), cette méthode a été développée en France essentiellement par J.-P. Benzécri durant la période 1970-1990.

L'analyse factorielle des correspondances est une méthode de réduction de dimension, elle permet d'analyser les informations contenues dans un tableau de contingence. Cette méthode consiste à étudier la liaison ou la correspondance existante entre deux variables qualitatives. L'analyse factorielle des correspondances est une extension de l'analyse en composantes principales, basée sur la distance du chi-deux.

Notre travail fait l'objet de trois chapitres, deux chapitres sur la partie théorique et un chapitre sur la partie pratique.

Le premier chapitre est consacré à la description des données et leurs caractéristiques, avec un passage décrivant l'analyse en composantes principales [10], c'est une méthode fondamentale en statistique descriptive multidimensionnelle. Elle permet de traiter simultanément un nombre quelconque de variables quantitatives et de résumer l'information en un nombre de composantes plus limitées que le nombre d'origine en générale 2 ou 3 afin de pouvoir visualiser graphiquement ces liaisons.

Dans le deuxième chapitre on va traiter l'analyse factorielle des correspondances [5, 9, 11], en expliquant le principe de cette méthode et montrer son interaction avec l'analyse en composantes principales. On interprétera ainsi les résultats de l'analyse

factorielle des correspondances.

Ce travail sera clôturé par un exemple pratique portant sur des données réelles d'un groupe de 1000 sujets féminins sur les fragrances de parfums.

On essayera de donner les clefs et les commandes nécessaires afin de pouvoir écrire un programme à partir du logiciel R [6, 12], et, ainsi de visualiser et d'interpréter ces données matricielles.

Chapitre 1

Notions générales

1.1 Introduction

L'analyse des données est une famille de méthodes statistiques se préoccupant principalement de la description de données appelées généralement "statistique multivariée"[4].

Par ces méthodes on cherche à montrer les liens pouvant exister entre les différentes données. De là, on arrive à tirer une information par l'utilisation de méthodes statistiques [8] comme : l'Analyse en Composantes Principales (ACP), et l'Analyse Factorielle des Correspondances (AFC), ...etc.

On consacre cette partie à la description des données et à leurs caractéristiques tels que : les individus, les variables , ...etc, ainsi qu'à l'introduction de la méthode de l'analyse en composantes principales nécessaires dans la définition de l'AFC.

1.1.1 Les données et leurs caractéristiques

Soit X un tableau rectangulaire de n lignes et p colonnes rassemblant les observations de p variables sur n individus.

$$X = \begin{array}{c} \\ e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & \cdots & X_j & \cdots & X_p \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \end{array}$$

Où x_{ij} la valeur de la variable X_j observée sur l'individu e_i.

Individus et variables :

◇ Les individus sont représentés par les lignes du tableau X, chaque individu est décrit par p composantes représentées par un vecteur de dimension $\mathbb{R}^p$ noté e_i tel que :

$$e_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{ip})^t \in \mathbb{R}^p; \quad i = 1, \cdots, n.$$

◇ Les variables sont représentées par les colonnes du tableau X, chaque variable est décrite par n valeurs représentées par un vecteur de dimension $\mathbb{R}^n$ noté X_j tel que :

$$X_j = (x_{1j}, \cdots, x_{ij}, \cdots, x_{nj})^t \in \mathbb{R}^n; \quad j = 1, \cdots, p.$$

On distingue divers types de variables selon la nature des données. Ainsi, une variable peut être qualitative ou quantitative.

Variables quantitatives :

Une variable quantitative ne peut pas être associée à une qualité, c'est un facteur ou une propriété quantifiable qui prend des valeurs numériques, telles que l'âge d'une personne ou le taux de glycémie dans le sang.

Variables qualitatives :

Les variables qualitatives sont des variables caractérisant l'appartenance de l'individu à un groupe (ou une catégorie). Elles expriment une caractéristique ou une qualité observable du sujet étudié. Comme par exemple, la couleur des yeux d'une personne ou encore son état civil.

1.1.2 Matrice de poids

Les individus jouent le même rôle dans la plupart des cas, Nous avons choisi implicitement cette situation, en affectant le même poids à chaque individu.

Par commodité, on choisit ces poids tels que la masse totale de ces individus soit égale à 1 ($\sum_{i=1}^{n} p_i = 1$), à chaque individu on associe alors le poids $\frac{1}{n}$.

Pour certaines applications [10] on travaille avec des poids p_i éventuellement différents d'un individu à l'autre, ils sont regroupés dans une matrice diagonale.

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & p_n \end{pmatrix}$$

Dans le cas ou tous les poids sont égaux on a $D = \frac{1}{n}\,\mathrm{Id}_n$.

Avec,

$$\mathrm{Id}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix}$$

1.1.2.1 Point moyen ou centre de gravité

Le vecteur g des moyennes arithmétiques de chaque variable $g = (\bar{X}_1, \bar{X}_2, ..., \bar{X}_p)^t$ définit le point moyen, ou centre de gravité du nuage.

◦ $\bar{X}_j$ désigne la moyenne de la variable X_j avec :

$$\bar{X}_j = \sum_{i=1}^{n} p_i x_{ij}$$

• Si $p_i = \frac{1}{n}, \quad \forall i \in \{1, \cdots, n\}$ alors $\bar{X}_j = \frac{1}{n} \sum\limits_{i=1}^{n} x_{ij}$.

La forme matricielle étant comme suit :

$$g = X^t D 1_n$$

En effet,

$$X^t D 1_n = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & p_n \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = g$$

◦ 1_n désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1.

Soit Y le tableau des variables aléatoires centrées associées à X avec :

$$y_{ij} = x_{ij} - \bar{X}_j$$

Alors on aura :

$$\begin{aligned} Y &= X - 1_n g^t \\ &= X - 1_n 1_n^t DX \\ &= (\mathrm{Id}_n - 1_n 1_n^t D)X \end{aligned}$$

1.1.3 Matrice de variance-covariance et matrice de corrélation

1.1.3.1 Matrice de variance-covariance

La matrice de variance-covariance est une matrice carrée symétrique d'ordre p qui collecte les variances dans la diagonale principale et les covariances dans les éléments extérieurs à la diagonale principale, notée V.

$$V = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ s_{p1} & \cdots & \cdots & s_p^2 \end{pmatrix}$$

• Si $p_i = \frac{1}{n}$, $\forall i \in \{1, \cdots, n\}$ la matrice de variance-covariance d'un tableau centré Y est définie par :

$$V = \frac{1}{n} Y^t Y$$

• Si $p_i \neq p_j$, i ≠ j on peut écrire V sous forme matricielle :

$$V = Y^t DY = X^t DX - gg^t$$

En effet ;

$$\begin{aligned} V &= Y^t DY \\ &= (X - 1_n g^t)^t D(X - 1_n g^t) \\ &= (X^t - g1_n^t) D(X - 1_n g^t) \\ &= X^t DX - (X^t D1_n)g^t - g(1_n^t DX) + g(1_n^t D1_n)g^t \\ &= X^t DX - gg^t \qquad car \quad 1_n^t D1_n = \sum_{i=1}^{n} p_i = 1 \end{aligned}$$

Cette formule est utilisée pour les calculs numériques.

Standardisation des données :

La standardisation est sans doute la transformation la plus efficace quand on veut comparer deux variables quantitatives.

Soit $D_{\frac{1}{s}}$ la matrice diagonale des inverses des écarts types.

$$D_{\frac{1}{s}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \frac{1}{s_p} \end{pmatrix}$$

Avant de commencer à analyser les données, il est pratique de transformer le tableau initial X en un tableau standard Z qui contient des données centrées et réduites tel que :

$$z_{ij} = \frac{x_{ij} - \bar{X}_j}{s_j}$$

Donc :

$$Z = YD_{\frac{1}{s}}$$

∘ s_j désigne l'écart type de la variable X_j avec :

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{X}_j)^2$$

1.1.3.2 Matrice de corrélation

La matrice de corrélation est une matrice carrée symétrique qui regroupe les corrélations de plusieurs variables entre elles, dont les termes diagonaux valent 1. Ses coefficients indiquent l'influence que les variables ont les unes sur les autres, elle est notée R.

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & \cdots & \cdots & 1 \end{pmatrix}$$

On peut écrire R sous forme matricielle : $R = D_{\frac{1}{s}} V D_{\frac{1}{s}} = Z^t DZ$.

En effet,

$$D_{\frac{1}{s}} V D_{\frac{1}{s}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \frac{1}{s_p} \end{pmatrix} \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ s_{p1} & \cdots & \cdots & s_p^2 \end{pmatrix} \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \frac{1}{s_p} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & \cdots & \cdots & 1 \end{pmatrix}$$

$$= R$$

$$= D_{\frac{1}{s}} Y^t D Y D_{\frac{1}{s}} \qquad car \quad V = Y^t D Y.$$

$$= Z^t D Z \qquad car \quad Z = Y D_{\frac{1}{s}}.$$

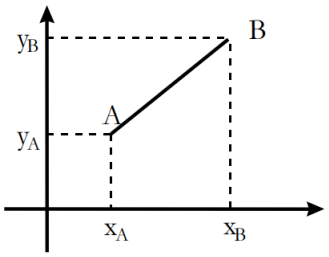
1.2 Espace des individus

On représente un individu comme un point de l'espace vectoriel à p dimension notée $\mathbb{R}^p$, dont chaque dimension représente une variable. L'ensemble des n individus constitue un nuage de points $\prec$ nuage des individus $\succ$.

Cet espace est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus.

1.2.1 Métrique

Pour mesurer la distance entre deux points on utilise la formule de Pythagore :



$$\boxed{d(A,B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}}$$

On généralisera cette notion à la distance euclidienne entre deux individus dans l'espace $\mathbb{R}^p$ qui s'écrit comme suit :

$e_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$

$e_{i'} = (x_{i'1}, x_{i'2}, \cdots, x_{i'p})$

$$\begin{aligned} d^2(e_i, e_{i'}) &= (x_{i1} - x_{i'1})^2 + (x_{i2} - x_{i'2})^2 + \cdots + (x_{ip} - x_{i'p})^2 \\ &= \sum_{k=1}^{p} (x_{ik} - x_{i'k})^2 \end{aligned}$$

Remarques :

• La formule de Pythagore n'est valable que si les individus sont représentés dans un repère orthonormé $\prec$ les axes sont perpendiculaires et de même longueur $\succ$ or ceci n'est pas vrai en réalité.

• Pour avoir la même unité on travaille avec le tableau Z au lieu de X c'est à dire on remplace les données par des données centrées et réduites.

On munit notre nuage des individus par un produit scalaire associé à la métrique M qui est une matrice carrée symétrique d'ordre p définie positive et ainsi on définit la distance entre deux individus e_i et $e_{i'}$ par :

$$\begin{aligned} d^2_M(e_i, e_{i'}) &= (e_i - e_{i'})^t M (e_i - e_{i'}) \\ &= < e_i - e_{i'}, e_i - e_{i'} >_M \\ &= ||e_i - e_{i'}||^2_M \end{aligned}$$

En théorie la métrique M dépend de l'utilisateur mais en pratique $M = \mathrm{Id}$, ce qui revient à utiliser le produit scalaire usuel.

Dans tous les logiciels [10] la métrique la plus utilisée est la métrique diagonale des inverses des variances est $D_{\frac{1}{s^2}}$.

$$M = D_{\frac{1}{s^2}} = \begin{pmatrix} \frac{1}{s_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2^2} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \frac{1}{s_p^2} \end{pmatrix}$$

Remarque :

• La métrique $M = \text{Id}$ est utilisée pour les données centrées réduites du tableau Z et la métrique $M = D_{\frac{1}{s^2}}$ pour les données du tableau initial X où les variables ne s'expriment pas avec les mêmes unités.

1.2.2 Inertie

On définit l'inertie totale d'un nuage de points par la moyenne pondérée des carrées des distances des points au centre de gravité g.

L'inertie mesure la dispersion totale du nuage de points.

$$I_g = \sum_{i=1}^{n} \frac{1}{n} d_M^2(e_i, g)$$

De façon générale :

$$I_g = \sum_{i=1}^{n} p_i d_M^2(e_i, g) \qquad avec \quad \sum_{i=1}^{n} p_i = 1$$

On peut aussi l'écrire de la façon suivante :

$$\begin{aligned} I_g &= \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) \\ &= \sum_{i=1}^{n} p_i < e_i - g, e_i - g >_M \\ &= \sum_{i=1}^{n} p_i ||e_i - g||_M^2 \end{aligned}$$

◦ Où p_i représente les poids des individus.

• Si $g = 0$ on a :

$$\begin{aligned} I_g &= \sum_{i=1}^{n} p_i e_i^t M e_i \\ &= \sum_{i=1}^{n} p_i ||e_i||_M^2 \end{aligned}$$

• Une autre définition de l'inertie [10]

$$I_g = tr(MV) = tr(VM)$$

.

⋆ Si $M = \mathrm{Id}$:

$$I_g = \sum_{i=1}^{p} s_i^2 = tr(V)$$

En effet,

• En dimension $\mathbb{R}^2$, la droite passe par le point moyen ; on place l'origine au centre de gravité en utilisant les nouvelles coordonnées centrées :

$$\begin{cases} X_i &= x_i - m_X \\ Y_i &= y_i - m_Y \end{cases}$$

$$\begin{aligned} I_g &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 + Y_i^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}((x_i - m_X)^2 + (y_i - m_Y)^2) \\ &= var(X) + var(Y) \\ &= tr(V) \end{aligned}$$

• En dimension $\mathbb{R}^p$, on a :

$$\begin{aligned} I_g &= \frac{1}{n}\sum_{i=1}^{n} ||e_i - g||^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} ||x_{ij} - \bar{X}_j||^2 \\ &= \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} ||x_{ij} - \bar{X}_j||^2 \\ &= \sum_{j=1}^{p} var(X_j) \\ &= tr(V) \end{aligned}$$

⋄ L'inertie étant aussi égale à la somme des variances des variables étudiées, avec V la matrice de variances-covariances.

⋆ Si $M = D_{\frac{1}{s^2}}$:

$$\begin{aligned} I_g &= tr(MV) \\ &= tr(D_{\frac{1}{s^2}} V) \\ &= tr(D_{\frac{1}{s}} V D_{\frac{1}{s}}) \\ &= tr(R) \\ &= p \end{aligned}$$

Remarque :

Dans le cas où les variables sont centrées réduites, la variance de chaque variable vaut 1 **l'inertie totale est alors égale au nombre de variables "p"**.

• L'inertie en un point quelconque "a" est définie par :

$$\begin{aligned} I_a &= \sum_{i=1}^{n} p_i d_M^2(e_i, a) \\ &= \sum_{i=1}^{n} p_i (e_i - a)^t M (e_i - a) \end{aligned}$$

• La relation de Huygens [10] :

$$I_a = I_g + ||g - a||_M^2$$

1.3 Espace des variables

Chaque variable X_j est associée à une suite de n valeurs numériques, elle peut être représentée comme un vecteur de l'espace vectoriel à n dimensions noté $\mathbb{R}^n$, dont chaque dimension représente un individu. L'ensemble de p variables constitue un nuage de points sur $\mathbb{R}^n$ appelé nuage des variables.

1.3.1 Métrique des variables

Pour étudier la proximité des variables entre elles, il faut munir cet espace d'une métrique, trouver une matrice d'ordre n symétrique et définie positive, le choix de cette matrice est évident et se porte sur la matrice diagonale des poids [5].

Ainsi on a :

$$\begin{aligned} < x_j, x_{j'} >_D &= \sum_{i=1}^{n} p_i x_{ij} x_{ij'} \\ &= x_j^t D x_{j'} \end{aligned}$$

⋆ Lorsque les variables sont centrées, le produit scalaire représente la covariance.

$$\begin{aligned} < x_j, x_{j'} >_D &= cov(x_j, x_{j'}) \\ &= s_{jj'} \end{aligned}$$

⋆ La norme représente la longueur d'une variable qui est égale à son écart type.

$$||x_j||_D^2 = s_j^2 \qquad \Rightarrow \qquad ||x_j||_D = s_j$$

Le produit scalaire est défini comme suit :

$$< x_j, x_{j'} >_D = ||x_j||_D ||x_{j'}||_D cos(x_j, x_{j'})$$

⋆ L'angle entre deux variables est donné par :

$$\begin{aligned} cos(x_j, x_{j'}) &= cos\theta_{jj'} \\ &= \frac{< x_j, x_{j'} >_D}{||x_j||_D ||x_{j'}||_D} \\ &= \frac{s_{jj'}}{s_j s_{j'}} \end{aligned}$$

Le coefficient de corrélation linéaire n'est autre que le cosinus d'angle entre les deux variables.

1.4 Analyse en Composantes Principales

L'analyse en composantes principales est l'une des méthodes d'analyse de données multivariées la plus fréquemment utilisée. Elle nous permet d'étudier des ensembles de données multidimensionnelles avec des variables quantitatives [3].

L'analyse en composantes principales est une méthode de base en statistique exploratoire multidimensionnelle, elle permet de représenter graphiquement les corrélations entre variables et les ressemblances entre individus, elle transforme un grand nombre de variables corrélées en un plus petit nombre de variables indépendantes les unes des autres appelées **composantes principales**.

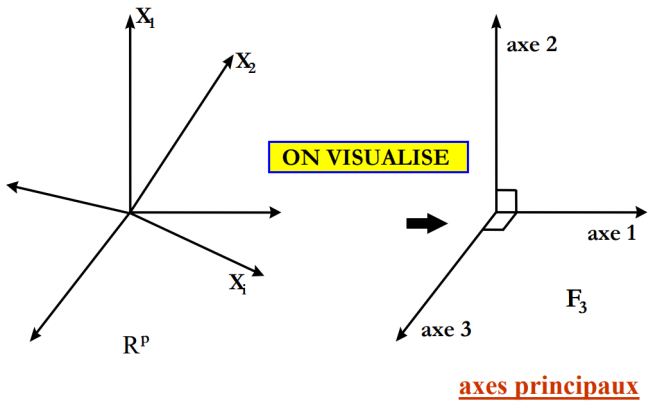
1.4.1 Principe de l'analyse en composantes principales

L'objectif de l'ACP c'est de fournir un outil de visualisation des données qui nous permet d'explorer les liaisons entre les variables et la ressemblance entre les individus en réduisant leurs dimensions afin de pouvoir construire une représentation graphique plus claire dans un sous-espace F_k de $\mathbb{R}^p$ de dimension k (k très petit 2 ou 3).

On définit k nouvelles variables qu'on appelle **composantes principales**, qui sont aussi des variables initiales contenant le plus d'informations possibles.

⋆ Les axes qu'elles déterminent sont appelés **axes principaux**.

⋆ Les formes linéaires associées sont appelées **facteurs principaux**.



1.4.2 Construction du sous-espace

Le critère de choix de l'espace de projection s'effectue tel que la moyenne des carrées des distances des projections soit la plus grande possible, il faut que l'inertie du nuage projeté sur le sous-espace F_k soit maximale.

⋆ On définit P l'opérateur de projection M-orthogonal sur le sous-espace F_k :

- $P^2 = P$ (P est idempotante).
- $P^tM = MP$ (P est M-symétrique).

⋆ Soit f_i la projection d'un individu e_i sur le sous espace F_k :

- $f_i = Pe_i$ d'où $f_i^t = e_i^tP^t$ c'est la i-ème ligne du tableau XP^t.

⋆ L'ensemble des individus sera associé au tableau initial X, le nuage projeté est alors associé au tableau X_{proj} tel que :

- $X_{proj} = XP^t$.

⋆ Le centre de gravité projeté est :

- $g_{proj} = Pg$.

En effet,

$$\begin{aligned} g_{proj} &= X_{proj}^t D 1_n \\ &= (XP^t)^t D 1_n \\ &= P(X^t D 1_n) \\ &= Pg \end{aligned}$$

⋆ La matrice de variance du tableau projeté est :

- $V_{proj} = PVP^t$.

En effet,

$$\begin{aligned} V_{proj} &= X_{proj}^t D X_{proj} - g_{proj} g_{proj}^t \\ &= PX^t D X P^t - P g g^t P^t \\ &= P(X^t D X - g g^t) P^t \\ &= PVP^t \end{aligned}$$

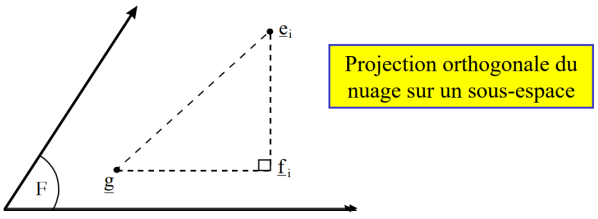
⋆ L'inertie du nuage projeté est :

- $I_{proj} = trace(VMP)$.

En effet,

$$\begin{aligned} I_{proj} &= tr(V_{proj} M) \\ &= tr(PVP^t M) \\ &= tr(PVMP) \quad car \quad P \quad \text{est M- symètrique} \\ &= tr(VMP^2) \\ &= tr(VMP) \quad car \quad P \quad \text{est idempotente} \end{aligned}$$

Déterminer l'espace de projection F_k revient à trouver P l'opérateur de projection M-orthogonal de rang k, maximisant l'inertie du nuage.



1.4.3 Axes principaux

On cherche une droite de dimension 1 passant par g qui maximise l'inertie du nuage projeté sur elle.

Soit $a_1 \in \mathbb{R}^p$ un vecteur directeur de Δ_1, P_1 est le projecteur M-orthogonale sur la droite Δ_1 tel que :

$$P_1 = a_1(a_1^t M a_1)^{-1} a_1^t M$$

Il est simple de vérifier que P_1 est un projecteur M-orthogonale, en effet :

$$\begin{aligned} P_1^2 &= a_1(a_1^t M a_1)^{-1} a_1^t M a_1 (a_1^t M a_1)^{-1} a_1^t M \\ &= a_1 \frac{a_1^t M a_1}{a_1^t M a_1} (a_1^t M a_1)^{-1} a_1^t M \\ &= a_1(a_1^t M a_1)^{-1} a_1^t M \\ &= P_1 \\ P_1^t M &= M a_1(a_1^t M a_1)^{-1} a_1^t M \\ &= M P_1 \end{aligned}$$

L'inertie du nuage projeté sur cette droite est :

$$I_{\Delta_1} = trace(VMP_1) = \frac{a_1^t MVMa_1}{a_1^t M a_1}$$

En effet,

$$\begin{aligned} trace(VMP_1) &= trace(VMa_1(a_1^t M a_1)^{-1} a_1^t M) \\ &= \frac{1}{a_1^t M a_1} trace(VMa_1 a_1^t M) \\ &= \frac{1}{a_1^t M a_1} trace(a_1^t MVMa_1) \\ &= \frac{a_1^t MVMa_1}{a_1^t M a_1} \end{aligned}$$

On pose $\frac{a_1^t MVMa_1}{a_1^t M a_1} = f(a_1)$, où f est une fonction (forme quadratique) définie sur $\mathbb{R}^p$.

Elle atteint son maximum en la dérivant par rapport à "a_1", puis en résolvant cette dernière en l'annulant.

En appliquant la règle de dérivation d'une forme quadratique par rapport à un vecteur, on obtient :

$$VMa_1 = (\frac{a_1^t MVMa_1}{a_1^t Ma_1})a_1$$

On pose, $\frac{a_1^t MVMa_1}{a_1^t Ma_1} = \lambda \in \mathbb{R}$, alors : $\quad VMa_1 = \lambda a_1$.

Où "a_1" est un vecteur propre de la matrice VM associée à la plus grande valeur propre λ.

Remarques :

• La matrice MVM est appelée matrice d'inertie du nuage, elle définit la forme quadratique associée à l'inertie projetée sur l'axe.

• Si la matrice d'inertie est $M = \text{Id}$ alors elle sera confondue avec la matrice de variance- covariance.

1.4.4 Facteurs principaux

On associe à l'axe principal a_k le facteur principal u_k [7] défini par :

$$u_k = Ma_k$$

Comme a_k est le vecteur propre de la matrice VM alors on remarque que u_k est le vecteur propre de la matrice MV car on a :

$VMa_k = \lambda_k a_k \Longrightarrow \quad MVMa_k = \lambda_k Ma_k$

$\Longrightarrow \quad MVu_k = \lambda_k u_k$

En effet, on a : $a_k^t MM^{-1}Ma_k = a_k^t Ma_k = 1$

Car a_k est M-orthonormé ce qui implique que u_k est M^{-1}-orthonormé.

Donc les facteurs principaux sont les vecteurs propres M^{-1}-normé et M^{-1} orthogonaux.

En effet, $\mathbb{R}^p$ est muni de la métrique M, son dual doit être muni de la métrique M^{-1} donc $u_k^t M^{-1}u_k = 1$ avec $u_k \in \mathbb{R}^{p*}$.

1.4.5 Composantes principales

Les composantes principales sont les nouvelles variables définies par les facteurs principaux tels que :

$$C_k = Xu_k$$

∘ X désigne la matrice utilisée dans les simulations (si les variables sont différentes on aura la matrice centrée-réduite).

∘ C_k désigne le vecteur des coordonnées de la projection M-orthogonale des n individus.

La variance d'une composante principale est égale à la valeur propre λ_k avec :

$$Var(C_k) = \lambda_k$$

En effet,

$$\begin{aligned} Var(C_k) &= C_k^t D C_k \\ &= (Xu_k)^t D(Xu_k) \\ &= u_k^t X^t D X u_k \\ &= u_k^t V u_k \\ &= u_k^t \lambda_k M^{-1} u_k \\ &= \lambda_k u_k^t M^{-1} u_k \\ &= \lambda_k \qquad car \qquad u_k^t M^{-1} u_k = 1 \end{aligned}$$

1.5 Résumé des éléments principaux d'une ACP

Axes principaux **a**	$VMa = \lambda a$	M-orthonormés
Facteurs principaux **u**	$MVu = \lambda u$	$M^{-1} - orthonormés$
Composantes principales **C**	$C = Xu$	avec $u = Ma$

Table 1.1 – Tableau des éléments principaux d'une ACP

Chapitre 2

Analyse factorielle des correspondances

2.1 Introduction

L'analyse factorielle des correspondances [2] est une méthode statistique qui permet d'étudier la liaison entre deux variables qualitatives, nous la notons AFC.

L'analyse factorielle des correspondances peut être étudiée comme une analyse des composantes principales (ACP) particulière qui utilise une métrique spéciale (celle du χ^2). Le but de cette méthode est la réduction de dimension.

2.2 Notations

2.2.1 Effectif total

L'effectif total est la somme de tous les effectifs observés tels que :

$$n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$$

2.2.2 Effectif marginal

On peut définir l'effectif marginal des lignes $n_{i.}$ et l'effectif marginal des colonnes $n_{.j}$ appelés respectivement **marges en lignes** et **marges en colonnes** comme suit :

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \qquad i = 1 \cdots p$$

$$n_{.j} = \sum_{i=1}^{p} n_{ij} \qquad j = 1 \cdots q$$

2.2.3 Fréquences

La fréquence d'une valeur, est le quotient (rapport) de l'effectif de chaque valeur n_{ij} par l'effectif total n tel que :

$$f_{ij} = \frac{n_{ij}}{n}$$

La fréquence totale est donnée par :

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

2.2.4 Fréquences marginales

On définit les fréquences marginales en lignes et en colonnes $f_{i.}$ et $f_{.j}$, respectivement par :

$$f_{i.} = \sum_{j=1}^{q} f_{ij} = \frac{n_{i.}}{n} \qquad i = 1 \cdots p$$

Et

$$f_{.j} = \sum_{i=1}^{p} f_{ij} = \frac{n_{.j}}{n} \qquad j = 1 \cdots q$$

2.2.5 Fréquences conditionnelles

On définit les fréquences conditionnelles aux profils-lignes $f_{i/j}$ par :

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

Et les fréquences conditionnelles aux profils-colonnes $f_{j/i}$ par :

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

2.3 Tableau de contingence et nuages associés

Le tableau de contingence (appelés aussi tableau de dépendance ou tableau croisé) est un moyen de représenter simultanément deux caractères observés sur une même population, c'est la matrice des effectifs observés de p lignes et q colonnes.

L'analyse factorielle est basée sur le nuage de points, le point de départ d'une AFC est le tableau de contingence obtenu en croisant les modalités de deux variables qualitatives.

Soit N^* la matrice des effectifs :

$$N^* = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1q} \\ n_{21} & n_{22} & \cdots & n_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ n_{p1} & n_{p2} & \cdots & n_{pq} \end{pmatrix}$$

Soient U et V deux variables qualitatives, p et q le nombre de modalités (catégories) décrivant un ensemble de n individus tel que :

U / V	y_1	$\cdots$	y_j	$\cdots$	y_q	marge i
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
marge j	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	n

Table 2.1 – Tableau de contingence de deux variables qualitatives

• Tableau des fréquences :

Le tableau des fréquences est un tableau à p lignes et q colonnes, noté N, définit comme suit :

U / V	y_1	$\cdots$	y_j	$\cdots$	y_q	marge i
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1q}	$f_{1.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{iq}	$f_{i.}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_p	f_{p1}	$\cdots$	f_{pj}	$\cdots$	f_{pq}	$f_{p.}$
marge j	$f_{.1}$	$\cdots$	$f_{.j}$	$\cdots$	$f_{.q}$	1

Table 2.2 – Tableau des fréquences

La matrice des fréquences observées N est de la forme :

$$N = \frac{1}{n}N^* = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1q} \\ f_{21} & f_{22} & \cdots & f_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ f_{p1} & f_{p2} & \cdots & f_{pq} \end{pmatrix}$$

2.3.1 Exemple

Pour bien illustrer ce qu'est un tableau de contingence, on peut l'expliquer par l'utilisation d'une base de donnée "CSP / Filières", qui est le résultat d'une enquête, où l'on a croisé l'origine sociale des étudiants (à travers la CSP- catégorie sociale professionnelle) avec les choix de filières à l'Université. Ce tableau est tiré de la page de cours de F-G. Carpentier de l'Université de Brest [9].

CSP / Filière	Droit	Sciences	Médecine	IUT
Exp.agri	80	99	65	58
Patron	168	137	208	62
Cadre.sup	470	400	876	79
Employé	145	133	135	54
Ouvrier	166	193	127	129

Table 2.3 – CSP et choix de filières - Tableau des effectifs observés

Cette base de donnée contient deux variables qualitatives, "CSP" comme la première variable de 5 modalités, et "Filière" comme la deuxième variable de 4 modalités, de taille $n = 3784$ individus.

2.4 Liaison entre les variables

L'analyse factorielle des correspondances a pour but d'étudier la liaison entre deux variables qualitatives (U et V), on exprime cette liaison dans un tableau de contingence, on s'intéresse à l'indépendance de ces variables. Si U et V sont indépendantes, alors l'AFC n'a aucun sens, pour cela, il faut étudier la significativité de la liaison entre les lignes et les colonnes.

On doit appliquer un test non paramétrique, on propose une hypothèse nulle et une autre alternative comme suit :

$$\begin{cases} H_0 : & \text{les variables U et V sont indépendantes (pas de correspondance)} \\ H_1 : & \text{les variables sont liées "dépendantes" (il y a une correspondance)} \end{cases}$$

On peut aussi l'écrire sous forme du test de chi-deux d'indépendance :

$$\begin{cases} H_0 : & f_{ij} = f_{i.} f_{.j} \\ H_1 : & f_{ij} \neq f_{i.} f_{.j} \end{cases}$$

Remarque :

• En analyse factorielle des correspondances, l'indépendance entre deux variables qualitatives U et V, se traduit par une proportionnalité entre les lignes et les colonnes des tableaux de contingences et des fréquences.

• S'il y a une dépendance entre les deux variables, alors, dans ce cas, l'analyse factorielle des correspondances nous sera utile afin de nous apporter les relations qui existent entre les différentes modalités de chaque variables.

2.4.1 Statistique du Chi-deux

Pour étudier la liaison entre les deux variables, il faut appliquer un test d'hypothèse. A l'aide de la statistique du χ^2, appliquée à la matrice des effectifs observés.

Cette statistique est une mesure de la différence entre les effectifs observés et les effectifs théoriques.

La statistique du χ^2 s'écrit sous la forme :

$$\chi^2 = \sum_{ij} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}}$$

Ce test a comme région critique au risque $\alpha = 0.05$:

$$D = \{|\chi^2| > \chi^2_{(p-1)(q-1)}\}$$

Remarque :

Si les deux variables qualitatives U et V sont indépendantes alors $\chi^2 = 0$.

2.5 Transformation des données

En analyse factorielle des correspondances, le tableau de contingence n'est pas analysé directement, c'est à dire au lieu de travailler avec le tableau N, on utilise d'autres tableaux qu'on appelle : **tableau des profils-lignes** et **tableau des profils-colonnes**, cette transformation découle de l'objectif qui vise à étudier la liaison entre les deux variables [5].

2.5.1 Matrices des profils-lignes et profils-colonnes

On définit, les matrices diagonales de profils-lignes et profils-colonnes, respectivement par :

$$D_l = \begin{pmatrix} f_{1.} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{p.} \end{pmatrix} \quad et \quad D_c = \begin{pmatrix} f_{.1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{.q} \end{pmatrix}$$

Avec p et q le nombre de catégories (modalités) .

2.5.2 Tableau des profils-lignes

On appelle tableau des profils-lignes, le tableau des fréquences conditionnelles $\frac{f_{ij}}{f_{i.}}$ qui est défini par :

$$X_l = D_l^{-1}N = \begin{pmatrix} \frac{f_{11}}{f_{1.}} & \cdots & \frac{f_{1q}}{f_{1.}} \\ \vdots & \ddots & \cdots \\ \frac{f_{p1}}{f_{p.}} & \cdots & \frac{f_{pq}}{f_{p.}} \end{pmatrix}$$

2.5.3 Tableau des profils-colonnes

On appelle tableau des profils-colonnes, le tableau des fréquences conditionnelles $\frac{f_{ij}}{f_{.j}}$ qui est défini par :

$$X_c = D_c^{-1}N^t = \begin{pmatrix} \frac{f_{11}}{f_{.1}} & \cdots & \frac{f_{p1}}{f_{.1}} \\ \vdots & \ddots & \cdots \\ \frac{f_{1q}}{f_{.q}} & \cdots & \frac{f_{pq}}{f_{.q}} \end{pmatrix}$$

∘ Avec N la matrice des fréquences observées.

∘ D_l et D_c les matrices diagonales à p lignes et q colonnes des fréquences marginales des variables qualitatives U et V.

Remarque :

Si on définit le profil marginal ligne et profil marginal colonne par les quantités $\frac{n_{i.}}{n}$ et $\frac{n_{.j}}{n}$ alors l'écriture matricielle sera : $\frac{D_l}{n}$ et $\frac{D_c}{n}$

2.5.4 Centre de gravité des profils

Le centre de gravité de profils-lignes et profils-colonnes, respectivement g_l et g_c est défini par :

$g_l = (f_{.1}, \cdots, f_{.q})^t \in \mathbb{R}^q$ et $g_c = (f_{1.}, \cdots, f_{p.})^t \in \mathbb{R}^p$

∗ Le centre de gravité du nuage g_l, est la moyenne pondérée de tous les points sur tous les axes j, tel que :

$$g_l = N^t 1_p$$

Où 1_p désigne le vecteur unité de $\mathbb{R}^p$ tel que $1_p = (1, \cdots, \cdots, 1)^t$.

$*$ Le centre de gravité du nuage g_c, est la moyenne pondérée de tous les points sur tous les axes i, tel que :

$$g_c = N1_q$$

Où 1_q désigne le vecteur unité de $\mathbb{R}^q$ tel que $1_q = (1, \cdots, \cdots, 1)^t$.

2.6 Métrique du chi-deux

La ressemblance entre deux lignes ou entre deux colonnes est définie par une distance entre profils. La distance employée est celle du χ^2, elle est définie de façon symétrique.

2.6.1 La distance entre deux profils-lignes

Pour calculer la distance entre deux profils-lignes i et i' on utilise la formule suivante :

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

La métrique correspondante est une matrice diagonale :

$$M_l = D_c^{-1} \quad Avec \quad D_c^{-1} = \begin{pmatrix} \frac{1}{f_{.1}} & \cdots & 0 \\ \vdots & \ddots & \cdots \\ 0 & \cdots & \frac{1}{f_{.q}} \end{pmatrix}$$

La distance du χ^2 entre le profil-ligne i et son centre de gravité g_l est donnée par la formule suivante :

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

2.6.2 La distance entre deux profils-colonnes

Par analogie la distance entre deux profils-colonnes j et j' est défini par la formule suivante :

$$d^2_{\chi^2}(j,j') = \sum_{i=1}^{p} \frac{n}{n_{i.}}(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}})^2$$
$$= \sum_{i=1}^{p} \frac{1}{f_{i.}}(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}})^2$$

La métrique correspondante est une matrice diagonale :

$$M_c = D_l^{-1} \quad Avec \quad D_l^{-1} = \begin{pmatrix} \frac{1}{f_{1.}} & \cdots & 0 \\ \vdots & \ddots & \cdots \\ 0 & \cdots & \frac{1}{f_{p.}} \end{pmatrix}$$

La distance du χ^2 entre le profil-colonne j et son centre de gravité g_c est donnée par la formule suivante :

$$d^2_{\chi^2}(j,g_c) = \sum_{i=1}^{p} \frac{1}{f_{i.}}(\frac{f_{ij}}{f.j} - f_{i.})^2$$

2.7 L'inertie totale

Nous allons définir ici la formule d'inertie totale des nuages de points profils-lignes et profils-colonnes par rapport aux centres de gravité respectivement par :

$$Inertie(X_l/g_l) = \sum_{i=1}^{p} f_{i.} d^2_{\chi^2}(i,g_l)$$

Et

$$Inertie(X_c/g_c) = \sum_{j=1}^{q} f_{.j} d^2_{\chi^2}(j,g_c)$$

L'inertie totale du nuage de point est une quantité qui mesure l'écart à l'indépendance, elle est donnée par la formule suivante :

$$\phi^2 = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

D'où,

$$\phi^2 = \frac{\chi^2}{n}$$

2.8 Analyse en composantes principales des nuages de points

L'analyse factorielle des correspondances est définie comme étant le résultat d'une double ACP, l'une portant sur les profils-lignes et l'autre sur les profils-colonnes.

2.8.1 ACP du nuage des profils-lignes et des profils- colonnes

Dans un tableau de contingence, les mots individus et variables n'ont pas la même signification que dans le tableau des données de l'analyse en composantes principales.

Dans le tableau de contingence, les lignes et les colonnes représentent les modalités de deux caractères. Pour conserver une homogénéité dans la présentation des deux analyses, les p modalités du caractère U en lignes portent le nom d'individus et les q modalités du caractère V en colonnes portent le nom de variables, c'est à dire les profils-lignes jouent le rôle des individus et les profils-colonnes le rôle des variables.

2.8.1.1 ACP du nuage des profils-lignes

* *Tableau des données :* $X_l = D_l^{-1}N$
* *Métrique :* $M_l = nD_c^{-1}$
* *Matrice de poids :* $D_{pl} = \frac{D_l}{n}$
* *Centre de gravité :* $g_l = X_l^t D_l 1_p$

2.8.1.2 ACP du nuage des profils-colonnes

* *Tableau des données :* $X_c = D_c^{-1}N^t$
* *Métrique :* $M_c = nD_l^{-1}$
* *Matrice de poids :* $D_{pc} = \frac{D_c}{n}$
* *Centre de gravité :* $g_c = X_c^t D_c 1_q$

Avec la matrice de variance-covariance d'un nuage de profil :

$$V = X^t DX - gg^t$$

2.8.2 ACP non centrées et facteur trivial

Le but de ce paragraphe est de montrer que du point de vue technique, il n'est pas nécessaire de centrer explicitement le nuage de point avant de l'analyser.

Comme g est orthogonal au nuage de points, cela signifie que g est un facteur principal c'est à dire g est un vecteur propre de la matrice VM associé à la valeur propre $\lambda = 0$.

En effet,dans le cas du nuage profils-lignes nous avons :
$VM_l g_l = V1_q$ car $M_l g_l = 1_q$

$$\begin{aligned} M_l g_l &= D_c^{-1} g_l \\ &= \begin{pmatrix} \frac{1}{f_{.1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{f_{.q}} \end{pmatrix} \begin{pmatrix} f_{.1} \\ \vdots \\ f_{.q} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1_q \end{aligned}$$

Donc,

$$\begin{aligned} VM_l g_l &= V1_q \\ &= (X_l^t D_l X_l - g_l g_l^t)1_q \\ &= X_l^t D_l X_l 1_q - g_l g_l^t 1_q \end{aligned}$$

Remarquons que,

$$\begin{aligned} g_l^t 1_q &= (f_{.1}, \cdots, f.q)1_q \\ &= f_{.1} + \cdots + f.q \\ &= 1 \end{aligned}$$

Donc, $g_l g_l^t 1_q = g_l$.

De plus, on a :$X_l 1_q = 1_p$ donc $D_l 1_p = g_c$

Ce qui donne : $X_l^t D_l X_l 1_q = X_l^t g_c$ avec $X_l^t g_c = g_l$

Donc, $X_l^t D_l X_l 1_q = g_l$

Ainsi, on a :

$$\begin{aligned} VM_l g_l &= X_l^t D_l X_l 1_q - g_l g_l^t 1_q \\ &= g_l - g_l \\ &= 0_{\mathbb{R}^q} \\ &= 0 \end{aligned}$$

Ce qui implique que $\lambda = 0$.

$\star$ Les vecteurs propres de la matrice VM sont les mêmes que ceux de la matrice $X^t DXM$ avec les mêmes valeurs propres, sauf pour g qui aura pour valeur propre $\lambda = 1$.

En effet,

$$\begin{aligned} X_l^t D_l X_l M_l &= (V + g_l g_l^t) M_l \\ &= V M_l + g_l g_l^t M_l \\ X_l^t D_l X_l M_l g_l &= V M_l g_l + g_l g_l^t M_l g_l \\ &= 0 + g_l ||g_l||_{M_l}^2 \\ &= g_l \end{aligned}$$

Ce qui implique que $\lambda = 1$.

2.8.3 ACP pour profils-lignes

Facteurs principaux :

Les facteurs principaux sont les vecteurs propres de la matrice $M_l X_l^t D_l X_l$ tels que :

$$\begin{aligned} M_l X_l^t D_l X_l &= (n D_c^{-1})(D_l^{-1} N)^t \frac{D_l}{n} (D_l^{-1} N) \\ &= D_c^{-1} N^t D_l^{-1} N \end{aligned}$$

Composantes Principales :

Les composantes principales sont les vecteurs propres de la matrice $X_l M_l X_l^t D_l$ tels que :

$$X_l M_l X_l^t D_l = D_l^{-1} N D_c^{-1} N^t$$

2.8.4 ACP pour profils-colonnes

Facteurs principaux :

Les facteurs principaux sont les vecteurs propres de la matrice $M_c X_c^t D_c X_c$ tels que :

$$M_c X_c^t D_c X_c = D_l^{-1} N D_c^{-1} N^t$$

Composantes Principales :

Les composantes principales sont les vecteurs propres de la matrice $X_c M_c X_c^t D_c$ tels que :

$$X_c M_c X_c^t D_c = D_c^{-1} N^t D_l^{-1} N$$

2.8.5 Résumé des deux ACP

Les résultats des deux ACP sont résumés comme suit :

	ACP des profils-lignes	ACP des profils-colonnes
Facteurs principaux	$D_c^{-1} N^t D_l^{-1} N$	$D_l^{-1} N D_c^{-1} N^t$
Composantes principales	$D_l^{-1} N D_c^{-1} N^t$	$D_c^{-1} N^t D_l^{-1} N$

Table 2.4 – Résumé des deux ACP.

On constate que les deux ACP conduisent aux mêmes valeurs propres et que les facteurs principaux de l'une sont les composantes principales de l'autre.

2.8.6 Formules de transition

Les facteurs sur les lignes et ceux sur les colonnes sont liés par des relations nommées de transition.

Soit $a = (a_1, a_2, \cdots, a_p)^t$ et $b = (b_1, b_2, \cdots, b_q)^t$ des composantes principales des profils-lignes et profils-colonnes, respectivement. Les deux formules de transition s'écrivent comme suit :

$$\boxed{a = \frac{1}{\sqrt{\lambda}} D_l^{-1} N b} \qquad \boxed{b = \frac{1}{\sqrt{\lambda}} D_c^{-1} N^t a}$$

En effet, nous cherchons une relation entre les vecteurs propres $a = (a_1, \cdots, a_p)^t$ et $b = (b_1, \cdots, b_q)^t$ afin d'éviter la diagonalisation de deux matrices, Une diagonalisation suffira celle de la matrice la plus petite.

On suppose que p < q alors on va diagonaliser la matrice $D_l^{-1} N D_c^{-1} N^t$ qui possède "a" comme vecteur propre alors :

$$D_l^{-1} N D_c^{-1} N^t a = \lambda a$$

On multiplie les deux côtés par $D_c^{-1} N^t$ on obtient un vecteur proportionnel a "b" $b = kb'$ tel que :

$$D_c^{-1} N^t D_l^{-1} N D_c^{-1} N^t a = \lambda D_c^{-1} N^t a$$
$$D_c^{-1} N^t D_l^{-1} N b' = \lambda b'$$

Avec $b' = D_c^{-1} N^t a$ un vecteur propre de la matrice $D_c^{-1} N^t D_l^{-1} N$.

Donc $b = k D_c^{-1} N^t a$, déterminons la valeur de k.

Pour que le vecteur "b" soit M_c -normé, il faut trouver une constante k telle que $||k D_c^{-1} N^t a||_{M_c} = 1$.

D'après la condition de normalisation, on sait que $\lambda = b^t \frac{D_c}{n} b$, donc :

$$\lambda = k a^t N D_c^{-1} \frac{D_c}{n} k D_c^{-1} N^t a$$
$$= \frac{k^2}{n} a^t N D_c^{-1} N^t a$$

D'après l'hypothèse $D_l^{-1} N D_c^{-1} N^t a = \lambda a$ On a :

$N D_c^{-1} N^t a = \lambda D_l$

On obtient $\lambda k^2 a^t \frac{D_l}{n} a = \lambda$ car $\lambda = a^t \frac{D_l}{n} a$.

$\Rightarrow k^2 \lambda = 1 \quad \Rightarrow k = \frac{1}{\sqrt{\lambda}}$

Donc on a bien :

$$a = \frac{1}{\sqrt{\lambda}} D_l^{-1} N b \qquad b = \frac{1}{\sqrt{\lambda}} D_c^{-1} N^t a$$

Où

$$a_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{q} \frac{n_{ij}}{n_{i.}} b_j \qquad b_j = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{p} \frac{n_{ij}}{n_{.j}} a_i$$

2.8.7 Décomposition de l'inertie

l'inertie totale est la somme des valeurs propres [11], elle est égale à ϕ^2 Comme il y a au plus min(p-1,q-1) valeurs propres, si p < q on obtient :

$$\phi^2 = \sum_{k=1}^{p-1} \lambda_k$$

2.8.8 Contribution des profils

Comme en ACP, pour interpréter correctement les graphiques, il faut tenir compte d'une part, de la proximité entre points et plans principaux, et d'autre part, du rôle joué par chaque point dans la détermination d'un axe [1].

Les données étant qualitatives, on n'utilisera pas les cercles de corrélations entre caractères et axes principaux, donc l'interprétation des composantes se fait essentiellement en utilisant les contributions des modalités aux inerties des axes factoriels.

La contribution totale est donnée par la formule suivante :

$$\lambda = \frac{1}{n}\sum_{i=1}^{p} n_{i.}a_i^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}b_j^2$$

On définit la contribution d'une ligne i par :

$$CTR(i) = \frac{n_{i.}}{n}\frac{a_i^2}{\lambda} = \frac{f_{i.}a_i^2}{\lambda} \qquad i = 1 \cdots p$$

Avec a_i la i-ème coordonnée de a, tel que a une composante principale du nuage des profils- lignes.

On définit la contribution d'une colonne j :

$$CTR(j) = \frac{n_{.j}}{n}\frac{b_i^2}{\lambda} = \frac{f_{.j}b_i^2}{\lambda} \qquad j = 1 \cdots q$$

Avec b_i la j-ème coordonnée de b, tel que b une composante principale du nuage des profils-colonnes.

En pratique, on considère les modalités ayant les plus fortes contributions, lorsqu'elle dépassent au minimum son poids :

$$CTR(i) > \frac{n_{i.}}{n} \quad \Rightarrow \quad CTR(i) > f_{i.}$$

Et

$$CTR(j) > \frac{n_{.j}}{n} \quad \Rightarrow \quad CTR(j) > f_{.j}$$

2.9 Exemple illustratif :

Considérons la matrice des données de l'exemple précédent de la base "CSP/filières" [9] :

CSP / Filière	Droit	Sciences	Médecine	IUT	$n_{i.}$
Exp.agri	80	99	65	58	302
Patron	168	137	208	62	575
Cadre.sup	470	400	876	79	1825
Employé	145	133	135	54	467
Ouvrier	166	193	127	129	615
$n_{.j}$	1029	962	1411	382	3784

$$N^* = \begin{pmatrix} 80 & 99 & 65 & 58 \\ 168 & 137 & 208 & 62 \\ 470 & 400 & 876 & 79 \\ 145 & 133 & 135 & 54 \\ 166 & 193 & 127 & 129 \end{pmatrix}$$

L'effectif total vaut $n = 3784$.

$p = 5$ et $q = 4$.

Les marges en lignes :

$n_{1.} = 302$, $n_{2.} = 575$.

$n_{3.} = 1825$,$n_{4.} = 467$.

$n_{5.} = 615$.

Les marges en colonnes :

$n_{.1} = 1029$, $n_{.2} = 962$.

$n_{.3} = 1411$, $n_{.4} = 382$.

Le tableau des fréquences :

CSP / Filière	Droit	Sciences	Médecine	IUT	$f_{i.}$
Exp.agri	0.0211	0.0262	0.0172	0.0153	0.0798
Patron	0.0444	0.0362	0.0550	0.0164	0.1520
Cadre.sup	0.1242	0.1057	0.2315	0.0209	0.4823
Employé	0.0383	0.0351	0.0357	0.0143	0.1234
Ouvrier	0.0439	0.0510	0.0336	0.0341	0.1625
$f_{.j}$	0.2719	0.2542	0.3729	0.1010	1

La matrice des fréquence s'écrit comme suit :

$$N = \begin{pmatrix} \frac{80}{3784} & \frac{99}{3784} & \frac{65}{3784} & \frac{58}{3784} \\ \frac{168}{3784} & \frac{137}{3784} & \frac{208}{3784} & \frac{62}{3784} \\ \frac{470}{3784} & \frac{400}{3784} & \frac{876}{3784} & \frac{79}{3784} \\ \frac{145}{3784} & \frac{133}{3784} & \frac{135}{3784} & \frac{54}{3784} \\ \frac{166}{3784} & \frac{193}{3784} & \frac{127}{3784} & \frac{129}{3784} \end{pmatrix}$$

Les fréquences marginales en lignes :

$$f_{1.} = \sum_{j=1}^{4} f_{1j} = \frac{302}{3784} = 0.0798, \qquad f_{2.} = \sum_{j=1}^{4} f_{2j} = \frac{575}{3784} = 0.1520,$$

$$f_{3.} = \sum_{j=1}^{4} f_{3j} = \frac{1825}{3784} = 0.4823, \qquad f_{4.} = \sum_{j=1}^{4} f_{4j} = \frac{467}{3784} = 0.1234,$$

$$f_{5.} = \sum_{j=1}^{4} f_{5j} = \frac{615}{3784} = 0.1625$$

Les fréquences marginales en colonnes :

$$f_{.1} = \sum_{i=1}^{5} f_{i1} = \frac{1029}{3784} = 0.2719, \qquad f_{.2} = \sum_{i=1}^{5} f_{i2} = \frac{962}{3784} = 0.2542,$$

$$f_{.3} = \sum_{i=1}^{5} f_{i3} = \frac{1411}{3784} = 0.3729, \qquad f_{.4} = \sum_{i=1}^{5} f_{i4} = \frac{382}{3784} = 0.1010$$

• La proportion de personnes qui ont choisi "Médecine" **et** qui sont des enfants de "Cadre.sup" est de 23.15%.

• La proportion des enfants de "Cadre.sup" est de 48.23%.

• La proportion des étudiants qui ont fait le choix de la filière "Médecine" est de 37.29%.

Statistique du χ^2 :

$$\chi^2 = n \sum_{i=1}^{5} \sum_{j=1}^{4} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \approx 320.27$$

Pour un risque $\alpha = 0.05$ on a $\chi^2_{12} = 13.84$, donc $\chi^2 > \chi^2_{12}$ ce qui montre qu'on a bien une dépendance entre les valeurs qualitatives.

Centres de gravité des profils-lignes :

$$g_l = (0.2719, 0.2542, 0.3729, 0.1010)^t$$

Centres de gravité des profils-colonnes :

$$g_c = (0.0798, 0.1520, 0.4823, 0.1234, 0.1625)^t$$

Matrices diagonales de profils-lignes :

$$D_l = \begin{pmatrix} 0.0798 & 0 & 0 & 0 & 0 \\ 0 & 0.1520 & 0 & 0 & 0 \\ 0 & 0 & 0.4823 & 0 & 0 \\ 0 & 0 & 0 & 0.1234 & 0 \\ 0 & 0 & 0 & 0 & 0.1625 \end{pmatrix}$$

Matrices diagonales de profils-colonnes :

$$D_c = \begin{pmatrix} 0.2719 & 0 & 0 & 0 \\ 0 & 0.2542 & 0 & 0 \\ 0 & 0 & 0.3729 & 0 \\ 0 & 0 & 0 & 0.1010 \end{pmatrix}$$

Tableau profils-lignes :

CSP / Filière	Droit	Sciences	Médecine	IUT	Total
Exp.agri	0.2649	0.3278	0.2152	0.1921	1
Patron	0.2922	0.2383	0.3617	0.1078	1
Cadre.sup	0.2575	0.2192	0.4800	0.0433	1
Employé	0.3105	0.2848	0.2891	0.1156	1
Ouvrier	0.2699	0.3138	0.2065	0.2098	1
Total	0.2719	0.2542	0.3729	0.1010	1

• La proportion de personnes qui ont choisi la filières "Sciences" est de 25.42%.

• La proportion de personnes qui ont choisi filière "Sciences" parmi les enfants "d'ouvrier" est de 31.38%.

Tableau profils-colonnes :

CSP / Filière	Droit	Sciences	Médecine	IUT	Total
Exp.agri	0.0777	0.1029	0.0461	0.1518	0.0798
Patron	0.1633	0.1424	0.1474	0.1623	0.1520
Cadre.sup	0.4568	0.4158	0.6208	0.2068	0.4823
Employé	0.1409	0.1383	0.0957	0.1414	0.1234
Ouvrier	0.1613	0.2006	0.0900	0.3377	0.1625
Total	1	1	1	1	1

Nous remarquons que les enfants de "Cadre.sup" ont une préférence pour la filière de "Médecine" avec une proportion de 62.08%, contrairement aux enfants "d'ouvriers" qui choisissent plutôt "IUT" avec une proportion de 33.77%.

Matrices profils-lignes :

$$X_l = D_l^{-1}N = \begin{pmatrix} \frac{40}{151} & \frac{99}{302} & \frac{65}{302} & \frac{29}{151} \\ \frac{168}{575} & \frac{137}{575} & \frac{208}{575} & \frac{62}{575} \\ \frac{94}{365} & \frac{16}{73} & \frac{12}{25} & \frac{79}{1825} \\ \frac{145}{467} & \frac{133}{467} & \frac{135}{467} & \frac{54}{467} \\ \frac{166}{615} & \frac{193}{615} & \frac{127}{615} & \frac{43}{205} \end{pmatrix}$$

$$= \begin{pmatrix} 0.2649 & 0.3278 & 0.2152 & 0.1921 \\ 0.2922 & 0.2383 & 0.3617 & 0.1078 \\ 0.2575 & 0.2192 & 0.4800 & 0.0433 \\ 0.3105 & 0.2848 & 0.2891 & 0.1156 \\ 0.2699 & 0.3138 & 0.2065 & 0.2098 \end{pmatrix}$$

Matrice profils-colonnes :

$$X_c = D_c^{-1}N^t = \begin{pmatrix} \frac{80}{1029} & \frac{8}{49} & \frac{470}{1029} & \frac{145}{1029} & \frac{166}{1029} \\ \frac{99}{962} & \frac{137}{962} & \frac{200}{481} & \frac{133}{962} & \frac{193}{962} \\ \frac{65}{1411} & \frac{208}{1411} & \frac{876}{1411} & \frac{135}{1411} & \frac{127}{1411} \\ \frac{29}{191} & \frac{31}{191} & \frac{79}{382} & \frac{27}{191} & \frac{129}{382} \end{pmatrix}$$

$$= \begin{pmatrix} 0.0777 & 0.1633 & 0.4568 & 0.1409 & 0.1613 \\ 0.1029 & 0.1424 & 0.4158 & 0.1383 & 0.2006 \\ 0.0461 & 0.1474 & 0.6208 & 0.0957 & 0.0900 \\ 0.1518 & 0.1623 & 0.2068 & 0.1414 & 0.3377 \end{pmatrix}$$

Distance du χ^2 :

$$\begin{aligned} d^2_{\chi^2}(Cadre.sup, Ouvrier) &= d^2_{\chi^2}(3,5) \\ &= \frac{1}{0.2719}(0.258-0.270)^2 + \frac{1}{0.2542}(0.219-0.314)^2 \\ &+ \frac{1}{0.3729}(0.480-0.207)^2 + \frac{1}{1010}(0.043-0.210)^2 \\ &= 0.5109 \end{aligned}$$

$d^2_{\chi^2}$ (Cadre.sup,Patron) $=d^2_{\chi^2}$ (3,2) = 0.0846.

On observe que le choix des filières des enfants de "Cadre.sup" est plus proche de celle des "Patrons" qu'elle ne l'est des enfants "d'ouvriers".

L'inertie totale des nuages de points :

$$\begin{aligned} Inertie(X_l/g_l) &= Inertie(X_c/g_c) \\ &= \frac{\chi^2}{n} \\ &= \frac{320.27}{3784} \\ &= 0.0846 \end{aligned}$$

Maintenant appliquons l'ACP a notre exemple, utilisons la plus petite matrice $A_l = X_l^t X_c^t$ tel que :

$$A_l = \begin{pmatrix} 0.2732 & 0.2730 & 0.2691 & 0.2759 \\ 0.2552 & 0.2611 & 0.2417 & 0.2800 \\ 0.3690 & 0.3546 & 0.4074 & 0.3012 \\ 0.1024 & 0.1111 & 0.0815 & 0.1427 \end{pmatrix}$$

Valeurs propres de la matrice A_l :

$\lambda_1 = 0.0823.$

$\lambda_2 = 0.0017.$

$\lambda_3 = 0.0005.$

$\lambda_4 = 1.$

On observe que le centre de gravité g_l est un vecteur propre de A_l associé à la valeur propre $\lambda_4 = 1$, c'est à dire $A_l g_l = g_l$.

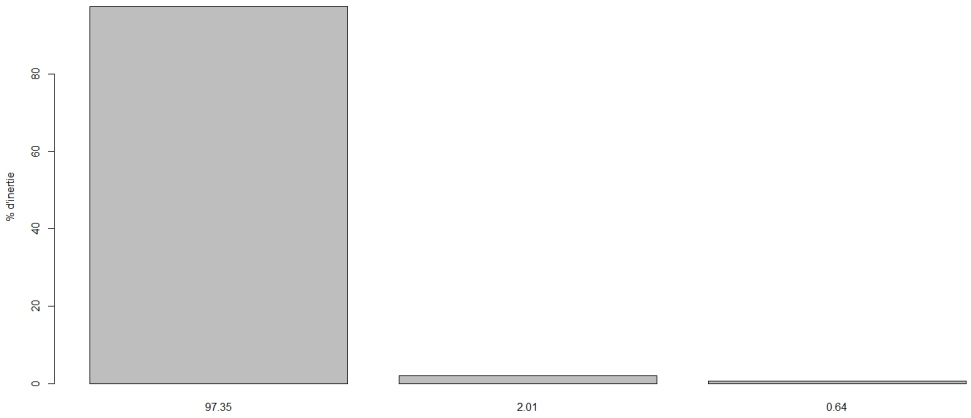


Figure 2.1 – Histogramme des valeurs propres

Donc l'inertie totale :

$$\begin{aligned}\phi^2 &= \sum_{k=1}^{3} \lambda_k \\ &= 0.0823 + 0.0017 + 0.0005 \\ &= 0.0846\end{aligned}$$

La variance du 1er facteur est égale à $\lambda_1 = 0.0823$ elle représente 97.35% de l'inertie totale.

Les vecteurs propres associés aux valeurs propres, respectivement, sont :

$$u_1 = \begin{pmatrix} 0.0556 \\ 0.2989 \\ -0.8272 \\ 0.4725 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 0.7982 \\ 0.0191 \\ -0.5276 \\ -0.2899 \end{pmatrix}$$

$$u_3 = \begin{pmatrix} 0.4097 \\ -0.8532 \\ 0.1682 \\ 0.2752 \end{pmatrix} \qquad u_4 = \begin{pmatrix} 0.5069 \\ 0.4738 \\ 0.6950 \\ 0.1880 \end{pmatrix}$$

Méthode de calcul du vecteur propre de la valeur propre $\lambda_4 = 1$:

$$A_l - \lambda_4 \,\mathrm{Id} = \begin{pmatrix} -0.7268 & 0.2730 & 0.2691 & 0.2759 \\ 0.2552 & -0.7389 & 0.2417 & 0.2800 \\ 0.3690 & 0.3546 & -0.5926 & 0.3012 \\ 0.1024 & 0.1111 & 0.0815 & -0.8573 \end{pmatrix}$$

$Au = \lambda u \Rightarrow (A - \lambda \,\mathrm{Id})u = 0$
C'est un système d'équations linéaires, nous pouvons résoudre le système par la méthode d'élimination de Gauss.
Ce qui nous donne le vecteur propre associé à la valeur propre λ_4, on utilise la même méthode pour les autres vecteurs propres.
On va normaliser les vecteurs par la métrique $M_l = D_c^{-1}$,
avec $u_i^* = \frac{u_i}{\sqrt{||u_i||^2_{M_l}}}$, i $= 1 \cdots 4$.

$$u_1^* = \frac{u_1}{\sqrt{u_1^t M_l u_1}} = \begin{pmatrix} 0.0265 \\ 0.1424 \\ -0.3940 \\ 0.2250 \end{pmatrix}$$

On utilise la même méthode pour trouver u_2^*, u_3^*, et u_4^*.

Les composantes principales :

$$C_1 = X_l M_l u_1^* = \begin{pmatrix} 0.4099 \\ 0.0199 \\ -0.2628 \\ 0.4513 \end{pmatrix}$$

Contribution des profils :

Calculons la contribution des cadres supérieurs :
$CTR(3) = \frac{f_3 . c_1^2(3)}{\lambda_1} = 40.43\%$
Calculons la contribution des exploitants agricoles :
$CTR(1) = \frac{f_1 . c_1^2(1)}{\lambda_1} = 16.29\%$

Représentation graphique :

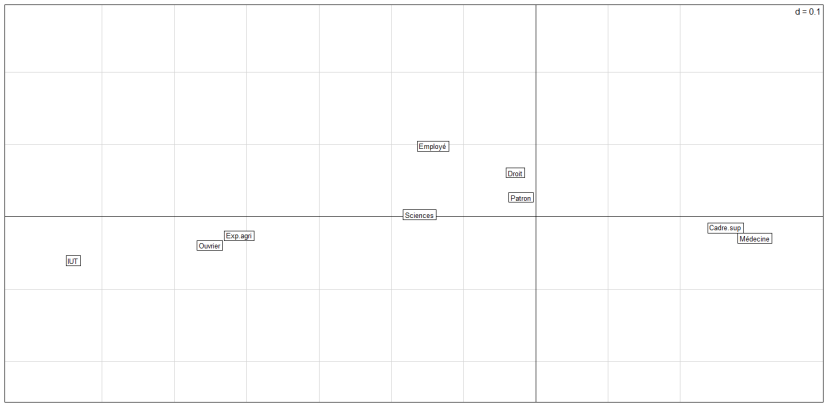


Figure 2.2 – Représentation Données "CSP/Filières"

Interprétation :

L'axe 1 reflète une hiérarchisation des classes sociales de "l'ouvrier" vers le "Cadre.sup", pour les filières de "l'IUT" vers les études en "Médecine" sachant que l'inertie apportée par cet axe est de 97.35%.
L'axe 2 est plus représenté par la classe des enfants de "Patron" et les études en "Droit" sachant que l'inertie apportée par cet axe est de 2.01%.

Axe 1 :

1ère variables :

(+)	(-)	Poids	Contribution	Qualité
	Exp.agri	7.98%	16.29%	98.73%
	Ouvrier	16.25%	40.20%	99.19%

2ème variables :

(+)	(-)	Poids	Contribution	Qualité
Médecine		37.29%	41.58%	99.02%
	IUT	10.10%	50.21%	98.87%

Axe 2 :

2ème variables :

(+)	(-)	Poids	Contribution	Qualité
Droit		27.19%	58.76%	77.69%

Conjonction (produit scalaire positif) :

Les enfants d'ouvriers et d'exploitants agricoles sont plus attirés par les études en institut universitaire technique "IUT" avec respectivement, une proportion de 33.77% et 15.18%.
Quant aux enfants de "Patrons", on observe graphiquement qu'ils sont intéressés par la filière de "Droit" avec une proportion de 16.33%, sauf qu'en réalité, ils sont mal représentés (qualité égale à 21.35%), au final, on conclut que cette affirmation n'est vraie que pour cet échantillon, vu la qualité de représentation.
De la même manière les enfants de cadres supérieurs ont une affinité avec les études en médecine (leur proportion est de 62.08%).
On observe donc, une conjoncture entre l'origine sociale des enfants et leurs choix de filières à l'université.

Opposition (produit scalaire négatif) :

De prime abord, on peut constater que suivant les filières d'études (médecine, institut universitaire technique) l'écart est visible et palpable dans leurs choix.

Quadrature (produit scalaire nul) :

On remarque graphiquement deux indépendances (angle droit) entre les enfants des patrons et ceux des cadres supérieurs et les enfants des patrons avec les enfants d'exploitants agricoles et des ouvriers.
En d'autres termes on ne peut rien conclure concernant le choix d'études des enfants de patrons (médecine ou IUT).

Chapitre 3

Applications sur R

Ce chapitre représente la partie pratique de notre travail nous allons appliquer la méthode de l'AFC sur des données réelles avec le logiciel R [12].
Dans notre exemple nous allons utilisés la version 4.0.5 du **logiciel R**, téléchargée depuis le site : https ://cran.r-project.org/bin/windows/base/old/4.0.5/.

3.1 Logiciel R

R est un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données. Le logiciel R est un environnement de manipulation de données [6], de calcul, de représentation graphiques et aide à la réalisation des analyses statistiques telles que : les méthodes de régression linéaire, les tests d'hypothèses, l'analyse de la variance $\cdots$ etc.

3.2 Les packages

On trouve plusieurs packages qui sont disponibles dans le logiciel R qui permettent la réalisation d'une AFC, tels que :
◦ Le package **FactoMineR** (Factor analysis and Data Mining with R).
◦ Le package **ade4** (Analysis of Environmental Data : Exploratory and Euclidean method).

3.3 Les fonctions liées a l'AFC

◦ La fonction **CA()** du package FactoMineR.
◦ La fonction **dudi.coa()** du package ade4.

3.4 Exemple d'application

On s'intéresse à la relation entre différents parfums et leurs fragrances selon 1000 sujets féminins. Les données sont résumées dans le tableau suivant :

Noms/Fragrances	Fruité	Doux	Fort	boisé
Giorgio Armani "Si"	45	127	23	5
Hermès "Terre"	87	95	19	1
Paco Rabanne "Black Xs"	0	0	52	146
Givenchy "L'interdit"	34	63	74	22
Chanel "N°5"	0	30	118	59

Table 3.1 – Parfums/Fragrances-Tableau des effectifs observés

Il s'agit d'une étude de liaison entre deux variables qualitatives :
U = parfums et V = fragrances.
Notre tableau de contingence contient deux variables qualitatives, les noms de parfums comme première variable de 5 modalités $p = 5$, et les fragrances de parfums comme deuxième variable de 4 modalités $q = 4$, de taille $n = 1000$.

Les données :

```
>♯ Entrée des données de l'exemple
> Effectif <- matrix(c(45,127,23,5,87,95,19,1,0,0,52,146, 34,63,74,22,0,30,118,
59),ncol=4,byrow=TRUE)
> colnames(Effectif) <- c("Fruité","Doux","Fort","Boisé")
> rownames(Effectif) <- c("Armani","Hermés","Black Xs","Givenchy","Chanel")
> Effectif
```

	Fruité	Doux	Fort	Boisé
Armani	45	127	23	5
Hermés	87	95	19	1
Black Xs	0	0	52	146
Givenchy	34	63	74	22
Chanel	0	30	118	59

```
> as.data.frame(as.table(Effectif))
> n <-sum(Effectif)
```

```
> n
[1] 1000
> l <- nrow(Effectif)
> l
[1] 5
> c <- ncol(Effectif)
> c
[1] 4
> ni. <-rowSums(Effectif)
> ni.
 Armani  Hermés  Black Xs  Givenchy  Chanel
    200     202       198       193     207
> n.j <-colSums(Effectif)
> n.j
 Fruité  Doux  Fort  Boisé
    166   315   286    233
```

⋆ Pour calculer l'effectif marginal, on peut aussi utiliser les commandes suivantes :

```
> margin.table(Effectif,1)
> margin.table(Effectif,2)
```

Tableau des fréquences :

```
> freqEffectif <- Effectif/n
> freqEffectif
           Fruité  Doux  Fort  Boisé
Armani      0.045 0.127 0.023 0.005
Hermés      0.087 0.095 0.019 0.001
Black Xs    0.000 0.000 0.052 0.146
Givenchy    0.034 0.063 0.074 0.022
Chanel      0.000 0.030 0.118 0.059
> fi.=ni./n
> fi.
 Armani  Hermés  Black Xs  Givenchy  Chanel
  0.200   0.202     0.198     0.193   0.207
> f.j=n.j/n
```

> f.j

Fruité	Doux	Fort	Boisé
0.166	0.315	0.286	0.233

⋆ La commande suivante peut aussi être utilisée :
> prop.table(Effectif)

Test du χ^2 :
La première étape consiste à étudier la liaison entre les deux variables : parfums et fragrances à l'aide du test du χ^2.

Pour cela on utilise la commande **chisq.test** pour effectuer le test.
> chisq.test(Effectif)
Pearson's Chi-squared test
data : Effectif
X-squared = 742.93, df = 12, p-value < 2.2e-16

Interprétation :
La p-value $< \alpha = 0.05$, ce qui indique qu'il y a une liaison entre les deux variables qualitatives (dépendantes).

Tableau des profils :
La deuxième étape consiste à analyser les profils-lignes et les profils-colonnes.

> ♯ *Profils-lignes*
> round(Effectif/ apply(Effectif,1,sum),3)

	Fruité	Doux	Fort	Boisé
Armani	0.225	0.635	0.115	0.025
Hermés	0.431	0.470	0.094	0.005
Black Xs	0.000	0.000	0.263	0.737
Givenchy	0.176	0.326	0.383	0.114
Chanel	0.000	0.145	0.570	0.285

>♯ *Profils-colonnes*
> round(t(t(Effectif)/apply(t(Effectif),1,sum)),3)

	Fruité	Doux	Fort	Boisé
Armani	0.271	0.403	0.080	0.021
Hermés	0.524	0.302	0.066	0.004
Black Xs	0.000	0.000	0.182	0.627
Givenchy	0.205	0.200	0.259	0.094
Chanel	0.000	0.095	0.413	0.253

```
> ♯ Poids des lignes et colonnes
> round(apply(Effectif,1,sum)/sum(Effectif),3)
```

Armani	Hermés	Black Xs	Givenchy	Chanel
0.200	0.202	0.198	0.193	0.207

```
> round(apply(Effectif,2,sum)/sum(Effectif),3)
```

Fruité	Doux	Fort	Boisé
0.166	0.315	0.286	0.233

Analyse Factorielle des Composantes :

La troisième étape consiste à calculer l'AFC, Pour cela on utilisera la fonction **dudi.coa**.

```
> ♯ Appliquer l'AFC
> afc <- dudi.coa(df=Effectif,scannf=F,nf=3)
```

Les valeurs propres :

Les valeurs propres sont utilisées pour déterminer le nombre d'axes.

```
> round(afc$eig,3)
[1] 0.594 0.118 0.032
```

Inertie :

La quatrième étape consiste à trouver l'inertie commune aux deux nuages de profils $\phi^2 = 0.744$ et qui s'accorde avec la somme des valeurs propres.

```
> sum(round(afc$eig,3))
[1] 0.744
> inertie<-round(afc$eig/sum(afc$eig)*100,3)
> inertie
 [1]  79.931  15.816  4.253
> cumsum(inertie)
 [1]  79.931  95.747  100.000
```

• L'inertie apportée par les deux axes est de 95.747%.

> barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,3))
> title("Valeurs propres")

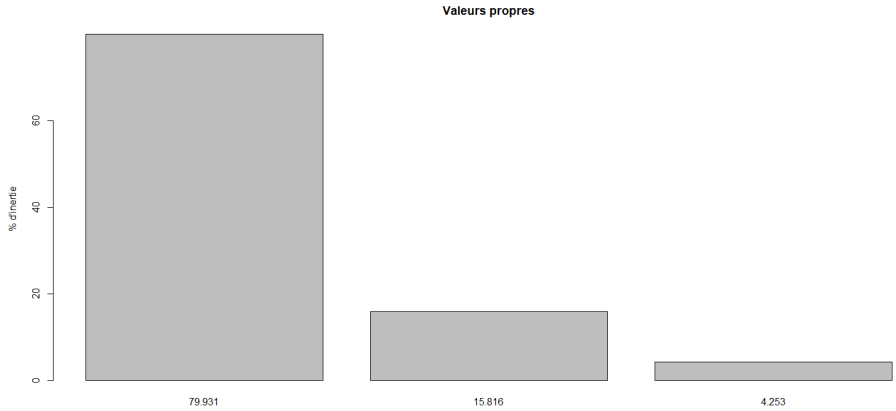


Figure 3.1 – Pourcentage des valeurs propres

• Le premier axe représente environ 80% de l'inertie totale λ_1.
• Le deuxième axe représente environ 16% de l'inertie totale λ_2.
• Le troisième axe représente environ 4% de l'inertie totale λ_3.
On prendra seulement les deux premiers axes principaux.

La dernière étape consiste à élaborer la représentation graphique.
> ♯ *Interprétation des facteurs*

Les composantes principales des profils-lignes :
> ♯ *Les coordonnées*
> round(afc$li,3)

	Axis1	Axis2	Axis3
Armani	-0.730	0.079	-0.309
Hermés	-0.856	0.247	0.239
Black Xs	1.173	0.446	-0.018
Givenchy	-0.143	-0.265	0.066
Chanel	0.552	-0.497	0.020

> ♯ *Les contributions des profils-lignes*
> ctrl=round(inertia.dudi(afc,row.inertia=TRUE)$ row.abs,3)
> ctrl

	Axis1	Axis2	Axis3
Armani	17.937	1.050	60.288
Hermés	24.917	10.484	36.553
Black Xs	45.863	33.515	0.205
Givenchy	0.666	11.519	2.681
Chanel	10.618	43.432	0.272

> ♯ *Qualité des lignes*
> round(inertia.dudi(afc, row.inertia = T)$row.rel,3)

	Axis1	Axis2	Axis3
Armani	-84.005	0.973	-15.023
Hermés	-86.110	7.169	6.721
Black Xs	87.349	12.630	-0.021
Givenchy	-21.563	-73.817	4.620
Chanel	55.226	-44.699	0.075

Les composantes principales des profils-colonnes :

> ♯ *Les coordonnées*
> round(afc$co,3)

	Comp1	Comp2	Comp3
Fruité	-0.877	0.281	0.311
Doux	-0.686	0.017	-0.209
Fort	0.374	-0.495	0.075
Boisé	1.092	0.384	-0.031

> ♯ *Les contributions des profils-colonnes*
> ctrc<-round(inertia.dudi(afc,col.inertia=TRUE)$col.abs,3)
> ctrc

	Axis1	Axis2	Axis3
Fruité	21.492	11.189	50.718
Doux	24.941	0.079	43.480
Fort	6.744	59.566	5.090
Boisé	46.822	29.166	0.712

```
> ♯ Qualité des colonnes
> round(inertia.dudi(afc, col.inertia = T)$col.rel,3)
```

	Axis1	Axis2	Axis3
Fruité	-81.396	8.385	10.219
Doux	-91.460	0.057	-8.483
Fort	35.871	-62.689	1.440
Boisé	88.963	10.965	-0.072

```
> ♯ Représentation du plan factoriel
> plot(afc$li[,1],afc$li[,2],type="n",xlab="Axe 1",ylab="Axe 2",
xlim=c(-1.2,1.6))
> text(afc$li[,1], afc$li[,2], label=row.names(Effectif))
> text(afc$co[,1], afc$co[,2], label= colnames(Effectif),col="red")
> title("Fragrances de Parfums")
> abline(h=0,v=0)
```

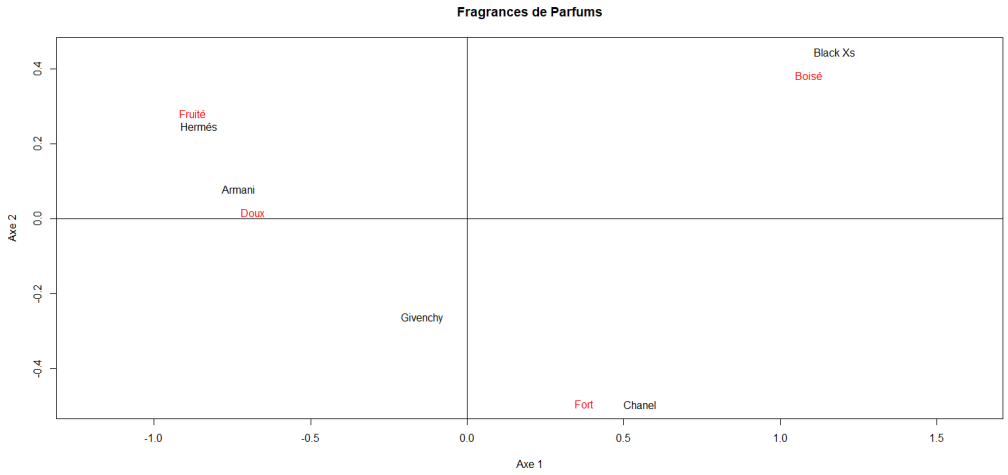


Figure 3.2 – Représentation simultanée des parfums et leurs fragrances

Le graphique 3.2 représente la projection des modalités sur le premier plan factoriel,la distance entre les points lignes ou entre les points colonnes, donne une mesure de leur similitude. Les points lignes avec un profil similaire sont proches sur le graphique. Il en va de même pour les points colonnes.

Interprétation :
L'axe 1 reflète une hiérarchisation des parfums "d'Hermés" vers "Black Xs", pour les fragrances de "Fruité" vers la senteur "Boisé".
L'axe 2 est plus représenté par le parfum "Chanel", et la fragrance "Fort".

Axe 1 :

1ère variables :

(+)	(-)	Poids	Contribution	Qualité
Black Xs		19.80%	45.86%	87.35%
	Hermés	20.20%	24.91%	86.11%

2éme variables :

(+)	(-)	Poids	Contribution	Qualité
Boisé		23.30%	46.82%	88.96%
	Fruité	16.60%	21.49%	81.39%

Axe 2 :

1ére variables :

(+)	(-)	Poids	Contribution	Qualité
	Chanel	20.70%	43.43%	44.70%

2éme variables :

(+)	(-)	Poids	Contribution	Qualité
	Fort	28.60%	59.56%	62.69%

Conjonction :

On remarque qu'un grand nombre de femmes trouve que le parfum "Black Xs" est associé à la senteur "Boisé".
Idem le parfum "Hermés" sera associé à la fragrance "Fruité".
Quant au parfum "Chanel", on a pu lui associer la fragrance "Fort".
Par contre, graphiquement, on remarque que dans cet échantillon, le parfum "Armani" est associé à la senteur "Doux" et le parfum "Gyvenchi" est associé à la senteur "Fort", cette assertion reste vraie seulement pour le panel, pas pour la population.

Opposition :

A priori les femmes qui achètent le parfum "Hermés" ne choisissent pas le parfum "Black Xs".
On remarque aussi, que les femmes qui aiment la senteur boisée vont détester les parfums fruités.

Quadrature :

Graphiquement on remarque une indépendance entre le parfum "black Xs" et "Chanel", "Gyvenchi" et "Hermés" donc les clientes qui achètent l'un peuvent aussi bien être attiré par l'autre.

Conclusion

L'analyse factorielle des correspondances est une méthode utilisée dans différents domaines tels que la biologie, la sociologie, la psychologie, l'économie, l'ingéniorat, $\cdots$ etc, elle a pour but d'étudier la liaison entre deux variables qualitatives ainsi que la réduction de dimension, à condition bien entendu, de conserver le maximum d'information.
Le seul inconvénient de l'analyse factorielle des correspondances est, qu'il nous ait impossible de travailler avec plus de deux variables (tableau de contingence), cependant d'autres méthodes existent pour traiter ce genre de problème tel que l'analyse factorielle multiple (AFM).
Pour conclure l'analyse factorielle des correspondances, nous permet d'obtenir des graphes simples, afin de mieux observer le phénomène étudié et d'avoir une meilleure interprétation, avec pour objectif, de classer un nombre d'individus en fonction de variables qui les décrivent.

Annexe A

Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :
ACP : Analyse en composantes principales.
AFC : Analyse factorielle des correspondances.
AFD : Analyse factorielle discriminante.
X : Tableau des données.
$\mathbb{R}$: Ensemble des nombres réels.
e_i : i-ème individu.
X_j : j-ème variable.
D : Matrice de poids.
g : Centre de gravité.
Y : Matrice centrée.
V : Matrice de variance-covariance.
$D_{\frac{1}{s}}$: Matrice diagonale des inverses.
Z : Tableau standard.
R : Matrice de corrélation.
d(A,B) : Distance entre deux points.
$d^2(e_i, e_j)$: Distance entre deux individus.
M : Métrique.
I_g : Inertie.
tr : Trace.
Var : Variance.
Cov : Covariance.
F_k : Sous-espace de projection de dimension k.
a_k : Axe principal.
u_k : Facteur principal.
C_k : Composante principale.
n : Effectif total.

n_{ij} : Effectif observé.
$n_{i.}$: Marges en lignes.
$n_{.j}$: Marges en colonnes.
f_{ij} : Fréquence observée.
$f_{i.}$: Fréquence marginale en lignes.
$f_{.j}$: Fréquence marginale en colonnes.
$f_{\frac{i}{j}}$: Fréquence conditionnelle aux profils-lignes.
$f_{\frac{j}{i}}$: Fréquence conditionnelle aux profils-colonnes.
N^* : Matrice des effectifs.
U, V : Variables qualitatives.
p, q : Nombres de modalités.
x_i et y_j : Modalités.
N : Matrice des fréquences.
χ^2 : La statistique du Chi-deux.
X_l : Tableau des profils-lignes.
X_c : Tableau des profils-colonnes.
g_l : Centre de gravité profils-lignes.
g_c : Centre de gravité profils-colonnes.
$d^2_{\chi^2}$: Distance du Chi-deux.
ϕ^2 : L'écart à l'indépendance.
CTR : Contribution.
AFM : Analyse factorielle multiple.

Bibliographie

[1] BAEY C. (2019-2020). Analyse de données. M2 Ingénierie Statistique et Numérique. Université de Lille (cour).

[2] BENZECRI J.-P. ET COLL. (1973). L'analyse des données. Tome 2 : L'analyse des correspondances. Dunod.

[3] BERTIER P.,BOUROCHE J.-M. (1975). Analyse des données multidimensionnelles. PUF, Paris.

[4] CAILLEZ F.,PAGES J.-P (1976). Introduction à l'analyse des données. Smash.

[5] ESCOFIER B.,PAGES J. (2008). Analyses factorielles simples et multiples.Dunod.

[6] HUSSON F., LÊ S., PAGES J. (2009). Analyse de données avec R. Presses universitaires de Rennes.

[7] LASGOUTTES J. M. (2022-2023). Cours d'analyse de données.

[8] LEBART L.,MORINEAU A.,FENELON J.-P. (1979). Traitement des données statistiques. Dunod.

[9] RAKOTOMALALA R. Pratique des Méthodes Factorielles avec Python, Université Lumière Lyon2. P.219.

[10] SAPORTA GILBERT. (2006). Probabilités et analyse des données statistique, 3ème édition, Edition Technip.

[11] THIARE, O. (16 avril 2020). Analyse factorielle des correspondance (AFC).

[12] TILLÈ Y. (2010). Résumé du cours de statistique descriptive.