République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Statistique**

Par

MASRI Maroua

Titre :

Analyse Factorielle des Correspondances : Etude de cas en utilisant le langage R

Membres du Comité d'Examen :

Pr.	**BENATIA Fateh**	UMKB	Président
Pr.	**NECIR Abdelhakim**	UMKB	Encadreur
Dr.	**BENAMEUR Sana**	UMKB	Examinateur

juin 2021

Dédicace

À mes chers parents, que Dieu les protège, pour leurs encouragements et leurs prières tout au long de mes études.

À mes frères et sœurs qui m'ont toujours encouragé durant ces années d'études.

À toute ma famille, je dédie cet humble travail.

REMERCIEMENTS

"Allah aime ceux qui s'en remettent à lui"

Grace à dieu et à son aide, et après l'effort et la persévérance, cet humble travail a été realisé. Je tiens à remercier Allah le tout puissant de m'avoir donné la santé et le courage pour accomplir ce travail.

Je tiens à exprimer toute ma reconnaissance à mon encadreur monsieur le Professeur **NECIR Abdelhakim,** je le remercie de m'avoir encadré, conseillé et aidé.

J'aimerais présenter mes remerciements aux membres du jury, monsieur le Professeur **BENATIA Fateh** et le Dr Madame **BENAMEUR Sana** pour le grand honneur qu'ils nous font en acceptant de juger ce travail.

Aussi, je souhaite adresser mes sincères remerciements à tous les enseignants de la Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie -Département de Mathématiques-.

Je remercie mes trés chers parents pour leurs encouragements et leur soutien.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'analyse des données est un terme regroupant plusieurs méthodes permettant d'extraire l'information contenue dans un jeu de données [2], lorsque ces jeux de données ont grand dimension, il faudra de réduire la dimension, en conservant au mieux l'information utile, pour cela nous nous intéresserons à la méthode d'analyse factorielle.

L'analyse factorielle des correspondances ou dite analyse factorielle simple (Correspondence Analysis en Anglais) a été proposé en France par **J.-P.** Benzécri à l'Université Pierre-et-Marie-Currie à Paris (ISUP et le Laboratoire de Statistique Théorique et Appliquée), elle a été mise au point durant la période 1970-1990 [11].

L'AFC est une technique statistique d'analyse des données, cette méthode consiste à étudier la liaison (dite encore correspondance) entre deux variables qualitatives (catégorielles), elle permet d'analyser les informations contenues dans un tableau de contingence. Le but de cette méthode est la réduction de la dimension. L'AFC est une extension de l'analyse en composantes principales (ACP), basée sur la distance du khi-deux.

Ce sujet est organisé en deux parties : théorique et pratique, la partie théorique se compose en deux chapitres, le premier chapitre rappelle les notions générales de cette méthode, le deuxième chapitre est consacré à la principe d'AFC et comment calculer l'AFC en utilisant l'analyse en composantes principales.

On finalise ce travail avec une application en utilisant le langage R, on charge un jeu

de données (Jeux Olympiques) de R, et on essayera d'appliquer l'AFC à ces données en utilisant les packages de R [**FactoMiner**] pour l'analyse et [**factoextra**] pour l'interprétation des résultas.

Chapitre 1

Notions générales

L'AFC étant une ACP particulière qui utilise une métrique spéciale. Dans ce chapitre, on souhaite donner des notations principales qu'on les utilise dans cette méthode [1].

1.1 Les données

1.1.1 Tableau de contingence

Soient V_1 et V_2 deux variables qualitatives ou bien catégorièlles à p et q catégories (modalités), respectivement, décrivant un ensemble de n individus [13]. L'Analyse Factorielle est basée sur le nuage de points, qu'on l'appelle tableau de contingence, on le note par N^*. C'est la matrice des effectifs obsérvés de p lignes et q colonnes. On croisant les deux variables V_1 et V_2 on obtient :

$$N^* := \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pq} \end{pmatrix} \in \mathcal{M}(p \times q),$$

où x_{ij} c'est l'effectif observé, c'est l'élément obtenu par l'intersection de la ligne i et la colonne j.

1.2 Exemple illustratif

Pour comprendre bien ce qu'est un tableau de contingence, on peut l'expliquer par utilisation une base de donnée «CSP Fillières», où l'on a croisé l'origine sociale des étudiants (à travers la CSP-catégorie social professionelles) 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 [12].

CSP\Fillière	Droit	Science	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

TAB. 1.1 – CSP et Choix de filières." Tableau des effectifs observés " [12].

Cette base de donnée contient deux variables qualitatives. "CSP" comme la première variable de 5 modalités V_1 :=(Exp.agri, Patron, Cadre.sup, Employé, Ouvrier), et "Fillère" comme la deuxième variable de 4 modalités V_2 :=(Droit, Science, Médecine, IUT), de taille ($n = 3784$) individus.

1.3 Quelques notations standards

1.3.1 Effectif total

Définition 1.3.1 *[11]L'effectif total est noté par n, c'est la somme de tout les effectifs observés*

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

1.3.2 Effectifs marginals

Définition 1.3.2 *[11]On peut définir les effectifs marginals des lignes par $X_{i.}$ avec*

$$X_{i.} = \sum_{j=1}^{q} x_{ij}, \ \ i = 1, ..., p,$$

et les effectifs marginals des colonnes $X_{.j}$ sont donnés par :

$$X_{.j} = \sum_{i=1}^{p} x_{ij}, \ \ j = 1, ..., q.$$

1.3.3 Distributions marginales

Définition 1.3.3 *[5]Le pourcentage de l'éffectif total est représenté par une case (i, j), c'est la fréquence observée, on la note par f_{ij}*

$$f_{ij} = \frac{1}{n} x_{ij},$$

la matrice des fréquences observées N est représentée comme suit :

$$N = \frac{1}{n}N^* = \begin{pmatrix} f_{11} & \cdots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pq} \end{pmatrix} \in \mathcal{M}(p \times q).$$

Définition 1.3.4 *[5]La somme des fréquences d'une même ligne i, représente le pourcentage global de cette ligne, c'est la fréquence marginale de la modalité i*

$$f_{i\cdot} = \sum_{j=1}^{q} f_{ij} = P(V_1 = i), \ \textit{pour } i = 1, ..., p.$$

On calcule de même la fréquence marginale de la modalité j

$$f_{\cdot j} = \sum_{i=1}^{p} f_{ij} = P(V_2 = j), \ \textit{pour } j = 1, ..., q.$$

Remarque 1.3.1 *[7]La somme des distributions marginales toujours égale à l'unité*

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

1.3.4 Distributions conditionnelles

Définition 1.3.5 *On définit les fréquences conditionnelles aux profils-lignes $f_{i/j}$ (lire " fréquence de i sachant j "), par*

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

De même, les fréquences conditionnelles aux profils-colonnes $f_{j/i}$ (lire " fréquence de j sachant i ")

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

On a aussi

$$\sum_{j=1}^{q} f_{j/i} = 1, \ \textit{pour } i = 1, ..., p,$$

et

$$\sum_{i=1}^{p} f_{i/j} = 1, \ \textit{pour } j = 1, ..., q.$$

Définition 1.3.6 *[11]On définit la fréquence théorique $\widetilde{f}_{ij}$ par :*

$$\widetilde{f}_{ij} = f_{i\cdot} f_{\cdot j}.$$

1.4 Liaison entre les variables

L'Analyse Factorielle a pour but d'étudier la liaison entre les deux variables (V_1 et V_2), dite encore correspondance. Lorsque on étudie un tableau de contingence (une population de n individus, à travers de ces variables qualitatives), on s'intéresse à l'indépendance de ces deux variables. Si ces variables sont indépendantes alors, l'AFC n'a aucun sens, pour cela, il est classique d'étudier la signficativité de la liaison entre les lignes et les colonnes. On doit faire un test non paramétrique. On propose une hypothèse nulle et une autre alternative comme suit :

$$\begin{cases} H_0 : \text{ les variables } V_1 \text{ et } V_2 \text{ sont indépendantes (pas de correspondance)} \\ H_1 : \text{ les variables sont liées (il y'a une correspondance)} \end{cases} \tag{1.1}$$

1.5 Test du χ^2 d'indépendance

Il y'a une autre façon d'écriture de l'équation (1.1). On peut le réécrire comme suit :

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

1.5.1 Statistique du χ^2

Avant d'accéder à cette méthode comme on a déjà dit il faut appliquer un test pour étudier la liaison entre ces deux variables. 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 [8].

Définition 1.5.1 *[8]La statistique du χ^2 s'écrit sous la forme :*

$$\chi^2 = \sum_{ij} \frac{(\textit{effectif observé} \ - \textit{effectif théorique})^2}{\textit{effectif théorique}} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n \ f_{ij} - n \ f_{i\cdot} f_{\cdot j})^2}{n \ f_{i\cdot} f_{\cdot j}}.$$

Remarque 1.5.1 *Si les deux variables V_1 et V_2 sont indépendantes alors $\chi^2 = 0$.*

En effet

$$V_1 \textit{ et } V_2 \textit{ sont indépendantes} \iff (f_{ij} = f_{i\cdot} f_{\cdot j}),$$

alors

$$\chi^2 = n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{0}{f_{i\cdot} f_{\cdot j}}$$

$$= 0.$$

1.5.2 Ecart à l'indépendance

Définition 1.5.2 *[11]Nous définissons l'écart à l'indépendance ϕ^2 par :*

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

Remarque 1.5.2 *[11]Lorsque n est assez grand, on a la convergence en distribution de la statistique du χ^2. Autrement dit, on a*

$$\chi^2 \underset{n\longrightarrow\infty}{\overset{D}{\longrightarrow}} \textit{loi de khi-deux}$$

à r ddl telque $(r = (p-1)(q-1))$.

L'écart à l'indépendance ϕ^2, qu'est la statistique du χ^2 divisé par n où n est l'effectif total, est appelée l'inertie totale en AFC. Il s'agit d'une mesure de la variance du tableau et ne dépend pas de la taille de l'échantillion. Cette quantité prend des autres noms tels que le coefficient de contingence quadratique moyenne [8].

Remarque 1.5.3 *[11]L'écart à l'indépendance ϕ^2 peut s'écrire en trois formes*

$$\begin{aligned}
\phi^2 &= \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot}f_{\cdot j})^2}{f_{i\cdot}f_{\cdot j}} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} f_{i\cdot}\frac{\left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}\right)^2}{f_{\cdot j}} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} f_{\cdot j}\frac{\left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot}\right)^2}{f_{i\cdot}}.
\end{aligned}$$

En effet,

$$\begin{aligned}\phi^2 &= \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(f_{ij}-f_{i\cdot}f_{\cdot j})^2}{f_{i\cdot}f_{\cdot j}} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(f_{i\cdot}\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)\right)^2}{f_{i\cdot}f_{\cdot j}} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{f_{i\cdot}^2\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)^2}{f_{i\cdot}f_{\cdot j}} = \sum_{i=1}^{p}\sum_{j=1}^{q}f_{i\cdot}\frac{\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)^2}{f_{\cdot j}}.\end{aligned}$$

D'où le résultat, même chose pour la $3^{ème}$ *forme.*

Chapitre 2

Principe de l'analyse factorielle des correspondances

L'analyse factorielle simple permet d'analyser le lien entre deux variables qualitatives. Sur le plan mathématique, on peut considérer l'AFC comme une ACP particulière qui utilise une métrique spéciale (la métrique du χ^2) [13]. Ce chapitre traite le principe de cette méthode et comment effectuer une AFC en utilisant l'analyse en composantes principales.

2.1 Transformation des données

En AFC, le tableau n'est pas analysé directement, c'est-à-dire au lieu de travailler avec le tableau N, on utilise des autres tableaux ce que nous appelons : 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 [7].

2.1.1 Tableau des profils-lignes

Définition 2.1.1 *[11]On définit la matrice diagonale de profils-lignes par*

$$D_r := \begin{pmatrix} f_{1\cdot} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{p\cdot} \end{pmatrix} \in \mathcal{M}(p \times p).$$

Définition 2.1.2 *[6]La matrice des profils-lignes X_r est obtenue en divisant chaque ligne i de N par son poids $f_{i\cdot}$*

$$X_r := D_r^{-1} N = \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix} \in \mathcal{M}(p \times q).$$

Le $i^{ème}$ profil-ligne est sous forme d'une distribution de fréquence conditionnelle de la variable V_2 sachant $V_1 = v_i$, le profil égale à :

$$f_i^{V2} := \left(\frac{f_{i1}}{f_{i\cdot}}, ..., \frac{f_{iq}}{f_{i\cdot}} \right)^t, \ i = 1, ..., p, \ [4].$$

2.1.2 Tableau des profils-colonnes

Définition 2.1.3 *[11]On définit la matrice diagonale de profils-colonnes par*

$$D_c := \begin{pmatrix} f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{\cdot q} \end{pmatrix} \in \mathcal{M}(q \times q).$$

Définition 2.1.4 *[6]La matrice des profils-colonnes X_c est obtenue en divisant chaque*

colonne j de N par son poids $f_{\cdot j}$

$$X_c := D_c^{-1} N^t = \begin{pmatrix} \frac{f_{11}}{f_{\cdot 1}} & \dots & \frac{f_{p1}}{f_{\cdot 1}} \\ \vdots & \ddots & \vdots \\ \frac{f_{1q}}{f_{\cdot q}} & \dots & \frac{f_{pq}}{f_{\cdot q}} \end{pmatrix} \in \mathcal{M}(q \times p).$$

De même à chaque modalité de V_2 on associe son profil, le $j^{ème}$profil colonne vaut

$$f_j^{V1} := \left(\frac{f_{1j}}{f_{\cdot j}}, ..., \frac{f_{pj}}{f_{\cdot j}} \right)^t, \ j = 1, ..., q, \ [4].$$

2.2 Centre de gravité de nuage de points

Définition 2.2.1 *[11]On définit le centre de gravité de profils-lignes et de profils-colonnes g_r et g_c respectivement par*

$$g_r = (f_{\cdot 1}, ..., f_{\cdot q})^t,$$

et

$$g_c = (f_{1\cdot}, ..., f_{p\cdot})^t,$$

où g_r est un vecteur de $q \times 1$ et g_c de $p \times 1$.

Remarque 2.2.1 *[11]Le centre de gravité g_r peut s'écrire aussi sous la forme suivante*

$$g_r = X_r^t D_r 1_p = N^t 1_p.$$

En effet,

$$\begin{aligned} X_r^t D_r 1_p &= \left(D_r^{-1} N\right)^t D_r 1_p = N^t \left(D_r^{-1}\right)^t D_r 1_p = N^t D_r^{-1} D_r 1_p = N^t 1_p \\ &= \begin{pmatrix} f_{11} & \cdots & f_{p1} \\ \vdots & \ddots & \vdots \\ f_{1q} & \cdots & f_{pq} \end{pmatrix} \times \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} f_{11} + \cdots + f_{p1} \\ \vdots \\ f_{1q} + \cdots + f_{pq} \end{pmatrix} \\ &= \begin{pmatrix} f_{\cdot 1} \\ \vdots \\ f_{\cdot q} \end{pmatrix} = g_r \in \mathcal{M}(q \times 1). \end{aligned}$$

De même, le centre de gravité de profils-colonnes g_c a une autre forme

$$g_c = X_c^t D_c 1_q = N 1_q.$$

En effet,

$$\begin{aligned} X_c^t D_c 1_q &= \left(D_c^{-1} N^t\right)^t D_c 1_q = N \left(D_c^{-1}\right)^t D_c 1_q = N D_c^{-1} D_c 1_q = N 1_q \\ &= \begin{pmatrix} f_{11} & \cdots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pq} \end{pmatrix} \times \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} f_{11} + \cdots + f_{1q} \\ \vdots \\ f_{p1} + \cdots + f_{pq} \end{pmatrix} \\ &= \begin{pmatrix} f_{1\cdot} \\ \vdots \\ f_{p\cdot} \end{pmatrix} = g_c \in \mathcal{M}(p \times 1). \end{aligned}$$

Exemple 2.2.1 *Considérons la matrice des données de la base «CSP Fillières» de*

l'exemple précédent :

$$N^* = \begin{pmatrix} 80 & 99 & 65 & 58 \\ 168 & 137 & 208 & 62 \\ 470 & 400 & 876 & 79 \\ 145 & 133 & 135 & 54 \\ 166 & 193 & 127 & 129 \end{pmatrix} \in \mathcal{M}(5 \times 4),$$

et l'effectif total vaut $n = 3784$. *Ainsi,*

$$N = \begin{pmatrix} 80/3784 & 99/3784 & 65/3784 & 58/3784 \\ 168/3784 & 137/3784 & 208/3784 & 62/3784 \\ 470/3784 & 400/3784 & 876/3784 & 79/3784 \\ 145/3784 & 133/3784 & 135/3784 & 54/3784 \\ 166/3784 & 193/3784 & 127/3784 & 129/3784 \end{pmatrix} \in \mathcal{M}(5 \times 4).$$

Les fréquences marginales-lignes sont

$$f_{1\cdot} = \sum_{j=1}^{4} f_{1j} = (80/3784 + 99/3784 + 65/3784 + 58/3784) = \frac{151}{1892}$$

$$f_{2\cdot} = \sum_{j=1}^{4} f_{2j} = (168/3784 + 137/3784 + 208/3784 + 62/3784) = \frac{575}{3784}$$

$$f_{3\cdot} = \sum_{j=1}^{4} f_{3j} = (470/3784 + 400/3784 + 876/3784 + 79/3784) = \frac{1825}{3784}$$

$$f_{4\cdot} = \sum_{j=1}^{4} f_{4j} = (145/3784 + 133/3784 + 135/3784 + 54/3784) = \frac{467}{3784}$$

$$f_{5\cdot} = \sum_{j=1}^{4} f_{5j} = (166/3784 + 193/3784 + 127/3784 + 129/3784) = \frac{615}{3784}.$$

Les fréquences marginales-colonnes sont

$$f_{\cdot 1} = \sum_{i=1}^{5} f_{i1} = (80/3784 + 168/3784 + 470/3784 + 145/3784 + 166/3784) = \frac{1029}{3784}$$
$$f_{\cdot 2} = \sum_{i=1}^{5} f_{i2} = (99/3784 + 137/3784 + 400/3784 + 133/3784 + 193/3784) = \frac{481}{1892}$$
$$f_{\cdot 3} = \sum_{i=1}^{5} f_{i3} = (65/3784 + 208/3784 + 876/3784 + 135/3784 + 127/3784) = \frac{1411}{3784}$$
$$f_{\cdot 4} = \sum_{i=1}^{5} f_{i4} = (58/3784 + 62/3784 + 79/3784 + 54/3784 + 129/3784) = \frac{191}{1892}.$$

On peut alors calculer la statistique du χ^2

$$\begin{aligned}
\chi^2 &= n \sum_{i=1}^{5} \sum_{j=1}^{4} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \\
&= n \times \left\{ \frac{\left(\frac{80}{3784} - \frac{151}{1892} \times \frac{1029}{3784}\right)^2}{\frac{151}{1892} \times \frac{1029}{3784}} + \frac{\left(\frac{168}{3784} - \frac{575}{3784} \times \frac{1029}{3784}\right)^2}{\frac{575}{3784} \times \frac{1029}{3784}} \right. \\
&\qquad \left. + \cdots + \frac{\left(\frac{129}{3784} - \frac{615}{3784} \times \frac{191}{1892}\right)^2}{\frac{615}{3784} \times \frac{191}{1892}} \right\} \\
&= 3784 \times 0.08464061 \\
&\simeq 320.28.
\end{aligned}$$

Les centres de gravités des profils-lignes et profils-colonnes, respectivement, sont

$$g_r = (1029/3784, 481/1892, 1411/3784, 191/1892)^t \in \mathcal{M}(4 \times 1),$$

et

$$g_c = (151/1892, 575/3784, 1825/3784, 467/3784, 615/3784)^t \in \mathcal{M}(5 \times 1).$$

Les matrices diagonales des profils-lignes et profils-colonnes respectivement sont

$$D_r = \begin{pmatrix} 151/1892 & 0 & 0 & 0 & 0 \\ 0 & 575/3784 & 0 & 0 & 0 \\ 0 & 0 & 1825/3784 & 0 & 0 \\ 0 & 0 & 0 & 467/3784 & 0 \\ 0 & 0 & 0 & 0 & 615/3784 \end{pmatrix} \in \mathcal{M}(5 \times 5),$$

et

$$D_c = \begin{pmatrix} 1029/3784 & 0 & 0 & 0 \\ 0 & 481/1892 & 0 & 0 \\ 0 & 0 & 1411/3784 & 0 \\ 0 & 0 & 0 & 191/1892 \end{pmatrix} \in \mathcal{M}(4 \times 4).$$

Les matrices profils-lignes et profils-colonnes, respectivement, sont

$$X_r = D_r^{-1}N = \begin{pmatrix} 40/151 & 99/302 & 65/302 & 29/151 \\ 168/575 & 137/575 & 208/575 & 62/575 \\ 94/365 & 16/73 & 12/25 & 79/1825 \\ 145/467 & 133/467 & 135/467 & 54/467 \\ 166/615 & 193/615 & 127/615 & 43/205 \end{pmatrix} \in \mathcal{M}(5 \times 4),$$

et

$$X_c = D_c^{-1}N^t = \begin{pmatrix} 80/1029 & 8/49 & 470/1029 & 145/1029 & 166/1029 \\ 99/962 & 137/962 & 200/481 & 133/962 & 193/962 \\ 65/1411 & 208/1411 & 876/1411 & 135/1411 & 127/1411 \\ 29/191 & 31/191 & 79/382 & 27/191 & 129/382 \end{pmatrix} \in \mathcal{M}(4\times5).$$

2.3 Métrique du χ^2

Il y'a plusieurs distances, comme la distance Euclidienne (Pythagore), on peut définir la distance euclidienne entre deux profils-lignes i et i' par la formule suivante :

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

de même pour la distance euclidienne entre deux profils-colonnes j et j'

$$d^2(j,j') = \sum_{i=1}^{p} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right) = \|j - j'\|^2 .$$

Question : Quelle est la bonne quantité qui mesure la dispersion des profils autour du centre de gravité ? [4]

Réponse : Dans l'AFC, pour mesurer la dispersion des profils autour du centre de gravité, on utilise la métrique du χ^2, qu' on l'appelle aussi la distance du χ^2 [4].

2.3.1 Distance du χ^2 entre deux profils-lignes

Définition 2.3.1 *[11]On définit la distance du χ^2 entre deux profils-lignes i et i' par :*

$$d^2_{\chi^2}(\textit{profil-ligne } i,\ \textit{profil-ligne } i') = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 = \|i - i'\|^2_{M_r} ,$$

où

$$M_r = D_c^{-1} = \begin{pmatrix} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{pmatrix} .$$

On définit aussi la distance du χ^2 entre le profil-ligne i et son centre de gravité g_r par la formule suivante :

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

2.3.2 Distance du χ^2 entre deux profils-colonnes

Définition 2.3.2 *[11]De la même manière, on définit la distance du χ^2 entre deux profils-colonnes j et j' par la formule suivante :*

$$d^2_{\chi^2}(\textit{profil-colonne } j,\ \textit{profil-colonne } j') = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 = \|j - j'\|^2_{M_c} ,$$

où

$$M_c = D_r^{-1} = \begin{pmatrix} 1/f_{1\cdot} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{p\cdot} \end{pmatrix} .$$

On définit aussi la distance du χ^2 entre le profil-colonne j et son centre de gravité g_c par la formule suivante :

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

Remarque 2.3.1 *[8]Cette métrique est similaire à la distance euclidienne avec la pondération du terme $1/f_{\cdot j}$ à chaque carrée de différence dans le cas du nuage profils-lignes et la pondération du terme $1/f_{i\cdot}$ dans le cas des profils-colonnes.*

Exemple 2.3.1 *La matrice des profils-lignes est*

$$X_r = \begin{pmatrix} 40/151 & 99/302 & 65/302 & 29/151 \\ 168/575 & 137/575 & 208/575 & 62/575 \\ 94/365 & 16/73 & 12/25 & 79/1825 \\ 145/467 & 133/467 & 135/467 & 54/467 \\ 166/615 & 193/615 & 127/615 & 43/205 \end{pmatrix},$$

la distance du χ^2 entre la première et la deuxième lignes est

$$\begin{aligned} d^2_{\chi^2}(1,2) &= \left\{ \frac{3784}{1029}(\frac{40}{151} - \frac{168}{575})^2 + \frac{1892}{481}(\frac{99}{302} - \frac{137}{575})^2 \right. \\ &\quad \left. + \frac{3784}{1411}(\frac{65}{302} - \frac{208}{575})^2 + \frac{1892}{191}(\frac{29}{151} - \frac{62}{575})^2 \right\} \\ &= 0.162\,12. \end{aligned}$$

2.4 Inertie totale

Nous allons présenter ici la formule d'inerties totales des nuages de points profils-lignes et profils-colonnes par rapport aux centres de gravité respectivement par

$$\text{Inertie } (X_r/g_r) := \sum_{i=1}^{p} f_{i\cdot} \times d^2_{\chi^2}(i, g_r),$$

et

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

Proposition 2.4.1 *[11]L'inertie totale d'un tableau de contingence est la statistique du χ^2 divisée par n, le total du tableau, qui est l'écart à l'indépendance ϕ^2.*

Preuve. [11]On a

$$\begin{aligned}
\text{Inertie } (X_r/g_r) &= \sum_{i=1}^{p} f_{i\cdot} \times d^2_{\chi^2}(i, g_r) = \sum_{i=1}^{p} f_{i\cdot} \times \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{i\cdot}}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{i\cdot}}{f_{\cdot j}} \left(\frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot}} \right)^2 \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{i\cdot}}{f_{\cdot j} \times f^2_{i\cdot}} \left(f_{ij} - f_{i\cdot} f_{\cdot j} \right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - f_{i\cdot} f_{\cdot j} \right)^2}{f_{\cdot j} f_{i\cdot}} \\
&= \frac{\chi^2}{n} = \phi^2.
\end{aligned}$$

De même,

$$\begin{aligned}
\text{Inertie } (X_c/g_c) &= \sum_{j=1}^{q} f_{\cdot j} \times d^2_{\chi^2}(j, g_c) = \sum_{j=1}^{q} f_{\cdot j} \times \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{\cdot j}}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{\cdot j}}{f_{i\cdot}} \left(\frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{\cdot j}} \right)^2 \\
&= \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{\cdot j}}{f_{i\cdot} \times f^2_{\cdot j}} \left(f_{ij} - f_{i\cdot} f_{\cdot j} \right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - f_{i\cdot} f_{\cdot j} \right)^2}{f_{i\cdot} f_{\cdot j}} \\
&= \frac{\chi^2}{n} = \phi^2.
\end{aligned}$$

D'où le résultat. ■

Remarque 2.4.1 *L'inertie I des deux nuages de points sont égaux, et vaut à l'écart à l'indépendance.*

$$\textit{Inertie } (X_r/g_r) = \textit{Inertie } (X_c/g_c) = \chi^2/n = \phi^2.$$

Exemple 2.4.1 *Dans notre exemple les inerties totales des nuages de points X_r et*

X_c sont

$$\begin{aligned} \textit{Inertie } (X_r/g_r) &= \textit{Inertie } (X_c/g_c) = \chi^2/n \\ &= 320.28/3784 \\ &\simeq 8.46 \times 10^{-2} = \phi^2. \end{aligned}$$

2.5 ACP des deux nuages

La méthode d'Analyse factorielle consiste à résumer les principales liaisons existantes entre les modalités de V_1 et V_2, et a un but de réduire la dimension. On peut considérer l'AFC comme une double ACP, une portant sur les profils-lignes et l'autre sur les profils-colonnes.

Définition 2.5.1 *[11]On définit le nuage profils-lignes centré par*

$$Y_r = X_r - 1_p g_r^t,$$

où 1_p est un vecteur unitaire de $(p \times 1)$ et

$$Y_r = \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} - f_{\cdot 1} & \cdots & \frac{f_{1q}}{f_{1\cdot}} - f_{\cdot q} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} - f_{\cdot 1} & \cdots & \frac{f_{pq}}{f_{p\cdot}} - f_{\cdot q} \end{pmatrix} \in \mathcal{M}(p \times q).$$

De façon symétrique, on définit le nuage profils-colonnes centré par

$$Y_c = X_c - 1_q g_c^t,$$

où 1_q *est un vecteur unitaire de* $(q \times 1)$ *et*

$$Y_c = \begin{pmatrix} \frac{f_{11}}{f_{\cdot 1}} - f_{1\cdot} & \cdots & \frac{f_{p1}}{f_{\cdot 1}} - f_{p\cdot} \\ \vdots & \ddots & \vdots \\ \frac{f_{1q}}{f_{\cdot q}} - f_{1\cdot} & \cdots & \frac{f_{pq}}{f_{\cdot q}} - f_{p\cdot} \end{pmatrix} \in \mathcal{M}(q \times p).$$

Proposition 2.5.1 *[11]Pour chercher les axes principaux de nuage des points des profils-lignes* Y_r*, il suffit de calculer les vecteurs propres de la matrice* $V_r M_r$*, où*

$$\begin{aligned} V_r &= Y_r^t D_r Y_r = X_r^t D_r X_r - g_r g_r^t \in \mathcal{M}(q \times q) \\ &= N^t D_r^{-1} D_r D_r^{-1} N - g_r g_r^t = N^t D_r^{-1} N - g_r g_r^t, \end{aligned}$$

et $M_r = D_c^{-1}$.

Preuve. [11]Soit E l'axe principale de l'ACP, et u son vecteur propre c'est-à-dire

$$\|u\|^2_{M_r} = u^t M_r u = 1,$$

la métrique utilisée n'est pas comme dans l'ACP classique, ici on utilise la métrique du χ^2. On note

$$\text{la projection du point } \mathbf{y}_i \text{ sur } E^\perp := \mathbf{proj}_{E^\perp, i}.$$

Nous définissons l'inertie du nuage Y_r par rapport à $E^\perp$ par

$$\text{Inertie}(Y_r \backslash E^\perp) = \sum_{i=1}^{p} f_{i\cdot} d^2_{\chi^2}(\mathbf{y}_i,\ \mathbf{proj}_{E^\perp, i}).$$

On note aussi

$$\text{La projection du point } \mathbf{y}_i \text{ sur } E := \frac{\langle y_i, u \rangle_{M_r} u}{\|u\|^2_{M_r}}.$$

A l'aide de la relation de Chasles, on a

$$\frac{\langle y_i, u\rangle_{M_r} u}{\|u\|^2_{M_r}} + \mathbf{proj}_{E^\perp,i} = \mathbf{y}_i,$$

ce qui implique

$$\mathbf{y}_i - \mathbf{proj}_{E^\perp,i} = \frac{\langle y_i, u\rangle_{M_r} u}{\|u\|^2_{M_r}},$$

alors

$$\begin{aligned}
d^2_{\chi^2}(\mathbf{y}_i,\ \mathbf{proj}_{E^\perp,i}) &= \left\|\mathbf{y}_i - \mathbf{proj}_{E^\perp,i}\right\|^2_{M_r} \\
&= \left\| \frac{\langle y_i, u\rangle_{M_r} u}{\|u\|^2_{M_r}} \right\|^2_{M_r} = \frac{\langle y_i, u\rangle^2_{M_r} \|u\|^2_{M_r}}{\|u\|^4_{M_r}} \\
&= \frac{\langle y_i, u\rangle^2_{M_r}}{\|u\|^2_{M_r}} = \langle y_i, u\rangle^2_{M_r} \,(\text{car } \|u\|^2_{M_r} = 1) \\
&= \left(y_i^t M_r u\right)^2 = \left(y_i^t M_r u\right)\left(y_i^t M_r u\right)^t \\
&= \left(y_i^t M_r u\right)\left(u^t M_r y_i\right) = \left(u^t M_r y_i\right)\left(y_i^t M_r u\right) \\
&= u^t M_r y_i y_i^t M_r u,
\end{aligned}$$

alors

$$\begin{aligned}
\text{Inertie}(Y_r \backslash E^\perp) &= \sum_{i=1}^{p} f_{i\cdot} d^2_{\chi^2}(\mathbf{y}_i,\ \mathbf{proj}_{E^\perp,i}) = \sum_{i=1}^{p} f_{i\cdot} u^t M_r y_i y_i^t M_r u \\
&= u^t M_r \left[\sum_{i=1}^{p} f_{i\cdot} y_i y_i^t\right] M_r u = u^t M_r \left[Y_r^t D_r Y_r\right] M_r u \\
&= u^t M_r V_r M_r u.
\end{aligned}$$

Maintenant, on va chercher le vecteur u qui maximise Inertie $(Y_r \backslash E^{\perp})$ sous la contrainte $\|u\|^2_{M_r} = 1$, à l'aide du multiplicateur de lagrange, on va maximiser la fonction

$$u \longrightarrow f(u) = u^t M_r V_r M_r u - \lambda(u^t M_r u - 1).$$

En dérivant cette fonction on obtient

$$f'(u) = 2M_r V_r M_r u - 2\lambda M_r u$$

puisque la matrice M_r est diagonale alors, elle est inversible. Donc on obtient

$$f'(u) = 0 \Longleftrightarrow V_r M_r u = \lambda u.$$

D'où le résultat. ■

Proposition 2.5.2 *[11]Les centres de gravités g_r et g_c sont des vecteurs propres de $V_r M_r$ et $V_c M_c$ respectivement associés à $\lambda = 0$.*

Preuve. On prend le cas de nuage profils-colonnes, on va démontrer que

$$V_c M_c g_c = 0_{\mathbb{R}^p} = 0 g_c.$$

En effet, on a

$$\begin{aligned} M_c g_c &= D_r^{-1} g_c \\ &= \begin{pmatrix} \frac{1}{f_{1\cdot}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{f_{p\cdot}} \end{pmatrix} \begin{pmatrix} f_{1\cdot} \\ \vdots \\ f_{p\cdot} \end{pmatrix} = 1_p. \end{aligned}$$

Alors

$$\begin{aligned} V_c M_c g_c = V_c 1_p &= \left(X_c^t D_c X_c - g_c g_c^t\right) 1_p \\ &= X_c^t D_c X_c 1_p - g_c g_c^t 1_p, \end{aligned}$$

où

$$\begin{aligned} g_c g_c^t 1_p &= g_c \times \begin{pmatrix} f_{1\cdot} & \ldots & f_{p\cdot} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= g_c \times (f_{1\cdot} + \ldots + f_{p\cdot}) \end{aligned}$$

alors

$$g_c g_c^t 1_p = g_c \times 1 = g_c.$$

Et

$$\begin{aligned} X_c^t D_c X_c 1_p &= X_c^t D_c \left(D_c^{-1} N^t\right) 1_p \\ &= X_c^t N^t 1_p \\ &= X_c^t \times \begin{pmatrix} f_{11} & \cdots & f_{p1} \\ \vdots & \ddots & \vdots \\ f_{1q} & \cdots & f_{pq} \end{pmatrix} \times \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= X_c^t \times \begin{pmatrix} f_{11} + \cdots + f_{p1} \\ \vdots \\ f_{1q} + \cdots + f_{pq} \end{pmatrix} \\ &= X_c^t \times \begin{pmatrix} f_{\cdot 1} \\ \vdots \\ f_{\cdot q} \end{pmatrix} \end{aligned}$$

Nous avons donc

$$X_c^t D_c X_c 1_p = X_c^t g_r.$$

Ce dernier égal à g_c. En effet,

$$\begin{aligned} X_c^t g_r &= \begin{pmatrix} \frac{f_{11}}{f_{\cdot 1}} & \cdots & \frac{f_{1q}}{f_{\cdot q}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{\cdot 1}} & \cdots & \frac{f_{pq}}{f_{\cdot q}} \end{pmatrix} \begin{pmatrix} f_{\cdot 1} \\ \vdots \\ f_{\cdot q} \end{pmatrix} \\ &= \begin{pmatrix} f_{11} + ... + f_{1q} \\ \vdots \\ f_{p1} + ... + f_{pq} \end{pmatrix} = \begin{pmatrix} f_{1\cdot} \\ \vdots \\ f_{p\cdot} \end{pmatrix} \\ &= g_c. \end{aligned}$$

Finalement, on trouve

$$\begin{aligned} V_c M_c g_c &= X_c^t D_c X_c 1_p - g_c g_c^t 1_p \\ &= g_c - g_c \\ &= 0. \end{aligned}$$

D'où le résultat. ■

2.6 Lien entre l'ACP des profils-lignes et des profils-colonnes

On peut faire l'ACP sans centrer le nuage des profils-lignes et des profils-colonnes, dans le cas de nuage profils-lignes, on travaille avec $A_r = X_r^t X_c^t = N^t D_r^{-1} N D_c^{-1}$, ($A_r \in \mathcal{M}(q \times q)$), c'est-à-dire au lieu de chercher les valeurs et les vecteurs propres de $V_r M_r$, il suffit de chercher les valeurs et les vecteurs propres de A_r.

De la même façon, on travaille avec $A_c = X_c^t X_r^t = N D_c^{-1} N^t D_r^{-1}$, $(A_c \in \mathcal{M}(p \times p))$ dans le cas des profils-colonnes.

Remarque 2.6.1 *[11]Les deux matrices A_r et A_c ont les mêmes valeurs propre non nulles, et on a*

$$\tau := rang\ V_r M_r = rang\ V_c M_c,$$

et

$$\tau + 1 := rang\ A_r = rang\ A_c,$$

De plus

$$0 < \tau \leqslant \min(p-1,\ q-1).$$

En effet, le rang d'une matrice carée égale au nombre des valeurs propres non nulles, on sait que $V_r M_r$ est une matrice carrée de $\mathcal{M}(q \times q)$, et admet g_r comme un vecteur propre associé à $\lambda = 0$, alors rang $V_r M_r \leqslant q-1$, et pour la matrice carrée $V_c M_c$ de $\mathcal{M}(p \times p)$, admet g_c comme un vecteur propre associé à $\lambda = 0$, alors rang $V_c M_c \leqslant p-1$, ce qui implique

$$0 < \tau \leqslant \min(p-1, q-1).$$

L'ACP des profils-lignes et l'ACP des profils-colonnes sont les mêmes. Dans le pratique, on fait l'ACP sur la plus petite matrice [2].

Remarque 2.6.2 *[11]La matrice A_r a les mêmes valeurs propres non nulles de la matrice $V_r M_r$ sauf*

$$(\lambda = 0,\ g_r)\ de\ V_r M_r \Leftrightarrow (\lambda = 1,\ g_r)\ de\ A_r.$$

de même, la matrice A_c a les mêmes valeurs propres non nulles de la matrice $V_c M_c$ sauf

$$(\lambda = 0,\ g_c)\ de\ V_c M_c \Leftrightarrow (\lambda = 1,\ g_c)\ de\ A_c.$$

Proposition 2.6.1 *[2]Si u est un vecteur propre de A_r associé à λ avec $\|u\|^2_{M_r} = 1$ alors*

$$\widetilde{u} = \frac{1}{\sqrt{\lambda}} X_c^t u$$

est un vecteur propre, de norme 1 pour la métrique M_c, pour A_c, pour la même valeur propre. De façon symétrique, si $\widetilde{u}$ est un vecteur propre de A_c associé à λ avec $\|\widetilde{u}\|^2_{M_c} = 1$ alors

$$u = \frac{1}{\sqrt{\lambda}} X_r^t \widetilde{u}$$

est un vecteur propre, de norme 1 *pour la métrique M_r, pour A_r, pour la même valeur propre.*

Preuve. $\widetilde{u}$ est un vecteur propre de A_c c'est-à-dire

$$\begin{aligned}
A_c\widetilde{u} &= \lambda\widetilde{u} \\
X_r^t\left(A_c\widetilde{u}\right) &= \left(\lambda\widetilde{u}\right)X_r^t \\
X_r^t X_c^t X_r^t\widetilde{u} &= \left(\lambda\widetilde{u}\right)X_r^t \\
A_r(N^t D_r^{-1}\widetilde{u}) &= \lambda(N^t D_r^{-1}\widetilde{u}),
\end{aligned}$$

donc $N^t D_r^{-1}\widetilde{u}$ est un vecteur propre de A_r, pour $N^t D_r^{-1}\widetilde{u}$ est de norme 1 pour la métrique M_r, il suffit qu'il existe un constant k telque

$$\begin{aligned}
\left(kN^t D_r^{-1}\widetilde{u}\right)^t M_r \left(kN^t D_r^{-1}\widetilde{u}\right) &= 1 \\
k\widetilde{u}^t D_r^{-1} N M_r \left(kN^t D_r^{-1}\widetilde{u}\right) &= 1 \\
k^2\widetilde{u}^t D_r^{-1}(X_c^t X_r^t)\widetilde{u} &= 1 \\
k^2\widetilde{u}^t D_r^{-1} A_c\widetilde{u} &= 1 \\
k^2\widetilde{u}^t D_r^{-1}\lambda\widetilde{u} &= 1 \\
k^2\lambda\widetilde{u}^t D_r^{-1}\widetilde{u} &= 1
\end{aligned}$$

puisque $\widetilde{u}$ est un vecteur propre, de norme 1 pour la métrique M_c , on a $\widetilde{u}^t D_r^{-1}\widetilde{u} = 1$, ce qui implique $k = \frac{1}{\sqrt{\lambda}}$, et $u = \frac{1}{\sqrt{\lambda}} X_r^t \widetilde{u}$. ■

Exemple 2.6.1 *Faisons l'ACP pour notre exemple, on travaille avec la plus petite matrice* $A_r = X_r^t X_c^t \in \mathcal{M}(4 \times 4)$, *rappelons que la matrice* X_r *est*

$$X_r = \begin{pmatrix} 40/151 & 99/302 & 65/302 & 29/151 \\ 168/575 & 137/575 & 208/575 & 62/575 \\ 94/365 & 16/73 & 12/25 & 79/1825 \\ 145/467 & 133/467 & 135/467 & 54/467 \\ 166/615 & 193/615 & 127/615 & 43/205 \end{pmatrix},$$

et la matrice X_c

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

alors

$$\begin{aligned} A_r &= X_r^t X_c^t \\ &= \begin{pmatrix} 0.273\,22 & 0.273\,03 & 0.269\,16 & 0.275\,94 \\ 0.255\,25 & 0.261\,14 & 0.241\,79 & 0.280\,01 \\ 0.369\,08 & 0.354\,65 & 0.407\,49 & 0.301\,26 \\ 0.102\,44 & 0.111\,19 & 8.156\,0 \times 10^{-2} & 0.142\,79 \end{pmatrix}. \end{aligned}$$

Les valeurs propres de A_r *sont*

$$\lambda_1 = 8.239\,5 \times 10^{-2},\ \lambda_2 = 1.704\,1 \times 10^{-3},\ \lambda_3 = 5.415\,1 \times 10^{-4},\ \lambda_4 = 1.$$

Les vecteurs propres associés sont

$$u_1 = \begin{pmatrix} 0.055\,67 \\ 0.298\,56 \\ -0.827\,13 \\ 0.472\,89 \end{pmatrix}, \; u_2 = \begin{pmatrix} 0.793\,77 \\ 3.184\,8 \times 10^{-2} \\ -0.531\,32 \\ -0.294\,30 \end{pmatrix}$$

$$u_3 = \begin{pmatrix} 0.400\,8 \\ -0.854\,85 \\ 0.174\,61 \\ 0.279\,45 \end{pmatrix}, \; u_4 = \begin{pmatrix} 0.506\,88 \\ 0.473\,88 \\ 0.695\,06 \\ 0.188\,17 \end{pmatrix}.$$

On remarque que le vecteur u_4 associé à $\lambda = 1$ est le centre de gravité de profils-lignes

$$g_r = \frac{u_4}{c}, \;\; avec \; c = 1.\,864.$$

On va normaliser les vecteurs précedents par la métrique $M_r = D_c^{-1}$, $u_i^* = u_i / \sqrt{\|u_i\|^2_{M_r}}$ pour $i = 1, 2, 3, 4$

$$u_1^* = \frac{u_1}{\sqrt{u_1^t M_r u_1}} = \frac{1}{\sqrt{4.\,411\,9}} \begin{pmatrix} 0.055\,67 \\ 0.298\,56 \\ -0.827\,13 \\ 0.472\,89 \end{pmatrix} = \begin{pmatrix} 2.\,650\,4 \times 10^{-2} \\ 0.142\,14 \\ -0.393\,79 \\ 0.225\,14 \end{pmatrix},$$

et

$$u_2^* = \frac{u_2}{\sqrt{u_2^t M_r u_2}} = \frac{1}{\sqrt{3.\,936}} \begin{pmatrix} 0.793\,77 \\ 3.184\,8 \times 10^{-2} \\ -0.531\,32 \\ -0.294\,30 \end{pmatrix} = \begin{pmatrix} 0.400\,10 \\ 1.\,605\,3 \times 10^{-2} \\ -0.267\,81 \\ -0.148\,34 \end{pmatrix},$$

et aussi,

$$u_3^* = \frac{u_3}{\sqrt{u_3^t M_r u_3}} = \frac{1}{\sqrt{4.320\,5}} \begin{pmatrix} 0.400\,8 \\ -0.854\,85 \\ 0.174\,61 \\ 0.279\,45 \end{pmatrix} = \begin{pmatrix} 0.192\,82 \\ -0.411\,27 \\ 8.400\,4 \times 10^{-2} \\ 0.134\,44 \end{pmatrix}.$$

Finalement,

$$u_4^* = \frac{u_4}{\sqrt{u_4^t M_r u_4}} = \frac{1}{\sqrt{3.474\,5}} \begin{pmatrix} 0.506\,88 \\ 0.473\,88 \\ 0.695\,06 \\ 0.188\,17 \end{pmatrix} = \begin{pmatrix} 0.271\,93 \\ 0.254\,23 \\ 0.372\,89 \\ 0.100\,95 \end{pmatrix} = g_r.$$

Les axes principaux de profils-lignes sont

$$E_i = \mathit{Vect}\ \{u_i^*\}, i = 1, 2, 3, 4.$$

Où Vect $\{u_i^*\}$ *est un sous espace vectoriel engendré par la famille de vecteurs* u_i^*, *c'est l'ensemble de toutes les combinaisons linéaires de vecteurs* u_i^*.

On a

$$\tau = \mathit{rang}\ A_r - 1 = 4 - 1 = 3.$$

L'inertie totale :

$$\begin{aligned} I_T &= \sum_{k=1}^{\tau=3} \lambda_k \\ &= 8.239\,5 \times 10^{-2} + 1.704\,1 \times 10^{-3} + 5.415\,1 \times 10^{-4} \\ &\simeq 8.46 \times 10^{-2}. \end{aligned}$$

Ensuite, on calcule les inerties du nuage de points de profils-lignes par rapport aux axes principaux

$$Inertie(X_r/E_1^{\perp}) = \lambda_1 = 8.\,239\,5 \times 10^{-2},$$

et

$$Inertie(X_r/E_2^{\perp}) = \lambda_2 = 1.\,704\,1 \times 10^{-3},$$

et finalement,

$$Inertie(X_r/E_3^{\perp}) = \lambda_3 = 5.\,415\,1 \times 10^{-4}.$$

2.7 Facteurs principaux et Composantes principales

Si les vecteurs propre sont identifiés, alors on peut déduire les facteurs principaux et les composantes principales [2].

Définition 2.7.1 *Soit u un vecteur propre de A_r associé à la valeur propre λ, le vecteur $w = M_r u$ est dite facteur principal pour le nuage des profils-lignes, et $c = Y_r w$ est son composante principale. Inversement, $\widetilde{u}$ un vecteur propre de A_c associé à la valeur propre λ, le vecteur $\widetilde{w} = M_c \widetilde{u}$ est le facteur principal pour le nuage des profils-colonnes, et $\widetilde{c} = Y_c \widetilde{w}$ est la composante principale correspondante.*

Remarque 2.7.1 *[11]Si u un vecteur propre de $V_r M_r$ associé à la valeur propre $\lambda \neq 0$, alors son composante principale vaut*

$$c = X_r w.$$

En effet

$$\begin{aligned} c &= Y_r w \\ &= \left(X_r - 1_p g_r^t\right) M_r u \\ &= X_r M_r u - 1_p g_r^t M_r u, \end{aligned}$$

comme g_r *est un vecteur propre de* $V_r M_r$ *associé à la valeur propre* $\lambda = 0$ *et* u *un vecteur propre de* $V_r M_r$ *associé à la valeur propre* $\lambda \neq 0$ *alors* g_r *et* u *sont* M_r*-orthogonaux, Autrement dit, le produit scalaire entre* g_r *et* u *par la métrique* M_r *égale à* 0 $\left(\langle g_r, u\rangle_{M_r} = g_r^t M_r u = 0\right)$. *Donc*

$$\begin{aligned} c &= X_r M_r u - 1_p g_r^t M_r u \\ &= X_r M_r u - 1_p \times 0 \\ &= X_r M_r u \\ &= X_r w. \end{aligned}$$

Et on a aussi si $u = g_r$ *vecteur propre de* $V_r M_r$ *associé à la valeur propre* $\lambda = 0$ *, alors* $c = 0$. *En effet,*

$$\begin{aligned} c &= Y_r w \\ &= X_r M_r g_r - 1_p g_r^t M_r g_r \\ &= X_r M_r g_r - 1_p \times 1, \ (\|g_r\|_{M_r}^2 = 1) \\ &= X_r M_r g_r - 1_p. \end{aligned}$$

On calcule $X_r M_r g_r$

$$\begin{aligned} X_r M_r g_r &= X_r D_c^{-1} g_r \\ &= X_r \times \begin{pmatrix} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{pmatrix} \begin{pmatrix} f_{\cdot 1} \\ \vdots \\ f_{\cdot q} \end{pmatrix} \\ &= X_r 1_q \end{aligned}$$

on trouve

$$\begin{aligned} X_r M_r g_r &= X_r 1_q \\ &= \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \\ &= \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} + \cdots + \frac{f_{1q}}{f_{1\cdot}} \\ \vdots \\ \frac{f_{p1}}{f_{p\cdot}} + \cdots + \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix} = \begin{pmatrix} \frac{f_{1\cdot}}{f_{1\cdot}} \\ \vdots \\ \frac{f_{p\cdot}}{f_{p\cdot}} \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1_p. \end{aligned}$$

Alors

$$\begin{aligned} c &= X_r M_r g_r - 1_p \\ &= 1_p - 1_p \\ &= 0. \end{aligned}$$

On résume que le nombre des composantes principales égale à τ.

Exemple 2.7.1 *Calculons les composantes principales* $c_k = X_r M_r u_k^*$, $k = \overline{1,\tau}$.

$$c_1 = X_r M_r u_1^* = \begin{pmatrix} 0.410\,12 \\ 2.014\,3 \times 10^{-2} \\ -0.262\,73 \\ 0.142\,09 \\ 0.451\,48 \end{pmatrix},$$

et

$$c_2 = X_r M_r u_2^* = \begin{pmatrix} -2.633\,7 \times 10^{-2} \\ 2.667\,7 \times 10^{-2} \\ -1.559\,6 \times 10^{-2} \\ 9.728\,3 \times 10^{-2} \\ -3.958\,3 \times 10^{-2} \end{pmatrix},$$

et finalement,

$$c_3 = X_r M_r u_3^* = \begin{pmatrix} -3.822\,9 \times 10^{-2} \\ 0.046\,82 \\ -6.177 \times 10^{-3} \\ -2.144\,6 \times 10^{-2} \\ 9.576\,2 \times 10^{-3} \end{pmatrix}.$$

Théorème 2.7.1 *Pour tout* $k = 1, ..., \tau$, *on a*

$$0 < \lambda_k \leq 1.$$

Preuve. (Voir [11].) ■

2.8 Contribution des profils

Définition 2.8.1 *[2]On définit la contribution d'une ligne i par*

$$CTR(i) = \frac{f_{i\cdot}c_i^2}{\lambda},\ i = 1, ..., p,$$

avec c_i la i-ème coordonnée de c.

De même, on définit la contribution d'une colonne j par

$$CTR(j) = \frac{f_{\cdot j}\widetilde{c}_j^2}{\lambda},\ j = 1, ..., q,$$

avec $\widetilde{c}_j$ la j-ème coordonnée de $\widetilde{c}$.

Exemple 2.8.1 *On calcule la contribution de $1^{ère}$ ligne par rapport à $E_1^{\perp}$:*

$$CTR(1,1) = \frac{f_{1\cdot}c_1^2(1)}{\lambda_1} = \frac{\frac{151}{1892} \times (0.410\,12)^2}{8.\,239\,5 \times 10^{-2}} = 0.162\,92 \simeq 16.3\%.$$

Chapitre 3

La mise en œuvre avec R

Ce chapitre traite la mise en œuvre dans l'environnement R, on peut le télécharger gratuitement. Nous montrons comment effectuer une AFC avec ce logiciel. Il est important de souligner que ce chapitre est la composition des deux documents suivants : [9] et [10].

3.1 Différents packages

Plusieurs packages sont disponibles dans le logiciel R pour appliquer une AFC : [4]

- Le package FactoMineR (Factor analysis and Data Mining with R).
- Le package ade4 (Analysis of Environmental Data : Exploratory and Euclidean method).
- Le package ca (Simple, Multiple and Joint correspondence Analysis).
- Le package MASS.

3.1.1 Les fonctions

Dans cette partie, on décrit les fonctions : [4]

– **CA()** [package FactoMineR], "Correspondence Analysis en Anglais".

– **ca**() [package ca].

– **corresp()** [package MAAS].

Nous utiliserons les deux packages FactoMineR (pour l'analyse) et factoextra (pour extraire et visualiser les résultats d'AFC).

3.1.2 Installation des deux packages

La première étape consiste à installer et charger ces deux packages comme suit :

– install.packages("FactoMineR") et library("FactoMineR").

– install.packages("factoextra") et library("factoextra").

3.2 Données Jeux Olympiques

Le jeu de données doit être un tableau de contingence, on utilise les données Jeux Olympiques (JO). Ces données sont disponibles dans le package FactoMineR.

> data(JO)

> JO

Description : Cette base de données est une table de contingence avec les événements d'athlétisme (en ligne) et les pays (en colonnes). Chaque cellule donne le nombre de médailles obtenues lors des 5 jeux olympiques de 1992 à 2008 (Barcelone 1992, Atlanta 1996, Sydney 2000, Athènes 2004, Pékin 2008).

	usa	ken	rus	gbr	eth	cub	mar	ger	jam	pol
10000 m	0	4	0	0	8	0	2	0	0	0
100 m	5	0	0	1	0	0	0	0	1	0
110 mH	9	0	0	0	0	3	0	1	0	0
1500 m	0	5	0	0	0	0	3	0	0	0
200 m	8	0	0	1	0	0	0	0	1	0
20Km	0	0	3	0	0	0	0	0	0	1
3000mSteeple	0	12	0	0	0	0	1	0	0	0
400m	11	1	0	1	0	0	0	0	1	0
400mH	7	0	0	1	0	0	0	0	2	0
4×100m	4	0	0	1	0	2	0	0	1	0
4×400m	5	0	1	2	0	1	0	0	2	0
5000m	0	5	0	0	4	0	3	1	0	0
50Km	0	0	4	0	0	0	0	1	0	3
800m	1	5	1	0	0	0	0	1	0	0
Decathlon	5	0	0	0	0	1	0	1	0	0
Disque	0	0	0	0	0	1	0	3	0	1
Hauteur	3	0	3	2	0	2	0	0	0	1
Javelot	0	0	2	3	0	0	0	0	0	0
Longueur	7	0	0	0	0	2	0	0	1	0
Marathon	1	3	0	0	3	0	1	1	0	0

TAB. 3.1 – Tableau représente les données JO de 10 pays, partie1.

	usa	ken	rus	gbr	eth	cub	mar	ger	jam	pol
Marteau	1	0	0	0	0	0	0	0	0	1
Perche	4	0	3	0	0	0	0	1	0	0
Poids	8	0	0	0	0	0	0	0	0	1
Triple saut	3	0	2	3	0	2	0	0	0	0

TAB. 3.2 – Tableau représente les données JO de 10 pays, partie2.

3.3 Code R pour calculer l'AFC

Il s'agit d'une étude de liaison entre les deux variables suivantes :

$$V_1 := \text{événements d'athlétisme de 24 modalités,}$$

et

$$V_2 := \text{pays de 58 modalités.}$$

Les individus sont les 360 médailles ($n = 360$).

3.3.1 Test du χ^2

La première étape consiste à étudier la liason entre les deux variables : épreuve et pays à l'aide de test du χ^2. On utilise la commande **chisq.test** pour effectuer le test. > test <-chisq.test(JO), on obtient

Pearson's Chi-squared test

data : JO

$$\text{X-squared} = 2122.2,\ \text{df} = 1311,\ \text{p-value} < 2.2e-16.$$

(La p-value $< \alpha = 0.05$), ce qui montre qu'il y a une liaison entre les deux variables (où la dépendance). Ici, on considère tous les éléments sont actifs c'est-à-dire les lignes et les colonnes supplémentaires sont nulles (row.sup=NULL, et col.sup=NULL). Pour effectuer une AFC, on utilise la fonction **CA.**

> res <- CA(JO).

Cette commande donne une liste contenant les valeurs propres, les pourcentages d'inerties associés à chaque dimension, les coordonnées des lignes et des colonnes, la qualité de représentation et les contributions de profils. Pour obtenir par exemple les contributions des lignes il suffit de taper la commande (>resrowcontib).

> summary.CA(res), # pour l'impression de résumés d'objets d'analyse des correspondances.

Les résultats pour profils-lignes (10 lignes) :

Les coordonnées

	Dim 1	Dim 2	Dim 3
10000m	-2.162	-0.330	-0.172
100m	0.678	-1.164	-0.407
110mH	0.593	-0.498	-0.395
1500m	-1.469	-0.185	0.373
200m	0.716	-1.084	-0.468
20km	0.284	1.037	1.476
3000mSt	-1.610	-0.147	0.127
400m	0.480	-0.736	-0.312
400mH	0.532	-0.785	-0.406
4×100m	0.550	-0.654	-0.397

Les coordonnées des lignes représentent les composantes principales c_1, c_2, c_3.

Les contributions

	Dim 1	Dim 2	Dim 3
10000m	23.850	0.730	0.227
100m	2.347	9.093	1.267
110mH	1.795	1.665	1.196
1500m	11.016	0.229	1.067
200m	2.612	7.889	1.679
20km	0.411	7.213	16.677
3000mSt	13.230	0.146	0.123
400m	1.177	3.639	0.746
400mH	1.444	4.137	1.265
4×100m	1.542	2.870	1.204

Le cosinus carré

	Dim 1	Dim 2	Dim 3
10000m	0.531	0.012	0.003
100m	0.073	0.215	0.026
110mH	0.093	0.066	0.041
1500m	0.266	0.004	0.017
200m	0.107	0.245	0.046
20km	0.010	0.135	0.274
3000mSt	0.399	0.003	0.002
400m	0.070	0.165	0.030
400mH	0.044	0.097	0.026
4×100m	0.062	0.088	0.032

Les résultats pour profils-colonnes (10 colonnes) :

Les coordonnées

	Dim1	Dim2	Dim3
alg	-0.997	-0.105	0.342
aus	0.446	0.594	0.959
bah	0.691	-0.645	0.442
bar	0.751	-1.477	-0.552
bdi	-2.066	-0.238	-0.120
blr	0.421	1.635	-1.376
bra	-0.016	-0.543	-0.516
brn	-1.626	-0.234	0.506
can	0.582	-0.406	0.094
chn	0.656	-0.632	-0.536

Les coordonnées des colonnes représentent les composantes principales $\widetilde{c}_1$, $\widetilde{c}_2$, $\widetilde{c}_3$.

Les contributions

	Dim1	Dim2	Dim3
alg	1.352	0.020	0.239
aus	0.406	0.948	2.815
bah	0.487	0.558	0.299
bar	0.192	0.977	0.155
bdi	1.452	0.025	0.007
blr	0.361	7.175	5.796
bra	0.000	0.395	0.408
brn	0.899	0.025	0.131
can	0.461	0.295	0.018
chn	0.147	0.179	0.147

Le cosinus carré

	Dim1	Dim2	Dim3
alg	0.199	0.002	0.023
aus	0.046	0.082	0.212
bah	0.039	0.034	0.016
bar	0.024	0.095	0.013
bdi	0.186	0.002	0.001
blr	0.031	0.472	0.334
bra	0.000	0.024	0.022
brn	0.115	0.002	0.011
can	0.068	0.033	0.002
chn	0.019	0.017	0.012

3.3.2 Représentation des données

Pour aider à l'extraction et la visualisation des résultats de l'analyse factorielle, on utilise le package [factoextra].

La fonction get_eigenvalue est disponible dans [factoextra] et elle a un rôle d'extraction les valeurs propres pour déterminer le nombre d'axes principaux.

> eig.val <- get_eigenvalue (res)

> eig.val, le résultat est de 23 dimensions (23 axes principaux) car le nombre des valeurs propres non nulles ou le nombre des inerties éxpliquées non nulles ne dépasse pas $\min(24-1, 58-1) = \min(23, 57) = 23$. On peut citer comme un exemple juste de la première à la quatrième dimension.

##	Eigenvalues	% of var	Cumulative % of var
Dim.1	0.82	13.85	13.85
Dim.2	0.62	10.53	24.38
Dim.3	0.54	9.23	33.62
Dim.4	0.48	8.16	41.78

TAB. 3.3 – Exraction les valeurs propres et les variances pour quatre dimensions.

Le premier et le deuxième plan expriment 24.4 % et 17.4 % de l'inertie totale. Il faut interpréter les axes suivants qui expriment un pourcentage important de l'inertie totale.

La visualisation des pourcentages d'inerties : pour visualiser les pourcentages d'inerties associés à chaque dimension, on utilise la commande suivante : > fviz_eig(res)

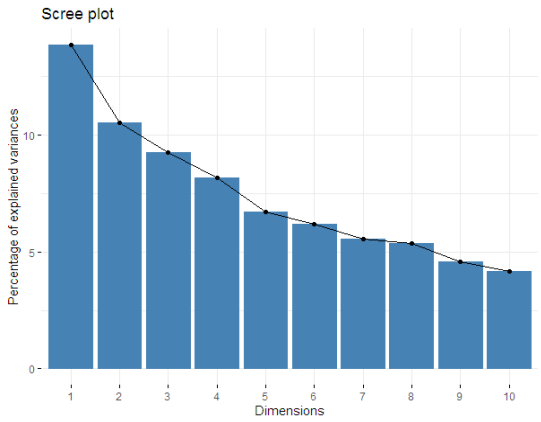

FIG. 3.1 – Données JO : les pourcentages d'inerties associés à chaque dimension.

La visualisation des valeurs propres : les valeurs propres nous donnent une idée sur la quantité d'informations retenue par chaque axe. On crée un diagramme en barres des valeurs propres avec barplot qui est disponible dans le package [graphics] avec les commandes suivantes :

> noms_barres<-c(1 :nrow(res$eig))

> barplot(res$eig[, 1], main="Valeur propres", names.arg=noms_barres, col="green").

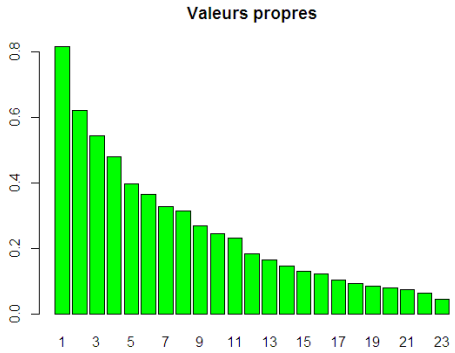

FIG. 3.2 – Données JO : valeurs propres associées à chaque dimension.

Visualisation de cos2 des lignes : la qualité de représentation est mésurée par cos2. Le code R suivant nous permet de créer un diagramme en barres de la qualité de représentation des lignes sur le premier plan.

> fviz_cos2(res, choice="row", axes=1 :2). Le point est parfaitement représenté sur l'axe, si la qualité est proche de 1.

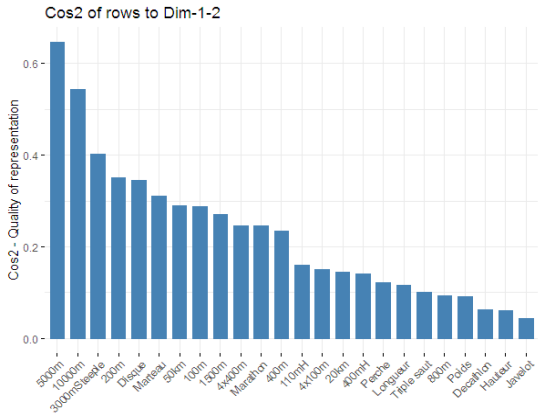

FIG. 3.3 – Données Jo : qualité de représentation des lignes sur le premier plan.

Visualisation des contributions : on tape la commande suivante pour obtenir un graphe de contribution des lignes sur le premier axe.

> fviz_contrib(res, choice = "row", axes = 1, top = 15), # (top=15) pour préciser le nombre des lignes.

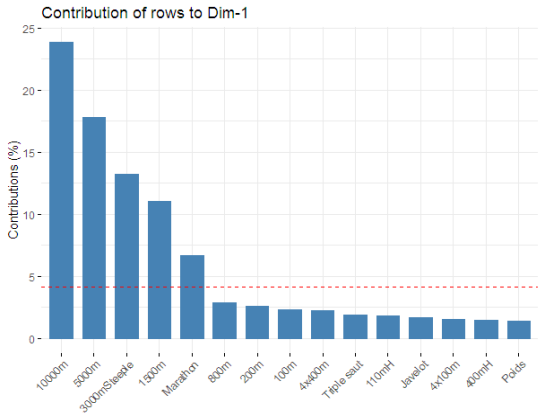

FIG. 3.4 – Données JO : contributions des lignes sur le premier axe.

La fonctin fviz_ca_biplot() est aussi disponible dans [factoextra], elle permet de faire la représentation superposée sur le plan.

> fviz_ca_biplot (res, repel = TRUE), # repel=TRUE pour éviter le chevauchement de texte.

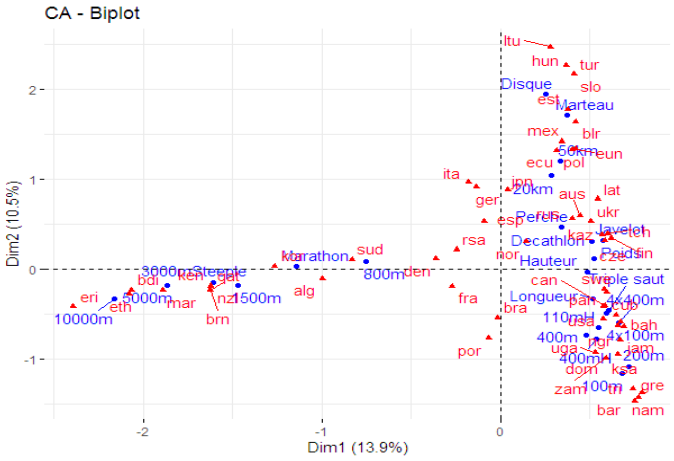

FIG. 3.5 – Données JO : repésentation sur le plan(1,2).

on remarque que les lignes sont représentées par des points bleus et les colonnes par des triangles rouges. Le plan(1,2) exprime 24.40% de l'inertie totale.

Si on veut tracer le graphe des points lignes ou colonnes, on utilise la foncion fviz_ca_row() et fviz_ca_col() [dans factoextra] et on tape : > fviz_ca_row(res, repel=TRUE).

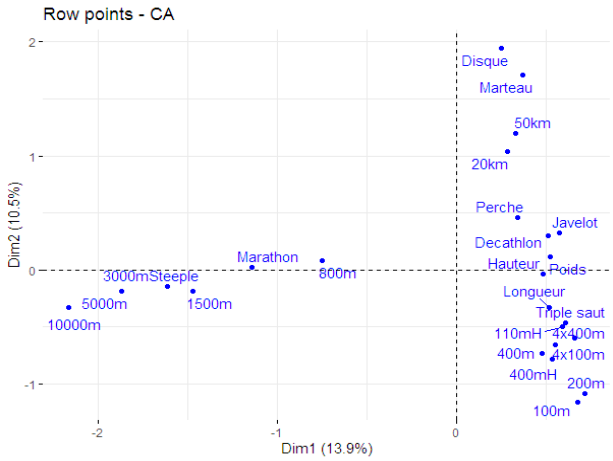

FIG. 3.6 – Données JO : graphique des points lignes.

3.3.3 Interprétation du plan factoriel

La figure (3.5) représente les projections des modalités des deux variables sur le premier plan factoriel. C'est la représentation d'AFC des profils-lignes et des profils-colonnes. Nous avons le premier axe principale de pourcentage d'inertie égale à 13.85% et le deuxième axe de pourcentage d'inertie égale à 10.53%. On peut citer quelques remarques comme suit :

- Les lignes 3000 m steeple, 10000 m, 5000 m et 1500 m, on peut les associer ensemble.
- Pour les épreuves du Disque et du Marteau, on retrouve que les pays de Estonie, Lituanie, Hongrie, Slovénie et Turquie sont les plus performants.
- Pour l'épreuve du Javelot, on retrouve que les pays Norvège, Finlande, Tchéquie, et Tchécoslovaquie sont les plus performants.
- Les lignes Marathon et 800 m sont associées le plus à la colonne Sud.
- Les lignes qui sont loins de l'origine sont bien représentés sur le graghique.
- On trouve des pays africains de même ensemble (l'Erithée, l'Ethiopie, le Burnudi, le Maroc, Qatar et Kenya) et aussi la nouvelle-Zélande.

Cette figure nous donne une idée sur le lien qui existe entre les modalités du même vecteur (ligne ou bien colonne) et aussi elle nous donne des informations sur la liaison existante entre les lignes et les colonnes.

Conclusion

En conclusion, ce mémoire nous s'a permis de comprendre la méthode d'analyse factorielle des correspondances, vu sa importance et sa utilisation dans beaucoup domaines telsque l'économie, gestion,...etc.

L'analyse des correspondances est la méthode privilégiée d'étude des liaisons entres deux variables qualitatives, et elle a un but de réduire la dimension. Dans ce mémoire, nous avons essayé comment effectuer et interpréter les résultats d'AFC appliquée à des données réelles en utilisant les différents packages de R.

Rappelons enfin, qu'il exsiste une autre extension de l'analyse des correspondances qui est l'analyse des correspondances multiples notée ACM. Elle n'est pas une nouvelle méthode mais une application particulière de l'AFC à des tableaux à plusieurs variables qualitatives.

Bibliographie

[1] Alain, B. (2010). Statistique Descriptive Multidimensionnelle, L'Institut de Mathématiques de Toulous.

[2] Baey, C. (2019). Analyse de donnée, https ://baeyc.github.io/teaching/.

[3] Bendjaballah, Ilhame. (2019). Analyses factorielles des correspondances, Mémoire Master de l'Université de Mohamed Khider Biskra.

[4] Boumaza, R. (2007). *Analyse des données* (Vol. 16). Centre de publication universitaire.

[5] Bry, X. (1995). *Analyse factorielle simple.*

[6] Chavent, M. (2014-2015). Notions de base pour l'analyse d'un tableau de contingence, Université de Bordeaux -MASTER MIMSE-2$^{\text{ème}}$année.

[7] Escofier, B., Pagès, J. (2008). *Analyses factorielles simples et multiples.* Dunod, Paris.

[8] Greenacre, M. (2017). *Correspondence analysis in practice.* CRC Press.

[9] Husson, F., Lê, S., Pagès, J. (2016). *Analyse de données avec R.* Presses universitaires de Rennes.

[10] Kassambara, A. (2017). *Practical guide to principal component methods in R : PCA, M (CA), FAMD, MFA, HCPC, factoextra* (Vol. 2). Sthda.

[11] Necir, A. (2020). Analyse factorielle des correspandances (Modèle linéaire), Cours de 1$^{\text{ère}}$Année Master, Université de Mohamed Khider Biskra.

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

[13] Saporta, G. (2006). *Probabilités Analyse des données et Statistique*, 2ème édition, Edition Technip.

Annexe : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

AFC	:	Analyse factorielle des correspondances.
ACP	:	Analyse en composantes principales.
x_{ij}	:	Effectif observé.
n	:	Effectif total.
$X_{i\cdot}$	:	Effectif marginal des lignes.
$X_{\cdot j}$	:	Effectif marginale des colonnes.
f_{ij}	:	Fréquence observé.
$f_{i\cdot}$	:	Fréquence marginale des lignes.
$f_{\cdot j}$	:	Fréquence marginale des colonnes.
$f_{i/j}$	:	Fréquence conditionnelle aux profils-lignes.
$f_{j/i}$	:	Fréquence conditionnelle aux profils-colonnes.
$\widetilde{f}_{ij}$	:	Fréquence théorique.
χ^2	:	La statistique du Khi-deux.
ϕ^2	:	L'écart à l'indépendance.
D	:	Distribution.

ddl	:	Degré de liberté.
g_r	:	Le centre de gravité de profils-lignes.
g_c	:	Le centre de gravité de profils-colonnes.
$d^2_{\chi^2}$	:	La distance du khi-deux.
I	:	Inertie.
I_T	:	Inertie totale.
CTR	:	Contribution.

ملخص

في هذه المذكرة، نهتم بطريقة التحليل العاملي، بدءاً بتقديم بعض الأساسيات التي نستخدمها في التحليل العاملي. على وجه الخصوص، جدول تقاطع البيانات، التوزيعات الهامشية وقياس مربع كاي. بعد ذلك، نهتم بمبدأ التحليل العاملي باستخدام تحليل المركبات الرئيسية. أخيرا لتوضيح عملنا، أضفنا تطبيقا للطريقة الإحصائية على البيانات الحقيقية الموجودة في برنامج R.

الكلمات المفتاحية: التحليل العاملي، تحليل المركبات الرئيسية، مقياس مربع كاي.

Résumé

Dans ce mémoire, nous sommes intéressés à la méthode d'analyse factorielle des correspondances. Nous avons commencé par une présentation de quelques notions de base que nous utilisons dans l'AFC, à savoir : tableau de contingence, les distributions marginales et la métrique de Khi-deux. Par la suite nous focalisons sur le principe d'AFC en utilisant l'analyse en composantes principales. Enfin, pour illustrer notre travail, nous avons ajouté une application de la méthode sur des données réelles trouvées dans le logiciel R.

Mots clés: Analyse factorielle, analyse en composantes principales, métrique de Khi-deux.

Abstract

In this work, we are interested to the correspondence analysis method. We started by presenting some basics that we use in CA, namely: contingency table, marginal distributions and Khi-square metric. Then we focus on the main procedure of CA by using the principal components analysis. Finally, to illustrate this method, we added an application of the CA method on real data that one find in the R software.

Key words: Correspondence analysis, principal component analysis, Khi-square metric.