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 pour obtenir le diplôme de

MASTER en Mathématiques

Option : **Statistique**

Par

ZIDI Imane

Titre :

Analyse des Correspondances Multiples

Membres du Comité d'Examen :

Pr.	MERAGHNI Djamel	UMKB	Président
Pr.	NECIR Abdelhakim	UMKB	Encadreur
Dr.	BENAMEUR Sana	UMKB	Examinatrice

Juin 2023

Dédicace

Je dédie ce humble travail à

Mon cher père : Abdesselem.

Ma chère mère : Bachir Louiza.

Mes chères sœurs :

Ahlem, Nada, Lina, Dhouha.

Mon cher frère : Ali.

Mon petit prince : Anas.

A tous mes amies : Loubna, Chaima, Widad, Moufida, Samia, Nesserine.

A toutes les personnes qui ont contribués de près ou de loin pour la réalisation de ce travail.

Merci du fond du cœur.

REMERCIEMENTS

Avant tout, je tiens à remercier "**ALLAH**" le Tout-Puissant, qui m'a accordé la santé, le courage, la patience et la volonté nécessaires pour accomplir ce travail.

Je tiens tout d'abord à remercier à mon encadreur le Professeur **NECIR Abdelhakim,** merci aucun mot ne vous rendra justice, si vous ne m'aviez pas soutenu, ce travail n'aurait pas été fait.

Par la suite, j'aimerais remercier tous mes enseignants qui m'enseignent dans tout au long de mon parcours d'études. En autre, je tiens aussi à remercier les membres de jury, le Professeur **MERAGHNI Djamel** et le Docteur **BENAMEUR Sana**, merci d'avoir accepté d'examiner et d'évaluer ce travail.

Je souhaite également remercier, tous nos enseignant (es) du département de Mathématiques à l'Université de Mohamed Khider, qui ont contribués à nos formations pendant les années de Licence et de Master.

Table des matières

Table des figures

Liste des tableaux

Introduction

Dans un contexte de croissance du volume de données collectées dans différents domaines, l'analyse de données a été développée en tant que concept dans le domaine de l'informatique et des sciences des données au début du 1900, avec le travail de statisticiens célèbres tels que **Ronald Fisher** et **Karl Pearson** [1]. Cela implique l'utilisation d'outils statistiques pour comprendre et analyser de grandes quantités de données, ainsi que pour extraire des informations. Parmi les outils d'analyse disponibles, on trouve deux méthodes : l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC). Cette dernière peut être généralisée au cas où il y a plus de deux variables qualitatives. Cette généralisation est appelée l'Analyse des Correspondances Multiples (ACM).

L'Analyse des Correspondances Multiples (Multiple Correspondence Analysis en Anglais) a été développée dans les années 1970, notamment par **J-P BENZECRI** à l'Université Pierre-et-Marie-Currie à Paris [15].

L'Analyse des Correspondances Multiples (ACM) est une méthode statistique utilisée pour analyser et visualiser des tableaux de données à plusieurs variables qualitatives à la fois et pour en extraire des informations significatives. Elle peut être considérée comme une généralisation de l'Analyse Factorielle des Correspondances (AFC) [5], qui permet d'analyser deux variables qualitatives. L'ACM combine deux méthodes, l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC), en utilisant l'ACP pour réduire les données multidimensionnelles en

2 ou 3 dimensions, et en utilisant l'AFC pour étudier les relations entre les variables qualitatives, les modalités et les individus. Cette technique est utilisée dans divers domaines tels que la psychologie, les sciences sociales, l'économie, le marketing, etc.

Les questions usuelles d'analyse factorielle sont :

- Y a-t-il des groupes d'individus qui se ressemblent ? sont-ils différents ?
- Quelles sont les associations entre les modalités ?
- Quelles sont les relations entre les variables ?

Ce travail est divisé en deux parties : théorique et pratique.

Partie théorique : cette partie est composée en deux chapitres. Dans le premier chapitre on va présenter le principe de deux méthodes : l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC). Dans le deuxième chapitre nous allons présenter les données de l'étude et les objectifs de l'ACM. Ainsi, nous allons exposer en détail le principe et les étapes de cette méthode et la relation entre l'ACM et l'AFC dans le cas deux variables qualitatives.

Partie pratique : dans cette partie, nous allons utilisé l'ACM pour analyser deux bases de données en utilisant les logiciels **R** et **SPSS**. Tout d'abord, le langage **R**, en utilisant le jeu de données poison disponibles dans le package **R FactoMineR**, et on essayera d'appliquer l'ACM à ces données en utilisant les deux packages **R FactoMineR** pour l'analyse et **factoextra** pour visualiser les résultats. Ensuite, on essayera d'appliquer l'ACM à une autre application en utilisant le logiciel **SPSS**. Cette application est décrite dans le document [5] à la page 63

Chapitre 1

Préliminaire

Dans ce chapitre nous allons présenter le principe de deux méthodes : l'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances (AFC).

1.1 Analyse en Composantes Principales

1.1.1 Matrice des observations

Les observations des p variables quantitatives sur n individus sont rassemblées dans une matrice individus×variables quantitatives notée Y^*, à n lignes et p colonnes [16],

$$Y^* = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix} \in \mathcal{M}(n \times p),$$

où y_{ij} désigne la valeur de la variable j mesurée sur l'individu i.

1.1.2 Centre de gravité

Définition 1.1.1 [8] *Le centre de gravité g est le vecteur dont la j-ème coordonnée g_j correspond à la moyenne arithmétique $\overline{y}_j$ de la variable j sur les n individus,*

$$g = (g_1, \ldots, g_p)^t = \left(\overline{y}_1, \ldots, \overline{y}_p\right)^t \in \mathcal{M}(p \times 1),$$

où $\overline{y}_j := \frac{1}{n} \sum_{i=1}^{n} y_{ij},\ j = 1, \ldots, p.$

1.1.3 Matrice centrée et matrice centrée-réduite

Il est important de centrer et réduire nos variables pour les rendre directement comparable parce qu'elles diffèrent en termes de moyenne et de variance.

Pour plus de détails voir [16] pages $6 - 7$.

Matrice centrée

C'est une matrice de dimension $n \times p$ notée Y,

$$Y = Y^* - \mathbf{1}_n g^t = \begin{bmatrix} y_{11} - \overline{y}_1 & \cdots & y_{1p} - \overline{y}_p \\ \vdots & \ddots & \vdots \\ y_{n1} - \overline{y}_1 & \cdots & y_{np} - \overline{y}_p \end{bmatrix} \in \mathcal{M}\left(n \times p\right).$$

Matrice centrée-réduite

C'est une matrice de dimension $n \times p$ notée Z,

$$Z = Y D_{1/s} = \begin{bmatrix} \frac{y_{11} - \overline{y}_1}{s_1} & \ldots & \frac{y_{1p} - \overline{y}_p}{s_p} \\ \vdots & \ddots & \vdots \\ \frac{y_{n1} - \overline{y}_1}{s_1} & \ldots & \frac{y_{np} - \overline{y}_p}{s_p} \end{bmatrix} \in \mathcal{M}\left(n \times p\right),$$

tel que :

– s_j désigne l'écart-type de la variable j où $s_j := \left(\frac{1}{n} \sum_{i=1}^{n} \left(y_{ij} - \overline{y}_j \right)^2 \right)^{\frac{1}{2}}$, $j = 1, \ldots, p$.

– $D_{1/s} = (1/s)\mathbf{I}_p$ désigne la matrice diagonale de $\mathbb{R}^p \times \mathbb{R}^p$ où $1/s = (1/s_1, \ldots, 1/s_p)^t$.

1.1.4 Matrice de variance-covariance et matrice de corrélation

Matrice de variance-covariance

[16] C'est une matrice carrée symétrique de dimension p notée V,

$$V = \frac{1}{n} Y^t Y \in \mathcal{M}\left(p \times p\right).$$

Matrice de corrélation

[16] C'est une matrice carrée symétrique de dimension p notée R,

$$R = \frac{1}{n} Z^t Z \in \mathcal{M}\left(p \times p\right).$$

1.1.5 Espace des individus et espace des variables

Un individu est représenté par le vecteur de $\mathbb{R}^p$, noté $\mathbf{e}_i$,

$$\mathbf{e}_i = \left(y_{i1}, \ldots, y_{ip}\right)^t \in \mathbb{R}^p.$$

Une variable est représentée par le vecteur de $\mathbb{R}^n$, notée Y_j,

$$Y_j = \left(y_{1j}, \ldots, y_{nj}\right)^t \in \mathbb{R}^n.$$

1.1.6 Moment d'inertie

[16] Nous mesurons la dispersion des points du nuage par rapport à son centre de gravité où l'inertie est grande, plus les points sont dispersés, et plus l'inertie est petite, plus il y a des points autour du centre de gravité.

Inertie totale du nuage des individus

Définition 1.1.2 [20] *L'inertie du nuage des individus par rapport à son centre de gravité est donnée par :*

$$I_T := \frac{1}{n}\sum_{i=1}^{n} d^2\left(\mathbf{e}_i, g\right) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{e}_i - g\|_M^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{e}_i - g)^t M (\mathbf{e}_i - g),$$

où M est une matrice carrée symétrique de dimension p définie positive.

Remarque 1.1.1 *Nous avons :*

- *Si $M = \mathbf{I}_p$, $I_T = trace(V) = \sum_{j=1}^{p} s_j$.*
- *Si $M = D_{1/s}$, $I_T = trace(R) = p$.*

Pour plus de détails voir [16] page 12.

Inertie par rapport à un sous-espace vectoriel E passant par g

Définition 1.1.3 [8] *L'inertie des individus par rapport à un sous-espace vectoriel* **E** *passant par g est donnée par :*

$$I_{\mathbf{E}} := \frac{1}{n}\sum_{i=1}^{n} d^2\left(\mathbf{e}_i, \mathbf{proj}_{\mathbf{E},i}\right),$$

où $\mathbf{proj}_{\mathbf{E},i}$ est la projection orthogonale de $\mathbf{e}_i$ sur le sous-espace vectoriel **E**.

Décomposition de l'inertie

[2] On note $\mathbf{E}^{\perp}$ l'espace orthogonal à $\mathbf{E}$ dans $\mathbb{R}^p$, et $\mathbf{proj}_{\mathbf{E}^{\perp},i}$ la projection orthogonale de $\mathbf{e}_i$ sur $\mathbf{E}^{\perp}$. Par le théorème de Pythagore, on a

$$d^2\left(\mathbf{e}_i, g\right) = d^2\left(\mathbf{e}_i, \mathbf{proj}_{\mathbf{E},i}\right) + d^2\left(\mathbf{e}_i, \mathbf{proj}_{\mathbf{E}^{\perp},i}\right).$$

On en déduit, c'est le théorème de Huygens, que :

$$I_T = I_{\mathbf{E}} + I_{\mathbf{E}^{\perp}}.$$

Ainsi,[8] dans le cas particulier où le sous-espace vectoriel de dimension $\mathbf{1}$, c'est-a-dire est un axe, $I_{\mathbf{E}^{\perp}}$ mesure l'étirement du nuage selon cet axe. On parle pour $I_{\mathbf{E}^{\perp}}$ les expressions "d'inertie portée par l'axe $\mathbf{E}$" ou bien "d'inertie expliquée par l'axe $\mathbf{E}$".

L'expression matricielle pour l'inertie expliquée par l'axe $\mathbf{E}$ est donnée par :

$$I_{\mathbf{E}^{\perp}} = \langle u, Vu \rangle = u^t V u, \tag{1.1}$$

où u est le vecteur directeur unitaire de l'axe $\mathbf{E}$ et $\|u\| = 1$.

Pour plus de détails voir [16] pages $13 - 14$.

On décompose l'espace $\mathbb{R}^p$ comme la somme de p sous-espaces de dimension $\mathbf{1}$ et orthogonaux entre eux :

$$\mathbb{R}^p = \mathbf{E}_1 \oplus \mathbf{E}_2 \oplus \ldots \oplus \mathbf{E}_p,$$

ainsi,

$$I_T = I_{\mathbf{E}_1^{\perp}} + I_{\mathbf{E}_2^{\perp}} + \ldots + I_{\mathbf{E}_p^{\perp}}.$$

1.1.7 Principes généraux de l'ACP

Le but de l'ACP est de réduire les données initiales dans la dimension p (p>3) à la dimension k ($k \leqslant 3$) en déformant le moins possible la réalité. En d'autres termes, minimiser la distance entre les points du nuage initial et leurs projections dans le sous-espace considéré. C'est-à-dire que nous recherchons le sous-espace $\mathbf{E}$ tel que $I_{\mathbf{E}}$ sont minimal et que ceci soit équivalent à $I_{\mathbf{E}^\perp}$ soit maximal.

Pour plus de détails voir [2] page 14.

Recherche des axes principaux

Le premier axe principal $\mathbf{E}_1$ Notons u_1 le vecteur directeur unitaire de $\mathbf{E}_1$, on cherche donc u_1 tel que $I_{\mathbf{E}_1^\perp}$ soit maximum sous la contrainte $\|u_1\|^2 = 1$, on obtient alors le problème suivant :

$$\begin{cases} \max_{u_1} I_{\mathbf{E}_1^\perp} \\ \|u_1\|^2 = 1 \end{cases} \Longleftrightarrow \begin{cases} \max_{u_1} u_1^t V u_1 \\ \|u_1\|^2 = 1 \end{cases}.$$

Après avoir résolu et simplifié le problème (voir [2] pages : $14-15$), nous trouvons :

$$V u_1 = \lambda_1 u_1.$$

où u_1 est le vecteur propre associé à la valeur propre λ_1 de la matrice V. D'après l'expression (1.1) on a :

$$I_{\mathbf{E}_1^\perp} = u_1^t V u_1 = u_1^t \lambda_1 u_1 = \lambda_1 \|u_1\|^2 = \lambda_1,$$

et donc pour maximiser $I_{\mathbf{E}_1^\perp}$ nous maximiserons λ_1 qui représente la plus grande valeur propre de la matrice V.

Le deuxième axe principal $\mathbf{E}_2$ Notons u_2 le vecteur directeur unitaire de E_2, $\|u_2\|^2 = 1$ et $u_2 \perp u_1$, on obtient le problème suivant :

$$\begin{cases} \max_{u_2} I_{\mathbf{E}_2^\perp} \\ \|u_2\|^2 = 1 \text{ et } u_2 \perp u_1 \end{cases} \iff \begin{cases} \max_{u_2} u_2^t V u_2 \\ \|u_2\|^2 = 1 \text{ et } u_2 \perp u_1 \end{cases}.$$

De la même manière on obtient le résultat suivant :

$$I_{\mathbf{E}_2^\perp} = \lambda_2.$$

Ainsi pour maximiser $I_{\mathbf{E}_2^\perp}$ nous maximiserons λ_2 qui représente la deuxième plus grande valeur propre de V.

La même façon de trouver les axes suivants, dont les vecteurs directeurs unitaires sont tous des vecteurs propres normés associés aux valeurs propres de la matrice V ordonnées par ordre décroissant.

Propriétés 1.1.1 [20] *On a les propriétés suivantes :*

- Les p vecteurs propres de V forme une base orthonormée de $\mathbb{R}^p$.
- Les axes $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_p$ sont appelés axes principaux ou axes factoriels.

1.1.8 Constribution des axes à l'inertie totale

Pourcentage d'inertie expliquée par l'axe j

On peut définir le pourcentage d'inertie expliquée par l'axe j par :

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}.$$

Pourcentage d'inertie expliquée par un sous-espace

De la même façon, le pourcentage d'inertie expliquée par un sous-espace $\mathbf{F}_h = \mathbf{E}_1 \oplus \mathbf{E}_2 \oplus \ldots \oplus \mathbf{E}_h$, est égale à :

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_h}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}.$$

Pour plus de détails voir [16] page 19.

1.2 Analyse Factorielle des Correspondances

Soient deux variables qualitatives (vqs) X_1 et X_2 ont respectivement (resp) m_1, m_2 modalités, décrivant un ensemble de n individus.

1.2.1 Tableau de contingence

Définition 1.2.1 [17] *On croisant les deux vqs X_1 et X_2, on obtient un tableau dit tableau de contingence avec m_1 lignes et m_2 colonnes, noté par N^*,*

$$N^* = \begin{bmatrix} n_{11} & \cdots & n_{1m_2} \\ \vdots & \ddots & \vdots \\ n_{m_1 1} & \cdots & n_{m_1 m_2} \end{bmatrix} \in \mathcal{M}(m_1 \times m_2), \tag{1.2}$$

À l'intersection de la ligne i et de la colonne j, on trouve l'effectif n_{ij}.

L'effectif total n est égal à $\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} n_{ij}$.

Les $n_{i.}$ et les $n_{.j}$ sont respectivement appelés les effectifs marginales des lignes et les effectifs marginales des colonnes,

$$n_{i.} = \sum_{j=1}^{m_2} n_{ij}, \text{ et } n_{.j} = \sum_{i=1}^{m_1} n_{ij}.$$

Tableau des fréquences observées

Définition 1.2.2 [17] *Le tableau des fréquences observées est une matrice notée F, avec m_1 lignes et m_2 colonnes,*

$$F = \frac{1}{n}N^* = \begin{bmatrix} f_{11} & \cdots & f_{1m_2} \\ \vdots & \ddots & \vdots \\ f_{m_1 1} & \cdots & f_{m_1 m_2} \end{bmatrix} \in \mathcal{M}(m_1 \times m_2),$$

tel que : $f_{ij} = \frac{n_{ij}}{n}$.

Les fréquences marginales des lignes (resp, colonnes) $f_{i.}$ (resp, $f_{.j}$) sont données par :

$$f_{i.} = \sum_{j=1}^{m_2} f_{ij} = P(X_1 = i) \text{ et } f_{.j} = \sum_{i=1}^{m_1} f_{ij} = P(X_2 = j),$$

et les fréquences conditionnelles des lignes (resp, colonnes) $f_{j/i}$ (resp, $f_{i/j}$),

$$f_{j/i} := \frac{P\left(X_1 = i, X_2 = j\right)}{P(V_1 = i)} = \frac{f_{ij}}{f_{i.}}.$$
$$f_{i/j} := \frac{P\left(X_1 = i, X_2 = j\right)}{P(X_2 = j)} = \frac{f_{ij}}{f_{.j}}.$$

Remarque 1.2.1 *On remarque que :*

$$\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} f_{ij} = \sum_{i=1}^{m_1} f_{i.} = \sum_{j=1}^{m_2} f_{.j} = 1, \text{ et } \sum_{i=1}^{m_1} f_{i/j} = \sum_{j=1}^{m_2} f_{j/i} = 1.$$

1.2.2 Liaison entre deux variables qualitatives

Pour appliquer l'AFC, nous mesurons la dépendance entre deux variables qualitatives en utilisant un test d'indépendance de khi2. Les hypothèses de ce test sont :

$$\begin{cases} H_0 : & X_1 \text{ et } X_2 \text{ sont indépendantes.} \\ H_1 : & X_1 \text{ et } X_2 \text{ sont dépendantes.} \end{cases}$$

La statistique de khi2

Pour $n > 30$ et $\widetilde{n}_{ij} \geq 5$, on utilise la statistique de khi2, notée χ^2, définie par la formule suivante :

$$\begin{aligned}\chi^2 &= n\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(P\left(X_1=i,X_2=j\right)-P\left(X_1=i\right)P\left(X_2=j\right)\right)^2}{P\left(X_1=i\right)P\left(X_2=j\right)}\\ &= n\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(\mathbf{f}_{ij}-\mathbf{f}_{i.}\mathbf{f}_{.j}\right)^2}{\mathbf{f}_{i.}\mathbf{f}_{.j}} = n\Phi^2,\end{aligned}$$

où :

– $\widetilde{n}_{ij}$ désigne l'effectif théorique, $\widetilde{n}_{ij} = n_{i.} * n_{.j}/n$.

– $\mathbf{f}_{ij}, \mathbf{f}_{i.}, \mathbf{f}_{i.}$: variables aléatoires des fréquences associées à $f_{ij}, f_{.j}, f_{i.}$respectivement.

– Φ^2 : mesure de l'écart à l'indépendance.

La convergence de khi2

[20] La statistique de khi2 (χ^2) converge en loi vers χ^2_r :

$$\chi^2 \xrightarrow{D} \chi^2_r, \text{ quand } n \longrightarrow +\infty,$$

où r désigne le dégré de liberté, $r = (m_1 - 1)(m_2 - 1)$.

La statistique de khi2 observée

La statistique de khi2 observée, notée χ^2_{obs}, est définie comme suit :

$$\chi^2_{obs} = n\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\left(f_{ij}-f_{i.}f_{.j}\right)^2}{f_{i.}f_{.j}} = n\Phi^2_{obs}.$$

Si p-value $< \alpha$: On rejette H_0. Cela signifie qu'il existe une dépendance entre X_1 et X_2, donc on peut appliquer l'AFC.

Pour plus de détails sur le test d'indépendance de khi2, voir [17] pages $4-5$.

1.2.3 Nuage des profils

[17] On définit respectivement les matrices diagonales des profils-lignes et profils-colonnes comme suit :

$$D_L = \begin{bmatrix} f_{1.} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_{m_1.} \end{bmatrix} \in \mathcal{M}(m_1 \times m_1),$$

et,

$$D_C = \begin{bmatrix} f_{.1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_{.m_2} \end{bmatrix} \in \mathcal{M}(m_2 \times m_2).$$

Nuage des profils-lignes

Les profils-lignes forment un nuage de m_1 points dans $\mathbb{R}^{m_2}$ avec des poids donnés par la matrice D_L.

Définition 1.2.3 [17] *On appelle le tableau des profils-lignes, la matrice de m_1 lignes et m_2 colonnes, notée P_L,*

$$P_L = D_L^{-1} F \in \mathcal{M}(m_1 \times m_2).$$

Le centre de gravité des profils-lignes est donné par :

$$g_L = (f_{.1}, \ldots, f_{.m_2})^t = P_L^t D_L \mathbf{1}_{m_1} \in \mathcal{M}(m_2 \times 1),$$

tel que $\mathbf{1}_{m_1} = (1, \ldots, 1)^t$, est le vecteur unitaire de dimension $m_1 \times 1$.

Nuage des profils-colonnes

Les profils-colonnes forment un nuage de m_2 points dans $\mathbb{R}^{m_1}$ avec des poids donnés par la matrice D_C.

Définition 1.2.4 [17] *On appelle le tableau des profils-colonnes, la matrice de m_2 lignes et m_1 colonnes, notée P_C,*

$$P_C = D_C^{-1} F^t \in \mathcal{M}(m_2 \times m_1).$$

Ainsi, le centre gravité des profils-colonnes est donné par :

$$g_C = (f_{1.}, \ldots, f_{m_1.})^t = P_C^t D_C \mathbf{1}_{m_2} \in \mathcal{M}(m_1 \times 1),$$

tel que $\mathbf{1}_{m_2} = (1, \ldots, 1)^t$, est le vecteur unitaire de dimension $m_2 \times 1$.

1.2.4 La métrique du khi2

En raison de différence de poids entre les lignes et les colonnes, en AFC, les résultats obtenus en utilisant la distance euclidienne ne sont pas satisfaisants en général.

Question : Quelle est la bonne métrique pour déterminer la distance entre deux points ?

Réponse : Dans l'AFC, la distance utilisée est la distance du khi2 et la métrique associée à cette distance est appelée la métrique du khi2.

Pour plus de détails sur cette métrique voir [6] page 92.

La distance du khi2 entre deux profils-lignes i et $i^{'}$

[17] La distance entre deux profils-lignes i et $i^{'}$ est définie comme suit :

$$d^2_{\chi^2}(i,i^{'}) = \sum_{j=1}^{m_2} \frac{1}{f_{.j}} \left(f_{j/i} - f_{j/i'}\right)^2 = \left\|i - i^{'}\right\|^2_{\mathbf{M}_L}, \tag{1.3}$$

où M_L est la métrique du khi2 pour les lignes, définie par :

$$M_L = D_C^{-1}.$$

Ainsi, on définit la distance entre un profil-ligne i et son centre de gravité g_L par :

$$d^2_{\chi^2}(i,g_L) = \sum_{j=1}^{m_2} \frac{1}{f_{.j}} \left(f_{j/i} - f_{.j}\right)^2 = \|i - g_L\|^2_{\mathbf{M}_L}. \tag{1.4}$$

La distance du khi2 entre deux profils-colonnes j et $j^{'}$

De la même manière, on définit la distance entre deux profils-colonnes j et $j^{'}$ par :

$$d^2_{\chi^2}(j,j^{'}) = \sum_{i=1}^{m_1} \frac{1}{f_{i.}} \left(f_{i/j} - f_{i/j'}\right)^2 = \left\|j - j^{'}\right\|^2_{\mathbf{M}_C}, \tag{1.5}$$

où M_C est la métrique du khi2 pour les colonnes, définie par :

$$M_C = D_L^{-1}.$$

La distance entre un profil-colonne j et son centre de gravité g_C, définie par :

$$d^2_{\chi^2}(j,g_C) = \sum_{i=1}^{m_1} \frac{1}{f_{i.}} \left(f_{i/j} - f_{i.}\right)^2 = \|j - g_C\|^2_{M_C}. \tag{1.6}$$

Les inerties totales

Les inerties totales des nuages des points profils-lignes et profils-colonnes par rapport aux centres de gravité correspondants sont définies respectivement par :

$$\text{Inertie}\,(P_L/g_L) = \sum_{i=1}^{m_1} f_{i.} d^2_{\chi^2}(i, g_L), \text{ et Inertie}\,(P_C/g_C) = \sum_{j=1}^{m_2} f_{.j} d^2_{\chi^2}(j, g_C). \tag{1.7}$$

Proposition 1.2.1 *L'inertie totale des nuages des points profils-lignes et profils-colonnes est mesurée par l'écart à l'indépendance* Φ^2 *:*

$$Inertie\,(P_L/g_L) = Inertie\,(P_C/g_C) = \Phi^2 = \frac{\chi^2}{n}.$$

La preuve de cette proposition est donnée dans [17] page 9.

1.2.5 ACP des deux nuages de profils

[20] L'objectif de l'AFC est de réduire les dimensions à 2 ou 3 dimensions comme l'ACP. L'AFC peut être considérée comme le résultat d'une double ACP, c'est-à-dire l'ACP des profils-lignes et l'ACP des profils-colonnes.

1. L'ACP des profils-lignes :
   - Matrice des observations : $Y^* = P_L$.
   - Nuage de profils-lignes centré : $Y = X_L = P_L - 1_{m_1} g_L^t$.
   - Matrice de variance-covariance associée à la matrice X_L : $V_L = X_L^t D_L X_L$.
2. L'ACP des profils-colonnes :
   - Matrice des observations : $Y^* = P_C$.
   - Nuage de profils-colones centré : $Y = X_C = P_C - 1_{m_2} g_C^t$.
   - Matrice de variance-covariance associée à la matrice X_C : $V_C = X_C^t D_C X_C$.

Les axes principaux des deux profils

Pour chercher les axes principaux des profils-lignes (resp, des profils-colonnes), nous allons appliquer l'ACP à la matrice $V_L M_L$ (resp, $V_C M_C$), c'est-à-dire nous cherchons les valeurs propres et les vecteurs propres associés à la matrice $V_L M_L$ (resp, $V_C M_C$). Pour plus de détails voir [17] pages $10 - 11$.

Corollaire 1.2.1 [17] *On pouvons réaliser l'ACP en travaillant directement avec la matrice A_L des profils-lignes ou avec la matrice A_C des profils-colonnes, tel que :*

$$A_L = P_L^t P_C^t, \text{ et } A_C = P_C^t P_L^t.$$

Quelques propriétés

1. $V_L M_L$ a les mêmes valeurs propres non nulles que $V_C M_C$.
2. Propriétés de la matrice $V_L M_L$ et g_L :
   - g_L est un vecteur propre de $V_L M_L$ associé à la valeur propre $\lambda = 0$.
   - g_L est un vecteur propre de A_L associé à la valeur propre $\lambda = 1$.
   - $V_L M_L$ a les mêmes valeurs propres que A_L sauf g_L qui a une valeur propre $\lambda = 1$.
   - g_L est M_L-orthonormés au X_L.
3. Propriétés de la matrice $V_C M_C$ et g_C :
   - g_C est un vecteur propre de $V_C M_C$ associé à la valeur propre $\lambda = 0$.
   - g_C est un vecteur propre de A_C associé à la valeur propre $\lambda = 1$.
   - $V_C M_C$ a les mêmes valeurs propres que A_C sauf g_C qui a une valeur propre $\lambda = 1$.
   - g_C est M_C-orthonormés au X_C.

La preuve des ces propriétés est donnée dans [17] pages $11 - 13$.

Relations entre les deux ACP

Formule de transition [2] Si μ_k est le $k^{ème}$ vecteur propre M_L-normé de A_L associé à la valeur propre $\lambda_k \neq 0$, alors :

$$\widetilde{\mu}_k = \frac{1}{\sqrt{\lambda_k}} P_C^t \mu_k,$$

est le $k^{ème}$ vecteur propre M_C-normé de A_C associé à la même valeur propre $\lambda_k \neq 0$. De même, si $\widetilde{\mu}_k$ est le $k^{ème}$ vecteur propre M_C-normé de A_C associé à la valeur propre $\lambda_k \neq 0$, alors :

$$\mu_k = \frac{1}{\sqrt{\lambda_k}} P_L^t \widetilde{\mu}_k,$$

est le $k^{ème}$ vecteur propre M_L-normé de A_L associé à la même valeur propre λ_k.

Proposition 1.2.2 [2] *On a les proposition suivants :*

1. *Les matrices $V_L M_L$ et $V_C M_C$ ont le même rang τ, tel que :*

$$0 < \tau \leq \min(m_1 - 1, m_2 - 1).$$

2. *Le nombre τ est égale au nombre des valeurs propres non nulles.*

3. *En notant λ_k la $k^{ème}$ valeur propre non nulle, on a alors :*

$$I_T = Inertie\,(P_L/g_L) = Inertie\,(P_C/g_C) = \Phi^2 = \sum_{k=1}^{\tau} \lambda_k.$$

Théorème 1.2.1 [2] *Chaque axe principal explique ainsi une partie de l'écart à l'indépendance entre les deux variables.*

1.2.6 Facteurs principaux et composantes principales

[17] Le facteur principal des profils-lignes est défini comme suit :

$$\omega_k = M_L \mu_k, \text{ pour } 1 \leq k \leq m_1,$$

de même, le facteur principal des profils-colonnes est défini comme suit :

$$\widetilde{\omega}_k = M_C \widetilde{\mu}_k, \text{ pour } 1 \leq k \leq m_2.$$

Propriétés des facteurs principaux Les facteurs principaux des profils-lignes (resp, profils-colonnes) sont les vecteurs propres M_L^{-1}-orthonormés (resp, M_C^{-1}-orthonormés) de la matrice $M_L V_L$ (resp, $M_C V_C$).

[17] La composante principale des profils-lignes est définie comme suit :

$$c_k = X_L \omega_k, \text{ pour } 1 \leq k \leq m_1,$$

de même, la composante principale des profils-colonnes est définie comme suit :

$$\widetilde{c}_k = X_C \widetilde{\omega}_k, \text{ pour } 1 \leq k \leq m_2.$$

Remarque 1.2.2 *On résume, le nombre des composantes principales non nulles est égale à* τ.

Propriétés des composantes principales Les composantes principales des profils-lignes (resp, profils-colonnes) peuvent également être définies comme suit :

$$c_k = P_L \omega_k \text{ et } \widetilde{c}_k = P_C \widetilde{\omega}_k, \text{ pour } 1 \leq k \leq \tau.$$

Ainsi, les composantes principales vérifient :

- $E(c_k) = E(\widetilde{c}_k) = 0$ (c_k et $\widetilde{c}_k$ sont centrées).
- $Var(c_k) = Var(\widetilde{c}_k) = \lambda_k$.
- $cov(c_k, c_j) = cov(\widetilde{c}_k, \widetilde{c}_j) = 0,\ j \neq k$.

Pour $1 \leq k \leq \tau$.

Pour plus de détails voir [17].

1.2.7 Les relations quasi-barycentriques

[17] Il est possible de représenter les deux profils sur le même graphe, ce qui permet une interprétation des relations entre eux. Les interprétations sont les suivantes :

1. Les points des nuage de profils-lignes sont au barycentre des points des nuage de profils-colonnes qu'ils possèdent,

$$c_k = \frac{1}{\sqrt{\lambda_k}} P_L \widetilde{c}_k,\ \text{pour } 1 \leq k \leq \tau.$$

2. Les points des nuage de profils-colonnes sont au barycentre des points des nuage de profils-lignes qu'ils possèdent,

$$\widetilde{c}_k = \frac{1}{\sqrt{\lambda_k}} P_C c_k,\ \text{pour } 1 \leq k \leq \tau.$$

Chapitre 2

Analyse des Correspondances Multiples

L'Analyse des Correspondances Multiples (ACM) est une méthode qui généralise l'AFC aux dimensions supérieurs, c'est-à-dire elle permet d'étudier l'association entre au moins deux variables qualitatives (vqs). Dans ce chapitre, nous allons exposer en détail le principe et les étapes de cette méthode d'analyse.

2.1 Les données

2.1.1 Tableau des données

Soient n individus et p vqs $X_1, \ldots, X_P$ ont resp $m_1, \ldots, m_p$ modalités. Chaque individu est décrit par les numéros des modalités des p vqs qu'il appartient [14]. Ces données sont présentées dans un tableau à n lignes et p colonnes, noté E,

$$E = \begin{bmatrix} e_{11} & \cdots & e_{1p} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{np} \end{bmatrix} \in \mathcal{M}(n \times p), \tag{2.1}$$

tel que e_{ij} représente la modalité k de la variable j prend par l'individu i.

2.1.2 Tableau disjonctif complet

Nous allons introduire un tableau appelé tableau disjonctif complet (TDC) qui permet de représenter les données qualitatives utilisé en analyse des données. Dans ce tableau, une variable qualitative à modalité est remplacée par variables binaires (0 ou 1), chacune correspondant à une des modalités. La définition suivante donne plus d'explication.

Définition 2.1.1 [2] *Le tableau disjonctif complet est un tableau à n lignes représentent les individus et m colonnes représentent les modalités des variables (voir le tableau (2.1)). Nous désignons ce tableau par X.*

	X_1	$\cdots$	X_j	$\cdots$	X_p
Individus	1	$\cdots$	1 k m_j	$\cdots$	m
1			$\vdots$		
$\vdots$			$\vdots$		
i	0100	$\cdots$	x_{ik}	$\cdots$	0010
$\vdots$			$\vdots$		
n			$\vdots$		

TAB. 2.1 – Tableau disjonctif complet : X = [X_1, ..., X_p].

Les éléments de ce tableau x_{ik} sont égales à 1 si l'individu i prend la modalité k de la variable j et 0 sinon.

Ce tableau contient p tableaux d'indicatrices et la somme des colonnes de chaque tableau est égale n.

Le nombre total des modalités m est égale à $\sum_{j=1}^{p} m_p$.

Tableau des fréquences F associée à TDC

Le tableau des fréquences F associée à TDC est défini par :

$$F = \frac{1}{np} X \in \mathcal{M}(n \times m).$$

Les éléments de ce tableau f_{ik} sont égales à $\frac{1}{np}$ si l'individu i prend la modalité k de la variable j et 0 sinon.

En reprenant les notations suivantes :

Appellations	Symboles
Effectif marginale des lignes	$n_{i.} = \sum_{k=1}^{m} x_{ik} = p$
Effectif marginale des colonnes	$n_{.k} = \sum_{i=1}^{n} x_{ik}$
Effectif total	$\sum_{i=1}^{n}\sum_{k=1}^{m} n_{ik} = np$
Fréquence marginale des lignes	$f_{i.} = \sum_{k=1}^{m} f_{ik} = \frac{1}{n}$
Fréquence marginale des colonnes	$f_{.k} = \sum_{i=1}^{n} f_{ik} = \frac{n_{.k}}{np}$

Exemple 2.1.1 *Afin d'illustrer l'ACM, nous utilisons l'exemple suivant [2] : soit 4 individus et 2 variables qualitatives X_1, X_2 à respectivement 2 et 3 modalités.*

Le tableau des données individus×variables et le TDC sont les suivants :

$$E = \begin{bmatrix} 1 & 3 \\ 2 & 2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \Rightarrow X = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Le nombre total des modalités est : $m = \sum_{j=1}^{2} m_j = 2 + 3 = 5$.

Nous avons $n = 4$ *et* $p = 2$*, donc l'effectif total est :* $np = 4 \times 2 = 8$.

Ainsi, les effectifs marginales des lignes (resp, des colonnes) :

$$n_{1.} = n_{2.} = n_{3.} = n_{4.} = 2.$$
$$n_{.1} = 3, n_{.2} = 1, n_{.3} = 2, n_{.4} = 1, n_{.5} = 1.$$

Le tableau des fréquences F est :

$$F = \frac{1}{8}X = \begin{bmatrix} \frac{1}{8} & 0 & 0 & 0 & \frac{1}{8} \\ 0 & \frac{1}{8} & 0 & \frac{1}{8} & 0 \\ \frac{1}{8} & 0 & \frac{1}{8} & 0 & 0 \\ \frac{1}{8} & 0 & \frac{1}{8} & 0 & 0 \end{bmatrix}.$$

Les fréquences marginales des lignes (resp, des colonnes) sont :

$$f_{1.} = f_{2.} = f_{3.} = f_{4.} = \frac{1}{4}.$$
$$f_{.1} = \frac{3}{8}, f_{.2} = \frac{1}{8}, f_{.3} = \frac{2}{8}, f_{.4} = \frac{1}{8}, f_{.5} = \frac{1}{8}.$$

2.1.3 Tableau de Burt

Il existe une autre façon de transformer le tableau E (2.1) en un tableau analysable appelé tableau de Burt, noté B avec $B = X^tX$ (voir le tableau (2.2)).

Le tableau de Burt est un tableau carré symétrique de taille $m \times m$, qui comporte tous les tableaux de contingence des variables prises deux à deux. Ce tableau peut être divisé en sous-tableaux. Les sous-tableaux sur la diagonale principale sont carrés, son diagonale représente les effectifs marginales des colonnes $n_{.k}$ et les éléments hors diagonale son nuls. Les sous-tableaux sur l'hors diagonale principale croisent les modalités de la variable j avec les modalités d'une autre variable.

Pour plus de détails voir [14].

		X_1					X_p	
		1	m_1		k			m
X_1	1	$n_{.1}$	0	$\cdots$	b_{1k}	$\cdots$	$\cdots$	b_{1m}
	m_1	0	$\ddots$					$\vdots$
		$\vdots$		$\ddots$		0		$\vdots$
	k	b_{k1}			$n_{.k}$			b_{km}
		$\vdots$		0		$\ddots$		$\vdots$
X_p		$\vdots$					$\ddots$	0
	m	b_{m1}	$\cdots$	$\cdots$	b_{mk}	$\cdots$	0	$n_{.m}$

TAB. 2.2 – Tableau de Burt : $\mathrm{B} = \mathrm{X}^t X$.

Les éléments $b_{kk'}$ sont égales à $\sum_{i=1}^{n} x_{ik}x_{ik'}$. Puisque ce tableau est symétrique, les effectifs marginales des lignes et des colonnes sont égaux et sont donnés par :

$$b_k = \sum_{k=1}^{m} b_{kk'} = pn_{.k},$$

et l'effectif total est donné par :

$$\sum_{k=1}^{m}\sum_{k'=1}^{m} b_{kk'} = p^2 n.$$

Dans ce tableau, aucune information n'est disponible sur les individus, il n'y a que de l'information liée aux relations entre les modalités, de sorte qu'il n'est pas possible d'effectuer une analyse des individus.

Remarque 2.1.1 *Le tableau de Burt peut être obtenu à partir du tableau de TDC, mais l'inverse n'est pas possible.*

Exemple 2.1.2 *Le tableau de Burt pour notre exemple est :*

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} \Rightarrow B = X^t X = \begin{bmatrix} 3 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 2 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

L'effectif total est : $p^2 n = 2^2 \times 4 = 16$, *ainsi les effectifs marginales sont :*

$$b_1 = 6, b_2 = 2, b_3 = 4, b_4 = 2, b_5 = 2.$$

2.2 Les objectifs de l'ACM

À première vue, nous pouvons dire que le problème de l'ACM est similaire au problème de l'ACP (étude d'un tableau individus×variables). D'autre part, il peut être considéré comme une généralisation du problème de l'AFC (étude de la liaison entre plusieurs variables qualitatives). Dans ce cas, nous résumons les objectifs de l'ACM dans trois familles : étude des individus, étude des variables et étude des modalités.

1. Étude des individus : l'ACM permet d'identifier les individus similaires ou différents en l'ensemble de ses réponses à plusieurs variables.
2. Étude des variables : l'ACM permet d'étudier les liens entre les variables qualitatives.
3. Étude des modalités : l'ACM permet d'identifier les similarités ou différences entre les modalités des différentes variables. On peut ainsi identifier les modalités communes et rares.

Pour plus de détails voir [9].

2.3 Principes de l'ACM

L'Analyse des Correspondances Multiples est basée sur l'application de l'AFC d'un TDC, donc en considérant X comme un tableau de contingence N^* (voir [14]), tel que les individus représentent ces lignes et les modalités représentent ces colonnes.

Nous allons énoncer le principe de cette analyse à partir du tableau disjonctif complet.

2.3.1 Nuage des profils

Nuage des profils des individus

En ACM, les individus représentent les lignes du tableau de contingence N^*, donc les profils des individus forment un nuage des n points dans $\mathbb{R}^m$. Les poids des individus sont donnés par la matrice D_L, définie comme suit :

$$D_L = diag\left[(f_{1.}, \ldots, f_{n.})^t\right] \in \mathcal{M}(n \times n),$$

comme $f_{i.} = \frac{1}{n}$ pour $i = 1, \ldots, n$, donc les poids des individus sont égales à $\frac{1}{n}$.

Ainsi, le tableau des profils des individus P_L est donné par :

$$P_L = D_L^{-1}F = \begin{bmatrix} \frac{f_{11}}{f_{1.}} & \ldots & \frac{f_{1m}}{f_{n.}} \\ \vdots & \ddots & \vdots \\ \frac{f_{n1}}{f_{1.}} & \ldots & \frac{f_{nm}}{f_{n.}} \end{bmatrix} = \begin{bmatrix} \frac{x_{11}}{p} & \ldots & \frac{x_{1m}}{p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}}{p} & \ldots & \frac{x_{nm}}{p} \end{bmatrix} \in \mathcal{M}\left(n \times m\right),$$

et le centre de gravité g_L est donné par :

$$g_L = (f_{.1}, \ldots, f_{.m})^t \in \mathcal{M}\left(m \times 1\right).$$

Nuage des profils des modalités

Aussi, en ACM, les modalités sont les colonnes du tableau de contingence N^*, donc les profils des modalités forment un nuage des m points dans $\mathbb{R}^n$. Les poids des individus sont donnés par la matrice D_C, définie comme suit :

$$D_C = diag\left[(f_{.1}, \ldots, f_{.m})^t\right] \in \mathcal{M}(m \times m).$$

comme $f_{.k} = \frac{n_{.k}}{np}$ pour $k = 1, \ldots, m$, donc le poids de la modalité k est égale à $\frac{n_{.k}}{np}$.
En outre, la matrice D_C peut être écrite comme suit :

$$D_C = \begin{bmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & D_p \end{bmatrix} \in \mathcal{M}(m \times m),$$

où chaque bloc D_j, $j = 1, \ldots, p$, représente une matrice diagonale des fréquences marginales des colonnes de la variable j.

Le tableau des profils P_C des modalités est donné par :

$$P_C = D_C^{-1} F^t = \begin{bmatrix} \frac{f_{11}}{f_{.1}} & \ldots & \frac{f_{n1}}{f_{.1}} \\ \vdots & \ddots & \vdots \\ \frac{f_{1m}}{f_{.m}} & \ldots & \frac{f_{nm}}{f_{.m}} \end{bmatrix} = \begin{bmatrix} \frac{x_{11}}{n_{.1}} & \ldots & \frac{x_{n1}}{n_{.1}} \\ \vdots & \ddots & \vdots \\ \frac{x_{1m}}{n_{.m}} & \ldots & \frac{x_{nm}}{n_{.m}} \end{bmatrix} \in \mathcal{M}\left(m \times n\right),$$

et, le centre de gravité g_C est donné par :

$$g_C = (f_{1.}, \ldots, f_{n.})^t \in \mathcal{M}\left(n \times 1\right).$$

Exemple 2.3.1 *Le tableau des fréquences F de l'exemple ci-dessus est :*

$$F = \frac{1}{8}X = \begin{bmatrix} \frac{1}{8} & 0 & 0 & 0 & \frac{1}{8} \\ 0 & \frac{1}{8} & 0 & \frac{1}{8} & 0 \\ \frac{1}{8} & 0 & \frac{1}{8} & 0 & 0 \\ \frac{1}{8} & 0 & \frac{1}{8} & 0 & 0 \end{bmatrix}.$$

Les fréquences marginales des lignes sont :

$$f_{1.} = f_{2.} = f_{3.} = f_{4.} = \frac{1}{4},$$

et les fréquences marginales des colonnes sont :

$$f_{.1} = \frac{3}{8}, f_{.2} = \frac{1}{8}, f_{.3} = \frac{2}{8}, f_{.4} = \frac{1}{8}, f_{.5} = \frac{1}{8}.$$

Les matrices diagonales des profils des individus et des profils des modalités sont :

$$D_L = \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}, \text{ et } D_C = \begin{bmatrix} \frac{3}{8} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{8} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{8} \end{bmatrix},$$

ainsi, la matrice des profils des individus est :

$$P_L = D_L^{-1}F = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix},$$

et la matrice des profils des modalités est :

$$P_C = D_C^{-1} F^t = \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Enfin, les centres de gravité de deux profils sont :

$$g_L = \left(\frac{3}{8}, \frac{1}{8}, \frac{2}{8}, \frac{1}{8}, \frac{1}{8}\right)^t, \text{ et } g_C = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^t.$$

2.3.2 Rappels sur la distance du khi2

Comme en AFC, pour comparer entre deux individus ou deux modalités, on utilise la distance du khi2 (voir (1.2.4)).

La distance du khi2 entre deux individus

Dans $\mathbb{R}^m$, nous rappeler la distance du khi2 entre deux individus (1.3),

$$d^2_{\chi^2}(i, i') = \left\| i - i' \right\|^2_{M_L} = \sum_{k=1}^{m} \frac{1}{f_{.k}} \left(\frac{f_{ik}}{f_{i.}} - \frac{f_{i'k}}{f_{i'.}} \right)^2.$$

La métrique $M_L = D_C^{-1}$.

Selon [15], si les deux individus i et i' choisissent la même modalité k à la variable j, la distance entre eux est nulle ($d^2_{\chi^2}(i, i') = 0$). Cela signifie que, deux individus seront proches s'ils possèdent les mêmes modalités, en particulier s'ils ont en commun des modalités rares.

Proposition 2.3.1 *On a l'égalité suivante :*

$$d^2_{\chi^2}(i,i^{'}) = \frac{n}{p}\sum_{k=1}^{m}\frac{1}{n_{.k}}\left(n_{ik}-n_{i^{'}k}\right)^2.$$

Preuve. Nous allons prouver l'égalité : $d^2_{\chi^2}(i,i^{'}) = \frac{n}{p}\sum_{k=1}^{m}\frac{1}{n_{.k}}\left(n_{ik}-n_{i^{'}k}\right)^2$. On a :

$$\begin{aligned} d^2_{\chi^2}(i,i^{'}) &= \left\|i-i^{'}\right\|^2_{M_L} = \sum_{k=1}^{m}\frac{1}{f_{.k}}\left(\frac{f_{ik}}{f_{i.}}-\frac{f_{i^{'}k}}{f_{i^{'}.}}\right)^2 \\ &= \sum_{k=1}^{m}\frac{np}{n_{.k}}\left(\frac{n_{ik}}{p}-\frac{n_{i^{'}k}}{p}\right)^2 \\ &= \frac{n}{p}\sum_{k=1}^{m}\frac{1}{n_{.k}}\left(n_{ik}-n_{i^{'}k}\right)^2, \end{aligned}$$

ce qu'il fallait démontrer. ■

Ainsi, la distance du khi2 entre un individu et le son centre de gravité g_L (1.4) est :

$$d^2_{\chi^2}(i,g_L) = \|i-g_L\|^2_{M_L} = \sum_{k=1}^{m}\frac{1}{f_{.k}}\left(\frac{f_{ik}}{f_{i.}}-f_{.k}\right)^2.$$

Selon [15], si l'individu i choisit des modalités rares, le point-individu se trouve loin du centre. Cela signifie que les individus qui choisissent des modalités rares tombent dans les extrémités du nuage.

La distance du khi2 entre deux modalités

De la même manière, dans $\mathbb{R}^n$ la distance du khi2 entre deux modalités (1.5) est :

$$d^2_{\chi^2}(k,k^{'}) = \left\|k-k^{'}\right\|^2_{M_C} = \sum_{i=1}^{n}\frac{1}{f_{i.}}\left(\frac{f_{ik}}{f_{.k}}-\frac{f_{ik^{'}}}{f_{.k^{'}}}\right)^2.$$

La métrique $M_C = D_L^{-1}$.

[15] Deux modalités sont proches si elles sont possédées par les mêmes individus.

Proposition 2.3.2 *On a l'égalité suivante :*

$$d^2_{\chi^2}(k,k') = n\sum_{i=1}^{n}\left(\frac{n_{ik}}{n_{.k}} - \frac{n_{ik'}}{n_{.k'}}\right)^2.$$

La distance du khi2 entre une modalité et le son centre de gravité (1.6) g_C est :

$$d^2_{\chi^2}(k,g_C) = \|k - g_C\|^2_{M_C} = \sum_{i=1}^{n}\frac{1}{f_{i.}}\left(\frac{f_{ik}}{f_{.k}} - f_{i.}\right)^2 = \frac{n}{n_{.k}} - 1.$$

Preuve. Nous allons prouver l'égalité : $d^2_{\chi^2}(k,g_C) = \frac{n}{n_{.k}} - 1$. On a :

$$\begin{aligned} d^2_{\chi^2}(k,g_C) = \|k - g_C\|^2_{M_C} &= \sum_{i=1}^{n}\frac{1}{f_{i.}}\left(\frac{f_{ik}}{f_{.k}} - f_{i.}\right)^2 \\ &= n\sum_{i=1}^{n}\left(\frac{n_{ik}}{n_{.k}} - \frac{1}{n}\right)^2 = n\sum_{i=1}^{n}\left[\left(\frac{n_{ik}}{n_{.k}}\right)^2 + \left(\frac{1}{n}\right)^2 - \frac{2}{n} * \frac{n_{ik}}{n_{.k}}\right] \\ &= n\sum_{i=1}^{n}\frac{n_{ik}^2}{n_{.k}^2} + \frac{1}{n} - \frac{2}{n}\sum_{i=1}^{n}\frac{n_{ik}}{n_{.k}}, \end{aligned}$$

comme n_{ik} prennent les valeurs 0 ou 1, alors $n_{ik}^2 = n_{ik}$, et on a $\sum_{i=1}^{n} n_{ik} = n_{.k}$, donc la dernière expression égale à

$$\frac{n}{n_{.k}} + \frac{1}{n} - \frac{2}{n} = \frac{n}{n_{.k}} - 1,$$

ainsi la preuve est terminée. ■

Exemple 2.3.2 *Pour notre exemple, la distance du Khi2 entre le premier et le deuxième individu de la matrice P_L est :*

$$d^2_{\chi^2}(1,2) = \frac{5}{6}(1-0)^2 + \frac{5}{2}(0-1)^2 + \frac{5}{4}(0-0)^2 + \frac{5}{2}(0-1)^2 + \frac{5}{2}(1-0)^2 = \frac{25}{3}.$$

Les Inerties totales

On va rappelé les inerties totales des nuages des points profils des individus (resp, des modalités) par rapport le centre de gravité (1.7) :

$$\text{Inertie}\,(P_L/g_L) = \sum_{i=1}^{n} f_{i.} d^2_{\chi^2}(i, g_L),$$
$$\text{Inertie}\,(P_C/g_C) = \sum_{k=1}^{m} f_{.k} d^2_{\chi^2}(k, g_C).$$

Proposition 2.3.3 *On a les égalités suivantes :*

$$\textit{Inertie}\,(P_L/g_L) = \textit{Inertie}\,(P_C/g_C) = \frac{m}{p} - 1.$$

Selon [2], l'inertie totale en l'ACM ne dépend donc que du nombre moyen de modalités par variable. Elle est dépendante du nombre de variables comme l'ACP et elle n'a pas d'interprétation statistique comme en l'AFC.

Preuve. 1. Nous allons prouver l'égalité : Inertie$(X_L/g_L) = \frac{m}{p} - 1$. On a :

$$\begin{aligned}
\text{Inertie}\,(P_L/g_L) &= \sum_{i=1}^{n} f_{i.} d^2_{\chi^2}(i, g_L) = \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{f_{i.}}{f_{.k}} \left(\frac{f_{ik}}{f_{i.}} - f_{.k}\right)^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{n_{i.}}{n_{.k}} \left(\frac{n_{ik}}{n_{i.}} - \frac{n_{.k}}{np}\right)^2 \\
&= p\sum_{i=1}^{n}\sum_{k=1}^{m} \frac{1}{n_{.k}} \left[\left(\frac{n_{ik}}{p}\right)^2 + \left(\frac{n_{.k}}{np}\right)^2 - 2\frac{n_{ik}n_{.k}}{np^2}\right] \\
&= \frac{1}{p}\sum_{i=1}^{n}\sum_{k=1}^{m} \frac{n_{ik}^2}{n_{.k}} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{m} n_{.k} - \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} n_{ik}.
\end{aligned}$$

Comme n_{ik} prennent les valeurs 0 ou 1, alors $n_{ik}^2 = n_{ik}$,

et on a,

$$\sum_{i=1}^{n}\sum_{k=1}^{m} n_{ik} = np,$$

et,

$$\sum_{i=1}^{n} n_{ik} = n_{.k},$$

donc la dernière expression égale à

$$\frac{1}{p}\sum_{k=1}^{m}\frac{n_{.k}}{n_{.k}} + \frac{1}{n^2}\sum_{i=1}^{n} np - \frac{2}{n}np = \frac{m}{p} - 1.$$

2. De la même façon on va prouver Inertie$(P_C/g_C) = \frac{m}{p} - 1$. ■

Exemple 2.3.3 *Pour notre exemple on a :*

$$Inertie\,(P_L/g_L) = \sum_{i=1}^{4} f_{i.} d^2_{\chi^2}(i, g_L) = \frac{5}{2} - 1 = 1.5 = Inertie\,(P_C/g_C)\,.$$

2.3.3 Axes principaux et facteurs

Étant donné que l'ACM est une application de l'AFC sur le TDC, cela signifie la réalisation de deux ACP sur le TDC (l'ACP sur les profils des individus et l'ACP sur les profils des modalités). Rappelons que l'un des deux ACP peut être déduit de l'autre et cela par des formules de transition (1.2.5), donc nous pouvons choisir de commencer par l'ACP des profils des individus et inférer l'autre avec des formules de transition.

D'après le corollaire précédentes (voir (1.2.1)), pour trouver les axes principaux, nous cherchons les valeurs propres et les vecteurs propres associés à la matrice A_L.

Pour plus de détails voir [2] page 32.

Exemple 2.3.4 *Faisons l'ACP des profils des individus pour notre exemple en utilisant le workplace, le calcul matriciel donne :*

Les transposées P_L et P_C sont :

$$P_L^t = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}, P_C^t = \begin{bmatrix} \frac{1}{3} & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

La matrice A_L est égale

$$A_L = P_L^t P_C^t = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{6} & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

Ainsi, les valeurs propres de A_L sont :

$$\lambda^*_{T_1} = \lambda^*_{T_2} = 1, \lambda^*_{T_3} = 0.5, \lambda^*_{T_4} = \lambda^*_{T_5} = 0.$$

D'après les propriétés (1.7), le centre de gravité g_L est un vecteur propre de A_L associé à la valeur propre $\lambda^*_{T_1} = 1$.

$$\begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{6} & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{3}{8} \\ \frac{1}{8} \\ \frac{2}{8} \\ \frac{1}{8} \\ \frac{1}{8} \end{bmatrix} = \begin{bmatrix} \frac{3}{8} \\ \frac{1}{8} \\ \frac{2}{8} \\ \frac{1}{8} \\ \frac{1}{8} \end{bmatrix}.$$

Les vecteurs propres associés aux valeurs propres respectivement sont :

$$v^*_{T_1} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, v^*_{T_2} = \begin{bmatrix} 3 \\ 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}, v^*_{T_3} = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}, v^*_{T_4} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, v^*_{T_5} = \begin{bmatrix} -3 \\ 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}.$$

Ces vecteurs propres sont M_L-orthogonaux, $\left\langle v^*_{T_i}/v^*_{Tj} \right\rangle_{M_L} = 0$, pour $i \neq j$.

On va normaliser les vecteurs précedents par la métrique M_L,

$$v_{T_i} = v^*_{T_i} / \sqrt{\left\| v^*_{T_i} \right\|^2_{M_L}}, i = 1, \ldots, 5.$$

Nous obtenons donc ce qui suit :

$$v_{T_1} = \begin{bmatrix} 0 \\ \frac{1}{4} \\ 0 \\ \frac{1}{4} \\ 0 \end{bmatrix}, v_{T_2} = \begin{bmatrix} \frac{\sqrt{3}}{4} \\ 0 \\ \frac{\sqrt{3}}{6} \\ 0 \\ \frac{\sqrt{3}}{12} \end{bmatrix}, v_{T_3} = \begin{bmatrix} 0 \\ 0 \\ -\frac{\sqrt{3}}{6} \\ 0 \\ \frac{\sqrt{3}}{6} \end{bmatrix}, v_{T_4} = \begin{bmatrix} 0 \\ -\frac{1}{4} \\ 0 \\ \frac{1}{4} \\ 0 \end{bmatrix}, v_{T_5} = \begin{bmatrix} -\frac{\sqrt{3}}{4} \\ 0 \\ \frac{\sqrt{3}}{6} \\ 0 \\ \frac{\sqrt{3}}{12} \end{bmatrix}.$$

Ensuite, les axes principaux des profils-lignes sont

$$E_i = Vect(v_{T_i}), i = 1, \ldots, 5.$$

La matrice $V_L M_L$ a les même valeurs propres que A_L,donc les valeurs propres associés à $V_L M_L$ sont :

$$\lambda_{T_1} = 1 > \lambda_{T_2} = 0.5 > \lambda_{T_3} = \lambda_{T_4} = \lambda_{T_5} = 0,$$

avec les mêmes vecteurs propres $\{v_{T_i},\ \textit{pour } i = 1, \ldots, 5\}$.

On conclus que :

$$\tau = rang\,(A_L) - 1 = 3 - 1 = 2,$$

c'est-à-dire, l'inertie totale est égale à :

$$I_T = Trace\,(A_L) - 1 = \sum_{k=1}^{2} \lambda_k = 1 + 0.5 = 1.5.$$

On calcule les inerties du nuage de points de profils-lignes par rapport aux axes principaux sont égales aux valeurs propres :

$$Inertie\left(P_L/E_1^{\perp}\right) = \lambda_{T_1} = 1,$$

et

$$Inertie\left(P_L/E_2^{\perp}\right) = \lambda_{T_2} = 0.5.$$

Facteurs et composantes principales

Le facteur principal des profils des individus et des profils des modalités est donné par :

$$\psi_{T_k} = M_L v_{T_k}, \text{ pour } 1 \leq k \leq n,$$

et,

$$\widetilde{\psi}_{T_k} = M_C \widetilde{v}_{T_k}, \text{ pour } 1 \leq k \leq m,$$

La composante principale des profils des individus et profils des modalités est donnée par :

$$c_k = X_L \psi_{T_k} = P_L \psi_{T_k}, \text{ pour } 1 \leq k \leq n,$$

et,

$$\widetilde{c}_k = X_C\widetilde{\psi}_{T_k} = P_C\widetilde{\psi}_{T_k}, \text{ pour } 1 \leq k \leq m.$$

Exemple 2.3.5 *Calculons les composantes principales des profils des individus*

$$c_k = P_L\psi_{T_k} = P_L M_L v_{T_k}, k = 1, \ldots, 5.$$

Nous obtenons :

$$c_1 = P_L M_L v_{T_1} = \begin{bmatrix} \frac{2}{3} \\ 0 \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \sqrt{3},$$

et,

$$c_2 = P_L M_L v_{T_2} = \begin{bmatrix} \frac{2}{3} \\ 0 \\ -\frac{1}{3} \\ -\frac{1}{3} \end{bmatrix} \sqrt{3},$$

remarquons, $c_3 = c_4 = c_5 = 0_{\mathbb{R}^4}$, *car les valeurs propres associées sont nulles.*

2.4 L'AFC du tableau de Burt

L'ACM peut être définie comme à l'AFC du tableau de Burt B [11]. En considérant le tableau B comme un tableau de contingence N^*, mais dans ce cas, les modalités représentent ces lignes et ces colonnes.

L'AFC du tableau B est équivalente à l'AFC du TDC et ont les mêmes vecteurs propres mais les valeurs propres associées sont diffèrent. Les valeurs propres de l'AFC du tableau de Burt, noté λ_B est égale au carré des valeurs propres de l'AFC du TDC,

c'est-à-dire :

$$\lambda_B = \lambda_T^2.$$

où λ_T désigne la valeur propre de l'AFC du TDC.

Preuve. Nous présentons d'abord les notations suivantes :

- Tableau des fréquences observées : $F_B = \frac{1}{np^2}B = \frac{1}{np^2}X^tX$.
- Matrice diagonale des profils des modalités associés à B : $D = D_C$.
- Tableaux de deux profils associés à B sont égales à : $P = D_C^{-1}F_B = \frac{1}{np^2}D_C^{-1}X^tX$.

L'AFC du tableau B revient à chercher les valeurs propres et les vecteurs propres du matrice du produit de deux profils associés à B, notée A

$$A = P^tP^t = \left(\frac{1}{np^2}\right)^2 X^tXD_C^{-1}X^tXD_C^{-1}.$$

D'autre part, l'AFC du tableau X revient à chercher les valeurs propres et les vecteurs propres du matrice A_C, après simplification (voir [14] page 115), on obtient :

$$\left(\frac{1}{np^2}\right) X^tXD_C^{-1}v_T = \lambda_T v_T. \tag{2.2}$$

En prémultipliant les deux membres de (2.2) par $\left(\frac{1}{np^2}\right)X^tXD_C^{-1}$ et après simplification, on obtient :

$$Av_T = \lambda_T^2 v_T = \lambda_B v_T.$$

Donc l'AFC du tableau de Burt a les même vecteurs propres de l'AFC du TDC mais les valeurs propres sont égales à λ_T^2. ■

2.5 Cas de deux variables

Dans le cas de deux variables qualitatives X_1 et X_2 à m_1,m_2 modalités, l'AFC du TDC et l'AFC du tableau de contingence N^* représentent des méthode équivalentes pour analyser les relations entre les variables, et on a l'AFC du tableau de Burt est équivalent l'AFC du TDC dans le cas des plusieures variables et cet équivalent permet aussi d'étudier des relations entre les modalités et les variables.

Equivalences analyses entre les 3 tableaux dans le cas de deux variables qualitatives

Il y a donc un équivalent analyses entre les 3 tableau dans le cas de deux variables :

- L'AFC du tableau de contingence $N^* = X_1^t X_2 \in \mathcal{M}(m_1 \times m_2)$.
- L'AFC du TDC $X = [X_1, X_2] \in \mathcal{M}(n \times m)$, où $m = m_1 + m_2$.
- L'AFC du tableau de Burt $B = X^t X \in \mathcal{M}(m \times m)$.

On résumons ces équivalences analyses dans le tableau (2.3).

La preuve de ces équivalences est donnée dans [14] pages $127 - 129$.

Tableau analysé	Valeur propre	Vecteur propre	Facteur
$N^* = X_1^t X_2$	λ	$\mu \in \mathbb{R}^{m_1}$ $\widetilde{\mu} \in \mathbb{R}^{m_2}$	$\omega \in \mathbb{R}^{m_1}$ $\widetilde{\omega} \in \mathbb{R}^{m_2}$
$X = [X_1, X_2]$	$\lambda_T = \frac{1 \pm \sqrt{\lambda}}{2}$	$v_T = \begin{pmatrix} \mu \\ \pm\widetilde{\mu} \end{pmatrix} \in \mathbb{R}^m$	$\psi_T = \begin{pmatrix} \omega \\ \pm\widetilde{\omega} \end{pmatrix} \in \mathbb{R}^m$
$B = X^t X$	$\lambda_B = \lambda_T^2$	$v_B = v_T$	$\psi_B = \psi_T \sqrt{\lambda_T}$

TAB. 2.3 – Equivalences analyses entre les 3 tableaux dans le cas de deux variables qualitatives.

2.6 Propriétés des valeurs propres

2.6.1 Choix du nombre d'axes

Dans l'ACM en raison de la nature des données du TDC, il est difficile de concentrer l'inertie dans les premiers facteurs, mais certaines valeurs propres peuvent être éliminer en prenant les valeurs propres supérieures à la moyenne qui est égale à $1/p$.

Les pourcentages d'inertie expliquées par les axes sont petits, c'est-à-dire cette critère n'est pas utilisé pour déterminer le nombre d'axes à retenir.

Pour plus de détails voir [19] page 312.

2.6.2 Corrections des valeurs propres

Correction de Benzécri

[2]Benzecri (1979) [3] a proposé la correction des valeurs propres et vise à améliorer les interprétations de l'ACM sans modifier les vecteurs propres. Tout d'abord, on va sélectionner les l valeurs propres supérieures à $1/p$. Ensuite, la correction des valeurs propres est obtenue en appliquant la formule suivante :

$$\widetilde{\lambda}_k = \left[\left(\frac{p}{p-1}\right)\left(\lambda_{T_k} - \frac{1}{p}\right)\right]^2,$$

où $\widetilde{\lambda}_k$ représent les valeurs propres corrigées pour $k = 1, \ldots, l$.

Ainsi, le pourcentage d'inertie expliquée corrigée par l'axe k est :

$$\frac{\widetilde{\lambda}_k}{\widetilde{\lambda}_1 + \widetilde{\lambda}_2 + \ldots + \widetilde{\lambda}_l}.$$

Correction de Greenacre

[2]Greenacre (1993) [12] a proposé une correction supplémentaire pour le calcul de l'inertie expliquée par chaque axe. Quand la correction de Benzécri propose de calculer le pourcentage d'inertie d'un axe k par $\widetilde{\lambda}_k / \sum_{k=1}^{l} \widetilde{\lambda}_k$ Greenacre propose d'utiliser :

$$\frac{\widetilde{\lambda}_k}{I_G}, \text{ ou } I_G = \left[\left(\frac{p}{p-1}\right)\left(\sum_{k=1}^{l} \widetilde{\lambda}_k^2 - \frac{m-p}{p^2}\right)\right]^2.$$

2.7 Aides à l'interprétation

2.7.1 Contributions relatives des individus et des modalités

La contribution relative (CTR) d'un individu i à l'axe factoriel l est donnée par :

$$CTR_l(i) = f_{i.}\frac{c_l(i)^2}{\lambda_{T_l}} = \frac{c_l(i)^2}{n\lambda_{T_l}},$$

où $c_l(i)$ est la coordonnée de l'individu i sur l'axe factoriel l

La contribution relative (CTR) d'une modalité k à l'axe factoriel l est donnée par :

$$CTR_l(k) = f_{.k}\frac{\widetilde{c}_l(k)^2}{\lambda_{T_l}},$$

où $\widetilde{c}_l(k)$ est la coordonnées de la modalité k sur l'axe factoriel l.

En ACM, la contribution d'une variable j à l'axe l est égale à la somme des contributions des modalités de cette variable est donnée par :

$$CTR_l(j) = \sum_{k=1}^{m_j} CTR_l(k)$$

Pour plus de détails voir [5].

2.7.2 Rapport de corrélation

[2] Le rapport de corrélation (RC) entre la variable j et l'axe factoriel l est donné par :

$$RCl(j) = \frac{1}{\lambda_{T_l}} \sum_{k=1}^{m_j} \frac{n_{.k}}{p}.$$

Lorsque ce ratio est proche de 1, il y a une liaison forte entre la variable j et l'axe factoriel l.

2.7.3 Qualité de représentation

[5] La qualité de représentation (QLT) d'un individu i à l'axe factoriel l est donnée par :

$$QLT(i) = \frac{c_l\,(i)^2}{\parallel i - g_L \parallel^2_{M_L}}.$$

[5] La qualité de représentation (QLT) d'une modalité k à l'axe factoriel l est donnée par :

$$QLT(k) = \frac{\widetilde{c}_l\,(k)^2}{\parallel k - g_C \parallel^2_{M_C}}.$$

Chapitre 3

Applications

Nous allons utilisé l'ACM pour analyser deux bases de données, et ce en utilisant les logiciels **R** et **SPSS**. Notre approche est s'inspirée des deux documents [13] et [18].

On note que langage **R** peut être téléchargé gratuitement, tandis que le logiciel **SPSS** n'est pas gratuit (pour plus de détails voir (3.3) et (3.4)).

3.1 L'ACM avec R

3.1.1 Différents packages R

Plusieurs packages sont disponibles dans le logiciel **R** pour le calcul de l'ACM (voir le tableau (3.1)).

Nous utiliserons les deux packages **FactoMineR** pour l'analyse et **factoextra** pour extraire et visualiser les résultats de l'ACM, ce dernier est basé sur le package **ggplot2**.

Voir les fonctions du package **factoextra** dans le tableau (3.5).

Packages	Le code R	Packages	Le code R
FactoMineR	MCA()	ade4	dudi.mca()
Factoshiny	Factoshiny()	MASS	mca()
FactoInvestigate	Investigate()	ExPosition	epMCA()

TAB. 3.1 – Packages R pour le calcul de l'ACM.

Installation et chargement les deux packages

Nous installons les deux packages **R** par la commande install.packages().

```
>install.packages(c("FactoMineR", "factoextra"))
```

Nous chargons les deux packages **R** par la commande library().

```
>library("FactoMineR")
>library("factoextra")
```

3.1.2 Données de l'étude

Nous utiliserons le jeu de données poison disponibles dans le package **FactoMineR**.

```
>data(poison)
```

Ces données proviennent d'une enquête menée auprès d'enfants de l'école primaire qui ont subi des intoxications alimentaires. Ils ont été interrogés sur leurs symptômes et sur ce qu'ils ont mangé.

Description

Ce tableau des données est de type "individus×variables qualitatives" et comporte 55 lignes et 15 colonnes. Les première et deuxième variables sont des variables quantitative supplémentaire (age et time), les troisième et quatrième variables (sick et sex) sont des variables qualitatives supplémentaires et les 11 dernières variables (nausea, vomiting, abdominals, fever, diarrhae, potato, fish, mayo, courgette, cheese et icecream) sont des variables actives.

Tableau disjonctif complet

La fonction tab.disjonctif() du package **FactoMineR** permet de créer le tableau disjonctif complet.

```
>TDC <- tab.disjonctif(poison)
>dim(TDC)
[1] 55 28
```

Ce tableau comporte 55 lignes représentant les individus et 28 colonnes représentant les modalités des variables.

3.1.3 Implémentation de l'ACM

Le code R pour implémenter l'ACM est la commande MCA() [**FactoMineR**].

```
>res.mca <- MCA(poison, ncp=2, quanti.sup=1 :2, quali.sup=3 :4, graph = FALSE)
```

- ncp : nombre de dimensions conservées dans les résultats finaux.
- graph : valeur logique. Si TRUE le graphique est affiché.

Le résultat de l'ACM est une liste stocké dans "res.mca". Cette liste contient les valeurs propres, les pourcentages d'inerties associés à chaque dimension, les coordonnées des individus et des modalités, la qualité de représentation et les contributions des individus et des modalités.

Par exemple : pour obtenir les valeurs propres il suffit de taper la commande :

```
>res.mca$eig
```

Résumer les résultats de l'ACM

Pour résumer les résultats de cette analyse, nous utilisons la commande summary().

```
>summary.MCA(res.mca)
```

Les résultats des valeurs propres (6 valeurs)			
	Eigenvalues	% of var	Cumulative % of var
Dim.1	0.335	33.523	33.523
Dim.2	0.129	12.914	46.437
Dim.3	0.107	10.735	57.172
Dim.4	0.096	9.588	66.760
Dim.5	0.079	7.883	74.643
Dim.6	0.071	7.109	81.752

TAB. 3.2 – Les valeurs propres et les pourcentages d'inerties expliquées.

Les résultats des individus (10 individus)						
	Dim.1			Dim.2		
ind	coord	ctr	cos2	coord	ctr	cos2
1	−0.453	1.111	0.347	−0.264	0.982	0.118
2	0.836	3.792	0.556	−0.032	0.014	0.001
3	−0.448	1.089	0.548	0.135	0.258	0.050
4	0.880	4.204	0.748	−0.085	0.103	0.007
5	−0.448	1.089	0.548	0.135	0.258	0.050
6	−0.359	0.701	0.025	−0.436	2.677	0.037
7	−0.448	1.089	0.548	0.135	0.258	0.050
8	−0.641	2.226	0.615	−0.005	0.000	0.000
9	−0.453	1.111	0.347	−0.264	0.982	0.118
10	−0.141	0.107	0.039	0.122	0.209	0.029

TAB. 3.3 – Les coordonnées, les contributions et le cosinus carré pour 10 individus.

Les résultats des modalités (10modalités)						
	Dim.1			Dim.2		
Modalités	coord	ctr	cos2	coord	ctr	cos2
Nausea_n	0.267	1.516	0.256	0.121	0.811	0.053
Nausea_y	−0.958	5.432	0.256	−3.720	2.906	0.053
Vomit_n	0.479	3.734	0.344	−0.409	7.072	0.251
Vomit_y	−0.719	5.601	0.344	0.614	10.608	0.251
Abdo_n	1.318	15.418	0.845	−0.036	0.029	0.001
Abdo_y	−0.641	7.500	0.845	0.017	0.014	0.001
Fever_n	1.172	13.541	0.785	−0.175	0.783	0.017
Fever_y	−0.670	7.738	0.785	0.100	0.447	0.017
Diarrhea_n	1.183	13.797	0.799	−0.003	0.000	0.000
Diarrhea_y	−0.676	7.884	0.799	0.002	0.000	0.000

TAB. 3.4 – Les coordonnées, les contributions et le cosinus carré pour 10 modalités.

3.1.4 Visualisation et interprétation des résultats

Fonctions	Description
get_eigenvalue()	Extraction des valeurs propres, ou variances des composantes principales.
fviz_eig() / fviz_screeplot()	Visualisation des valeurs propres.
get_mca_ind()	Extraction des résultats pour les individus.
get_mca_var()	Extraction des résultats pour les variables.
fviz_mca_ind()	Visualisation des résultats des individus.
fviz_mca_var()	Visualisation des résultats des variables.
fviz_mca_biplot()	Création d'un biplot des individus et des variables.

TAB. 3.5 – Les fonctions R dans le package factoextra utilisé dans l'ACM.

Nous utiliserons le package **factoextra** pour aider à interpréter et visualiser l'ACM. Les fonctions **R** fournies dans le package **factoextra** utilisé dans l'ACM permettent une extraction et un affichage faciles des résultats d'analyse **FactoMineR**.

Ces fonctions sont résumées dans le tableau (3.5).

Graphiques des valeurs propres

Les valeurs propres peuvent être extraites en utilisant la fonction get_eigenvalue().

```
>eig.val <- get_eigenvalue(res.mca)
>eig.val
```

Le résultat renvoie une liste contient les valeurs propres et les pourcentages d'inerties expliquées associés à 11 dimensions (voir le tableau (3.2)).

Maintenant on va visualiser les valeurs propres et les pourcentages d'inerties expliquées associés à chaque axe de l'ACM en utilisant respectivement les fonctions fviz_eig() et fviz_screeplot(), comme suit :

```
>fviz_eig(res.mca, choice = "eigenvalue")
>fviz_screeplot (res.mca, addlabels = TRUE, ylim = c (0, 12))
```

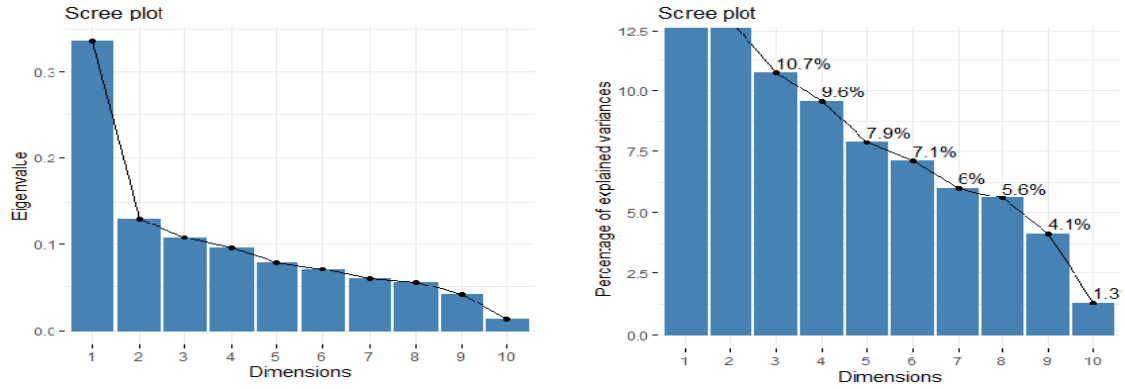

FIG. 3.1 – Données poison : valeurs propres et pourcentages d'inerties expliquées associés à chaque axe.

Interprétation :

La figure (3.1) montre les pourcentages d'inerties expliquée associés à chaque axe.

On a :

– Le 1^{er} axe représente environ 33% d'inertie totale.

– Le $2^{ème}$ axe représente environ 12% d'inertie totale.

Graphique des variables

La fonction get_mca_var() [**factoextra**] renvoie une liste contient les coordonnées, la cos2 et les contributions des modalités (voir le tableau (3.4)).

Maintenant on visualise ces résultats :

1. **Visualisation des modalités sur le premier plan** :

   On utilise la fonction fviz_mca_var() [**factoextra**] comme suit :

   ```
   >fviz_mca_var(res.mca, repel = TRUE)
   ```

   ▶L'option [repel=TRUE] permets d'éviter que les étiquettes ne se superposent.

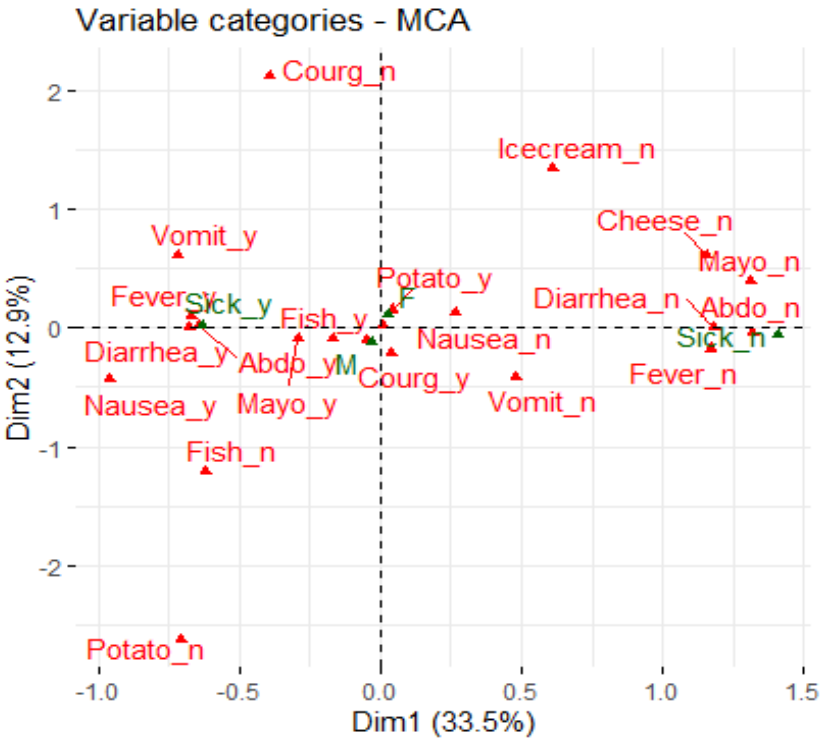

FIG. 3.2 – Données poison : représentation des modalités sur le premier plan.

2. **Visualisation de cos2 des modalités** :

   La qualité de représentation est mésurée par cos2. On utilise la fonction fviz_cos2() [**factoextra**] pour faire un barplot du cos2 des modalités sur le premier plan.

```
>fviz_cos2(res.mca, choice = "var", axes = 1 :2)
```

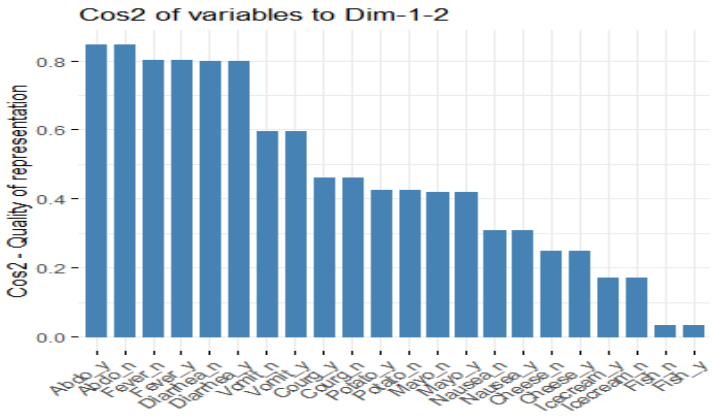

FIG. 3.3 – Données poison : qualité de représentation des modalités sur le premier plan.

3. **Visualisation des contributions des modalités** :

   La fonction fviz_contrib() [**factoextra**] nous permet de créer un barplot de la contribution des modalités sur toutes les dimensions.

```
>fviz_contrib(res.mca, choice = "var", axes = 1 :2, top = 15)
```

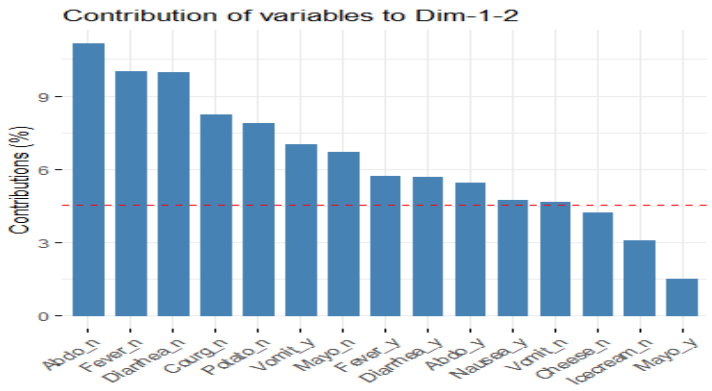

FIG. 3.4 – Données poison : contributions des top 15 modalités sur le premier plan.

4. **Les corrélations entre les variables et les axes (1,2)** :

   Pour visualiser la corrélation entre les varibles et les axes (1,2) de l'ACM, nous écrivons en **R** :

```
>fviz_mca_var(res.mca, choice = "mca.cor")
```

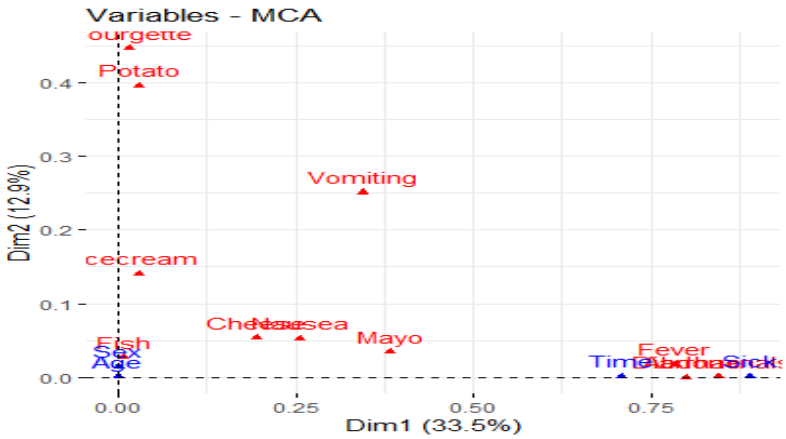

FIG. 3.5 – Données poison : les corrélations entre les variables et les axes (1,2).

Interprétation :

La figure (3.5) permet d'identifier les variables les plus corrélées. On a les variables Diarrhae, Abdominals et Fever sont les plus corrélées avec l'axe 1. De même, les variables Courgette et Potato sont les plus corrélées avec l'axe 2.

Graphique des individus

La fonction get_mca_ind() [**factoextra**] renvoie aussi une liste contient les coordonnées, la cos2 et les contributions des individus (voir le tableau (3.3)).

```
>get_mca_ind(res.mca)
>get_mca_ind
```

De la même manière, nous allons visualise ces résultats :

1. **Visualisation des individus sur le premier plan** :

   Nous écrivons en R la fonction fviz_ma_ind() [**factoextra**] comme suit :

```
>fviz_mca_ind(res.mca, repel = TRUE)
```

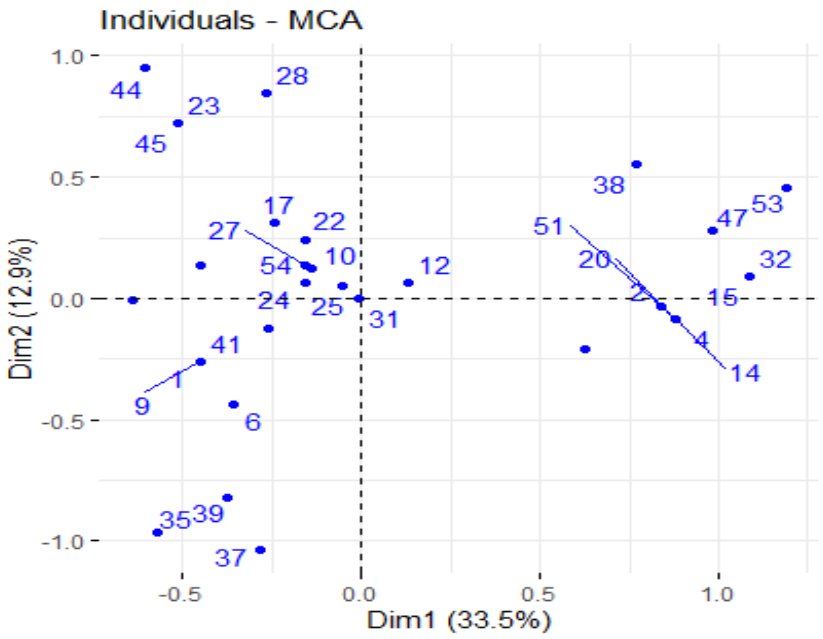

FIG. 3.6 – Données poison : représentation des individus sur le premier plan.

2. **Visualisation de cos2 des individus** :

   On utilise la fonction fviz_cos2() [**factoextra**] comme suit :

```
>fviz_cos2(res.mca, choice = "ind", axes = 1 :2, top = 20)
```

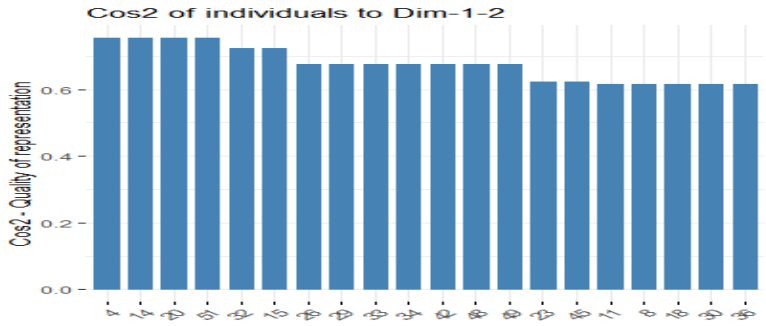

FIG. 3.7 – Données poison : qualité de représentation des top 20 individus sur le premier plan.

3. **Visualisation des contributions des individus** :

On utilise la fonction fviz_contrib() [**factoextra**] comme suit :

```
>fviz_contrib(res.mca, choice = "ind", axes = 1 :2, top = 20)
```

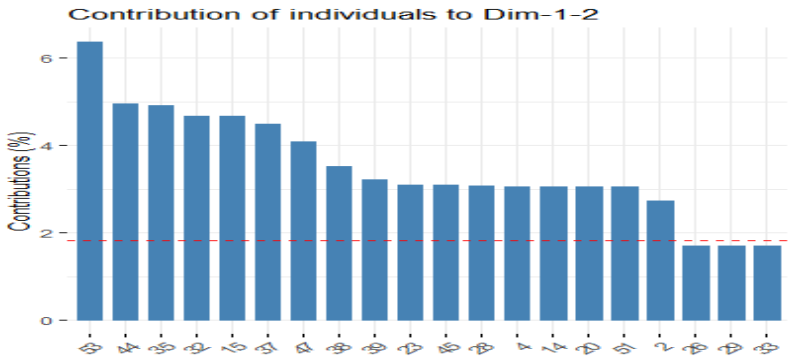

FIG. 3.8 – Données poison : contributions des top 20 individus sur le premier plan.

Visualisation des individus et des modalités sur le premier plan

La fonction fviz_mca_biplot() [**factoextra**] permet de visualiser le biplot des individus et des modalités sur le premier plan,

```
>fviz_mca_biplot (res.mca, repel = TRUE)
```

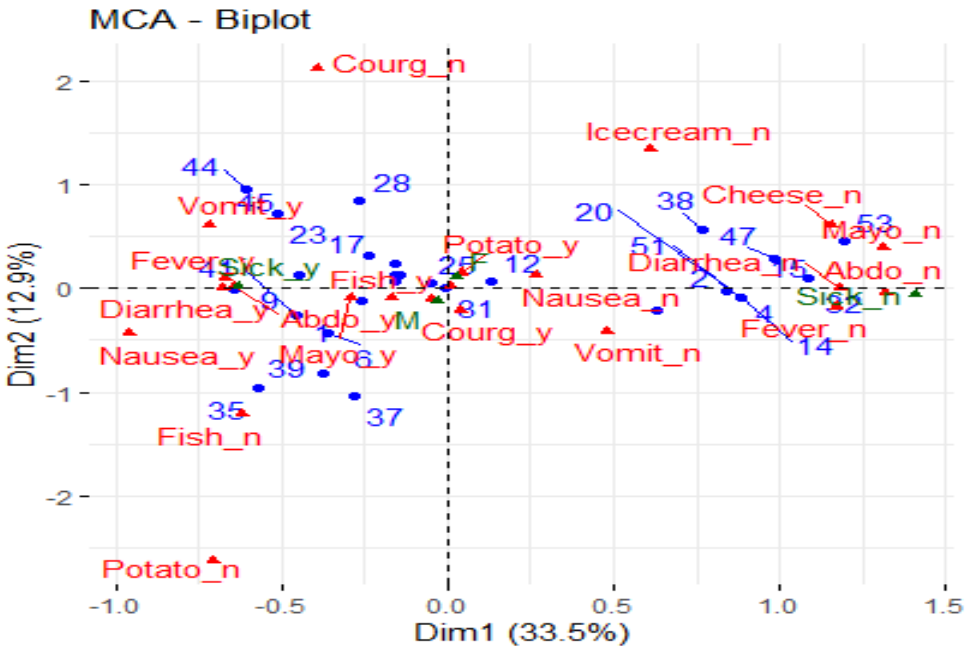

FIG. 3.9 – Données poison : représentation des individus et des modalités sur le premier plan.

Interprétation :

La figure (3.9) montre les deux premiers axes de l'ACM expriment 46.4% du jeu de données poison, cela signifie que de 46.4% de l'inertie totale du nuage des individus (ou des modalités) est représentée dans ce plan. Nous prenons quelques remarques :

- Les individus $4, 2, 14, 20$ et 51 sont similaires et partagent des modalités Diarrhae_n, Abdo_n, Fever_n, Vomit_n, Nausea_n.
- Les individus $27, 54, 24, 22, 17, 25$ et 41 sont similaires et partagent des modalités Diarrhae_y, Abdo_y, Fever_y, Vomit_y, Fish_y, Mayo_y.
- Le groupe des individus $35, 37$ et 39 prend une modalité rare : Potato_n.
- Les modalités Diarrhae_n, Abdo_n et Fever_n on peut les associer ensemble (aussi pour les modalités Diarrhae_y, Abdo_y et Fever_y).
- L'axe 1 oppose les enfants ayant des subi intoxications alimentaires, sur la gauche, contre ceux n'en présentant pas à droit.

3.1.5 Corrections des valeurs propres

En utilisant le logiciel **R** pour calculer la correction de Benzécri et la correction de Greenacre des valeurs propres, nous obtenons les résultats suivants :

Dim	% of var (sans corr)	% of var (Benzécri)	% of var (Greenacre)
1	33.523	91.641	62.472
2	12.914	4.958	3.380
3	10.735	2.102	1.433
4	9.588	1.084	0.739
5	7.883	0.188	0.128
6	7.109	0.024	0

TAB. 3.6 – Correction des valeurs propres (pourcentages d'inerties expliquées).

3.2 L'ACM avec SPSS

3.2.1 Données de l'étude

Ces données se trouvent dans le document [5] à la page 63.

Cette étude porte sur le temps de travail personnel hebdomadaire consacré par 210 étudiants d'une promotion à l'approche d'une session d'examen. Les questions posées sont les suivantes :

Les variables qualitatives	Les modalités
Sexe	1-Masculin, 2-Féminin.
Catégorie socio-professionnelle du père	1-Sans profession ou chômeur, 2-Salarié, 3-Cadre salarié, 4-Profession libérale, 5-Commerçant ou artisan, 6-Agriculteur.
Catégorie socio-professionnelle du mère	1-Sans profession ou chômeur, 2-Salarié, 3-Cadre salarié, 4-Profession libérale, 5-Commerçant ou artisan, 6-Agriculteur.
Êtes-vous membre d'une association sportive, musicale ou autre ?	1-Oui, 2-Non.
Pratiquez-vous souvent des activités de bricolage au lecture non scolaire ou autre ?	1-Oui, 2-Non.
Combien d'heures de travail personnel avez vous consacré à vos études ?	1-Moins de 10h, 2-Entre 10 et 20h, 3-Entre 20 et 30h, 4-Plus de 30h.

3.2.2 Lancement de l'ACM

Nous allons maintnenat lancer l'ACM avec le **SPSS**. Tout d'abord, nous devons sélectionner les options suivantes à partir du menu principal :

>>**Anayse** >> **Réduction des dimensions** >> **Codage optimal**.

La première fenêtre qui s'affiche est le codage optimale, et elle affichera plusieurs

choix, telles que, pour le niveau de codage optimale, nous choisissons le premier choix et pour le nembre d'ensembles de variables, nous choisissons "un ensemble". Aprés cela, nous cliquons sur "**Définir**".

Nous arrivons maintenant à une deuxième fenêtre. Nous allons sélectionner l'ensemble des variables que nous souhaitons analyser en les choisissant toutes, puis nous les ajoutons dans la case "**Variable d'analyse** :". Ensuite, nous spécifions deux dimensions à retenir. Après, nous cliquons sur "**Variable**"et sélectionnons toutes les variables, puis nous les ajoutons dans la case "**Joindre les tracés des catégories**". Enfin, nous cliquons sur "**Poursuivre**" puis sur "**OK**".

3.2.3 Représentation des résultats

Les valeurs proprs

Les résultats des valeurs propres		
	Eigenvalues	% of var
Dim.1	0, 323	32, 345
Dim.2	0, 306	30, 554
Total	0, 629	62, 899

TAB. 3.7 – SPSS : les valeurs propres et les pourcentages d'inerties expliquées.

Interprétation :

Le tableau (3.7) montre les valeurs propres et les pourcentages d'inerties expliquées associés aux deux premiers axes de l'ACM. Il faut dire que le premier plan résume 62.8% de l'information données et on a :

– Le 1^{er} axe représente environ $32,3\%$ d'inertie totale.

– Le $2^{ème}$ axe représente environ $30,5\%$ d'inertie totale.

Représentation des modalités

La figure suivante donne une représentation des modalités sur le premier plan.

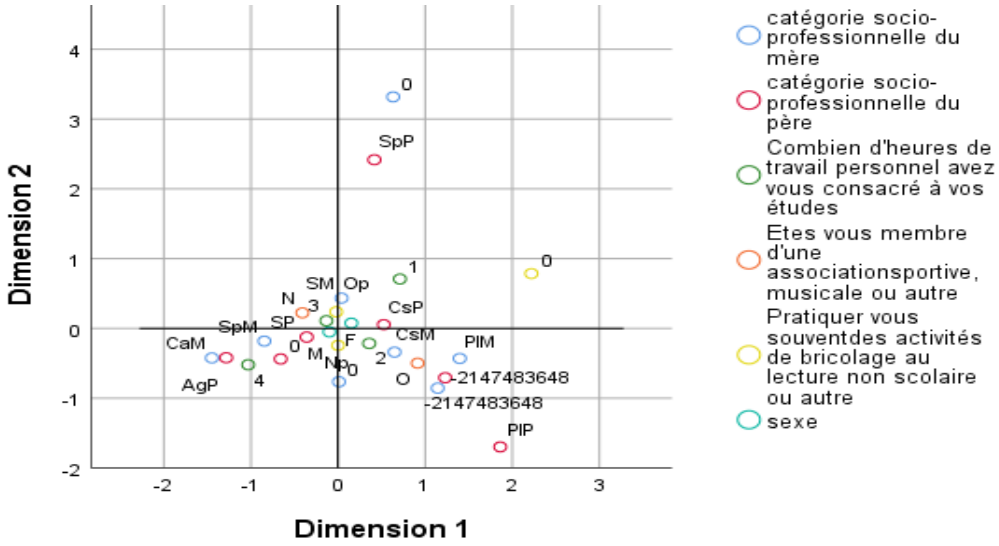

FIG. 3.10 – SPSS : représentation des modalités sur le premier plan.

Représentation des individus

La figure suivante donne une représentation des individus sur le premier plan.

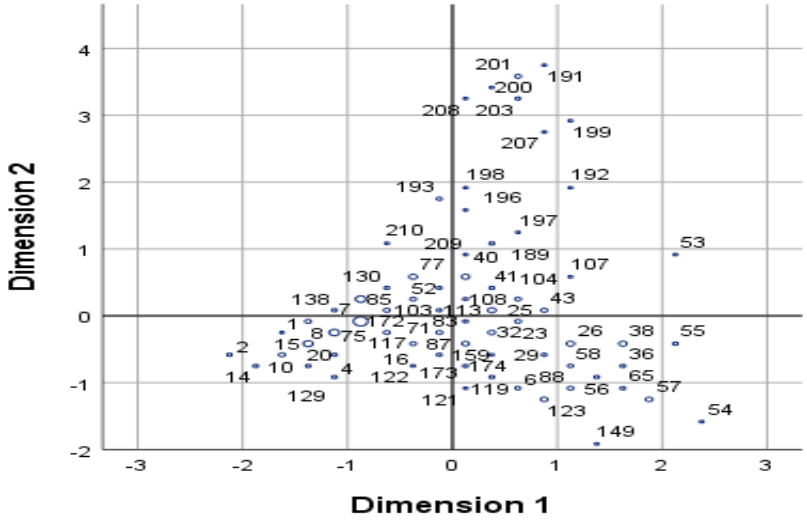

FIG. 3.11 – SPSS : représentation des individus sur le premier plan.

Les corrélations entre les variables et les axes (1,2)

La figure suivante donne les corrélations entre les variables et les axes (1,2).

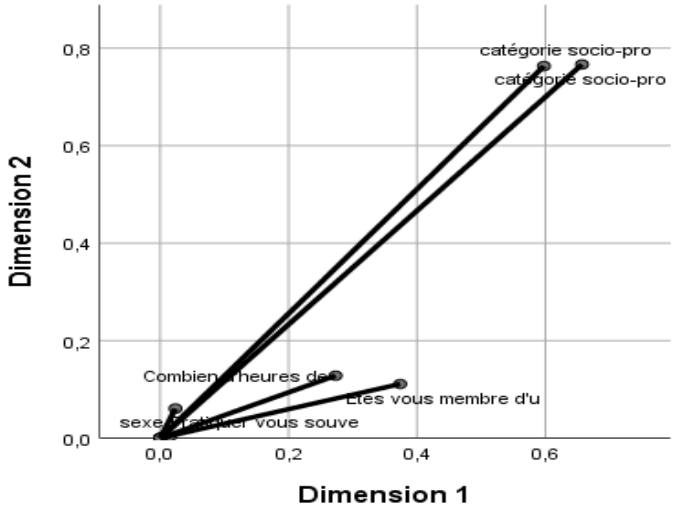

FIG. 3.12 – SPSS : les corrélations entre les variables et les axes (1,2).

Interprétation :

La figure (3.12) montre les variables catégorie socio-professionnelle du père et catégorie socio-professionnelle du mère sont les plus corrélées avec l'axe 2.

Conclusion

L'Analyse des Correspondances Multiples (ACM) est une généralisation de l'Analyse Factorielle des Correspondances (AFC), qui permet d'explorer les relations entre plusieurs variables qualitatives et de les représenter graphiquement dans un espace de dimensions réduit. Cette technique trouve des applications dans de nombreux domaines tels que les sciences sociales, les statistiques et le marketing.

L'ACM permet de visualiser graphiquement les relations entre les modalités des variables et entre les individus, ce qui facilite la compréhension des relations complexes au sein des données. Cette capacité de visualisation offre des avantages précieux pour l'interprétation des résultats et la communication des informations. Dans ce mémoire, nous avons utilisé les logiciels **R** et **SPSS** pour analyser des données réelles, tracer des graphiques et interpréter les résultats de l'ACM.

Cependant, il est important de souligner que l'interprétation des résultats de l'ACM nécessite une compréhension approfondie du contexte des données et des variables étudiées. Il est également essentiel d'être prudent dans l'extraction de conclusions.

Enfin, malgré cela, l'ACM reste une méthode précieuse pour l'analyse exploratoire et la prise de décision éclairée dans de nombreux domaines.

Bibliographie

[1] Ambapour, S. (2003). Introduction à l'analyse des données. Document de travail, Bamsi reprint.

[2] Baey, C. (2019). *Analyse de données*. M2 Ingénierie Statistique et Numérique. Université de Lille.

[3] Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]. *Cahiers de l'Analyse des Données*, 4(3), 377-378.

[4] Blasius, J., Greenacre, M. (Eds.). (2014). *Visualization and verbalization of data*. CRC Press.

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

[6] Bouroche, J.-M., Saporta, G. (Novembre 1992) L'analyse des données (5 éme édition), collection que sais-je ? PUF, Paris.

[7] Chavent, M. (2015). L'Analyse des Correspondances Multiples (ACM). *Université de Bordeaux. Consulté à l'adresse https ://www. math. u-bordeaux. fr/˜machaven/wordpress/wpcontent/uploads/2013/10/ACM. pdf.*

[8] Duby, C., Robin, S. (10 Juillet 2006). Analyse en composantes principales. Institut National Agronomique. Paris - Grignon.

[9] Escofier,B., Pagés,J. (2008). *Analyse factorielle simples et multiples*. Dunod, Paris.

[10] Gibrat, R. (1978). L'analyse des données. *Journal de la société française de statistique*, 119(3), 201-228.

[11] Greenacre, M., Blasius, J. (Eds.). (2006). *Multiple correspondence analysis and related methods*. CRC press.

[12] Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics*, 20(2), 251-269.

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

[14] Lebart, L., Morineau, A., & Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Paris : Dunod.

[15] Le Roux, B., Rouanet, H. (2010). *Multiple correspondence analysis* (Vol. 163). Sage.

[16] Necir, A. (2022.a). Analyse en Composantes Principales (Modèle linéaire), Cours de $1^{ère}$ Année Master, Université de Mohamed Khider Biskra.

[17] Necir, A. (2022.b). Analyse factorielle des correspandances (Modèle linéaire), Cours de $1^{ère}$ Année Master, Université de Mohamed Khider Biskra.

[18] Plaisent, M., Bernard, P. (2008). *Introduction à l'analyse des données de sondage avec SPSS : guide d'auto-apprentissage*. PUQ.

[19] Rakotomalala, R. (2020). Pratique des méthodes factorielles avec python, Université Lumière Lyon2.

[20] Saporta, G. (2006). *Probabilités Analyse des données et Statistique,* $2^{ème}$ édition, Editions technip.

Annexe A : Logiciel R

3.3 Qu'est-ce-que le langage R ?

• Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Logiciel SPSS

3.4 Qu'est-ce-que le SPSS ?

•Le SPSS (Statistical Package for the Social Sciences) est un logiciel statistique utilisé pour l'analyse de données dans les sciences sociales. Il est largement utilisé dans les domaines de la psychologie, de la sociologie, de l'économie et des sciences politiques, entre autres. Le SPSS offre une gamme d'outils pour l'analyse des données, y compris des statistiques descriptives, des tests d'hypothèses, des analyses de variance, des régressions et des analyses factorielles. Il permet également la manipulation de données, le nettoyage, la transformation et la création de variables, ainsi que la visualisation des résultats.

• Le SPSS a été développé par Norman H. Nie, C. Hadlai Hull et Dale H. Bent en 1968 à l'Université de Chicago. Initialement, le logiciel a été conçu pour simplifier l'analyse statistique des données dans les sciences sociales, en particulier pour les chercheurs et les professionnels qui n'avaient pas de formation approfondie en statistiques. Au fil du temps, le SPSS est devenu l'un des logiciels statistiques les plus populaires et les plus largement utilisés dans les sciences sociales et les domaines connexes. En 2009, SPSS Inc. a été acquise par IBM, qui a continué à développer et à soutenir le logiciel en tant que produit phare de sa division Analytics.

Le logiciel est disponible en plusieurs versions, chacune adaptée à des besoins spécifiques, la première version est apparue en 1968 et la dernière version en 2022.

Annexe C : 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.
ACM	:	Analyse des Correspondances Multiples.
TDC	:	Tableau disjonctif complet.
$\overline{y}_j$	:	Moyenne arithmétique de la variable j.
s_j	:	Ecart-type de de la variable j.
$X_1, \ldots, X_P$	:	Variables qualitatives.
$m_1, \ldots, m_p$	:	Modalités.
m	:	Nombre total des modalités.
n	:	Nombre d'individus/ Effectif total.
$n_{i.}$	:	Effectif marginale des lignes.
$n_{.j}$	:	Effectif marginale des colonnes.
$f_{i,j}$	:	Fréquence observée.
$f_{i.}$	:	Fréquence marginale des lignes.
$f_{.j}$	:	Fréquence marginale des colonnes.
$f_{j/i}$	:	Fréquence conditionnelle des lignes.

$f_{i/j}$	:	Fréquence conditionnelle des colonnes.
$\mathbf{f}_{i,j}$, $\mathbf{f}_{i.}$, $\mathbf{f}_{.j}$	:	Variable aléatoire de la fréquance associée à $f_{ij}, f_{.j}, f_{i.}$respectivement.
b_k	:	Effectifs marginales des lignes et des colonnes du tableau de Burt.
H_0	:	Hypothèse nulle.
H_1	:	Hypothèse alternative.
χ^2	:	Statistique de khi2.
χ^2_{obs}	:	Statistique de khi2 observée.
$\mathbf{\Phi}$	:	L'écart à l'indépendance.
$\mathbf{\Phi}_{obs}$	:	L'écart à l'indépendance observé.
g	:	Centre de gravité.
g_L	:	Centre de gravité profils-lignes/ profils des individus.
g_C	:	Centre de gravité profils-colonnes/ profils des modalités.
d^2	:	La distance euclidienne.
$d^2_{\chi^2}$	:	La distance du khi2.
I_T	:	Inertie totale.
I_E	:	Inertie par rapport à un sous-espace vectoriel E.
$I_{E^\perp}$	:	L'inertie expliquée par l'axe E.
Inertie(P_L/g_L)	:	Linertie totale des nuages des points profils-lignes/ profils des individus.
Inertie(P_C/g_C)	:	Linertie totale des nuages des points profils-colonnes/ profils des modalités.
CTR	:	La contribution relative.
QLT	:	Qualité de représentation.
RC	:	Rapport de corrélation.
$E\,(.)$	:	Espérance mathématique.
$Var\,(.)$	:	Variance.
$cov\,(.)$	:	Covariance.
$\langle ., . \rangle$	:	Produit scalaire.
vqs	:	Variables qualitatives.

Résumé

L'Analyse des Correspondances Multiples (ACM) est une méthode d'analyse statistique adaptée aux données qualitatives. Elle permet d'étudier plus de deux variables contrairement à l'Analyse Factorielle des Correspondances (AFC). Dans ce mémoire, nous avons présenté une introduction sur le principe de cette méthode ainsi que les différents packages des langage R et SPSS associés. Pour illustrer notre travail, nous avons appliqué cette méthode aux données réelles.

Mots clés: Analyse des Correspondances Multiples. Analyse Factorielle des Correspondances. Analyse en Composantes Principales. Métrique de Khi-deux. Réduction de dimension. Méthodes d'analyse des données multidimensionnelles.

Abstract

Multiple Correspondence Analysis (MCA) is a statistical analysis method pertaining to qualitative data. It allows the study more than two variables, unlike Correspondence Analysis (CA). In this memory, we presented an introduction to the main principle of this method as well as various packages and syntaxes of R and SPSS software corresponding to these one. To illustrate our work, we applied this method to real data.

Key words: Multiple Correspondence Analysis. Factorial Correspondence Analysis. Principal Component Analysis. Khi-square metric. Dimension reduction. Multidimensional data analysis methods.

ملخص

تحليل الارتباطات المتعددة هو طريقة تحليل إحصائي ملائمة للبيانات النوعية. تسمح بدراسة أكثر من متغيرين، على عكس التحليل العاملي. في هذه المذكرة، قدمنا مقدمة حول مبدأ طريقة تحليل المراسلات المتعددة، بالإضافة إلى الحزم المختلفة في لغة R و SPSS المرتبطة به. ولتوضيح عملنا قمنا بتطبيق هذه الطريقة على بيانات حقيقية.

الكلمات المفتاحية: تحليل الارتباطات المتعددة. التحليل العاملي. تحليل المركبات الرئيسية. مقياس Khi-deux. تقليل الأبعاد. أساليب تحليل البيانات متعددة الأبعاد.