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

BENDJABALLAH Ilhame

Titre :

Analyses factorielles des corespondances

Membres du Comité d'Examen :

Dr.	**BENELMIR Imen**	UMKB	Encadreur
Dr.	**KHEIREDDINE Souraya**	UMKB	Président
Dr.	**DJABER Ibtisem**	UMKB	Examinateur

Juin 2019

DéDICACE

A*u nom du Dieu clément et misécordieux*

J'ai l'immense honneur de dédier ce modeste travail

A *mes parents pour leur amour inestimable, leur confiance,*

leur soutien,leurs sacrifices

M*a famille,*

pour les encouragements tout au long de

mes études et leur amour

A *mes soeurs ainsi qu'à mes beaux frères pour leur tendresse*

leur complicité et leur présence,

Abdelkader Drifa Karima Salim Ibrahime Toufik Khawter

E*t à tous leurs enfants*

A *mes chères amies Hamza Nafla khaoula Hanan et Roufida*

Pour les bons moments passés ensemble

E*nfin, d'un point de vue personnel, je remercie tous mes amis et*

proches qui ont été à mes cotés pendant ces années

études

REMERCIEMENTS

Je tiens tout d'abord à remercier Dieu le tout puissant et miséricordieux

Qui m'a donné la force et la patience d'accomplir

Ce modeste travail.

Je tiens à remercier mon encadreur **Benelmire. Imane**

Pour ses précieux conseils et ses aides durant

Toute la période du travail.

Je n'oublie pas de remercier vivement **Meraghni. Djamel.** pour son soutien,

Ses conseilsjudicieux, et son aide précieuse.

Je tiens à remercier avec ma plus grande gratitude les membres du jury :

Kheireddine Souraya et **Djaber Ebtisem**

Pour l'honneur qu'ils m'ont accordée acceptant

De juger mon travail.

Je n'oublie pas de remercier vivement **Djabrane. Yahya** pour son soutien, ses conseille

Judicieux, et son aide précieuse.

Table des matières

Table des figures

Liste des tableaux

Introduction

La statistique est une science qui vise à collecter et à manipuler l'analyse des données à l'aide de méthodes statistiques, y compris des méthodes statistiques descriptives. Le but de la statistique descriptive est d'étudier la liaison pouvant exister entre les variables, pour structurer l'information contenue dans les données et les représenter simultanément. Les méthodes varient selon la nature des variables étudiées (qualitative, quantitative, ordinale,...ect). Lorsque les variables sont toutes qualitatives, on adopte une mesure différente de toutes celles qui la précèdent qui est l'écart à l'indépendance. C'est ce qu'on va détailler dans le deuxième chapitre. Les deux méthodes les plus courantes de la statistique descriptive multidimensionnelle sont l'Analyse en Composantes Principales (chapitre 1) et l'Analyse Factorielle des Correspondances (chapitre 2).

Le premier chapitre est consacré à l'Analyse en Composantes Principales notée ACP, c'est une méthode fondamentale en statistique descriptive multidimensionnelle, en passant d'abord par les préliminaires. Elle permet de traiter simultanément un nombre quelconque de variables, toutes quantitatives et résumer l'information en un nombre de composantes plus limités que le nombre d'origine des variables.

Dans le deuxième chapitre, on va traiter une méthode d'analyse des données qui est l'analyse factorielle des correspondances notée AFC, appelée aussi analyse des correspondances simples. C'est une méthode exploratoire à l'origine, elle a été conçue pour étudier et analyser des tableaux appelés couramment tableaux de contingence ou tableaux croisés. Elle a été développée essentiellement par **J.-P. Benzecri** durant la période 1970-1990, elle a pour but d'étudier les liaisons dites aussi correspondances existant entre deux variables qualitatives. C'est un outil qui permet de réduire la dimension des données en conservant le plus

d'information possible.

On finalise ce travail, avec une application de la méthode statistique AFC sur des données réelles (Tâches ménagères), pour une meilleure compréhension au fonctionnement de cette dernière.

Chapitre 1

Préliminaire

1.1 Données statistiques essentielles

L'analyse des données est une famille de méthodes statistiques qui se préoccupe de la description de données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différentes données ainsi qu'à en tirer une information par utilisation de mé thodes statistiques comme : ACP, AFC...ect.

Cette partie est consacrée à la description des données et à leur caractéristiques tel que : les individus, les variables, les poids,...ect.

1.1.1 Tableau des données

Les observations de p variables sur n individus sont rassemblées sur un tableau rectangulaire X à n lignes et p colonnes :

$$X = \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ . & & . & & . \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ . & & . & & . \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \in M_{\mathbb{R}}(n,p).$$

On note x_{ij} la valeur de la variable x_j observée sur l'individu e_i.

Exemple 1.1 *Le tableau X ci-dessous représente le résultat observé de trois personnes ou individus $(n = 3)$ selon deux critères (âge, poids) appelés aussi variables $(p = 2)$:*
Comme $x_1 = (25, 30, 35)^t \in \mathbb{R}^3$ et $x_2 = (52, 56, 60)^t \in \mathbb{R}^3$, alors :

$$X = \begin{bmatrix} 25 & 52 \\ 30 & 56 \\ 35 & 60 \end{bmatrix}.$$

Individus-variables

Les lignes du tableau X représentent les individus. Chaque individu est décrit par p variables, formant un vecteur de dimension p appelé vecteur individu$(ADD1 - MAB)$

$$e_i = \begin{bmatrix} x_{i1} & ... & x_{ij} & ... & x_{ip} \end{bmatrix}^t \in \mathbb{R}^p;\ i = 1, ..., n.$$

Tandis que les colonnes du tableau X représentent les variables. Chaque variable peut être représentée par un vecteur de dimension n appelé vecteur variable

$$x_j = \begin{bmatrix} x_{1j} & ... & x_{ij} & ... & x_{nj} \end{bmatrix}^t \in \mathbb{R}^n;\ j = 1, ..., p.$$

1.1.2 Standardisation des données

Avant de commencer à analyser les données, il est utile de transformer le tableau initial X en un tableau standard Z. Ce tableau contient des données centrées réduites de terme général.

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j},$$

où

$\begin{cases} \overline{x}_j : \text{moyenne arithmétique.} \\ s_j : \text{écart-type de la variable } x_j. \end{cases}$

Avec

$$\begin{cases} \overline{x}_j = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij}. \\ s_j^2 = \dfrac{1}{n}\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 . \end{cases}$$

1.1.3 Matrice des poids

Il arrive que les individus n'aient pas la même importance dans une population dans ce cas, chaque individu e_i est doté d'un poids p_i représentant son importance. Ces poids sont regroupés dans une matrice diagonale D de taille n appelée matrice des poids définit comme suit

$$D = \begin{bmatrix} p_1 & & \dots & 0 \\ 0 & p_2 & \dots & \\ & & . & \\ 0 & & \dots & p_n \end{bmatrix},$$

avec $p_i \succ 0$ et $\sum_{i=1}^{n} p_i = 1$.

Dans le cas général, les individus ont la même importance $\dfrac{1}{n}$. Donc la matrice diagonale devient égale à $\dfrac{1}{n} I_n$.

Preuve.

$$\begin{aligned} \sum_{i=1}^{n} p_i &= \sum_{i=1}^{n} p_1 \\ &= p_1 \sum_{i=1}^{n} 1 \\ &= p_1 n \\ &= 1. \end{aligned}$$

Donc

$p_1 = p_i = \dfrac{1}{n}$.

Alors

$$D = \begin{bmatrix} \frac{1}{n} & \dots & \dots & 0 \\ \dots & \ddots & \dots & \dots \\ \dots & \dots & \ddots & \dots \\ 0 & \dots & \dots & \frac{1}{n} \end{bmatrix} = \frac{1}{n} I_n. \quad \blacksquare$$

1.1.4 Centre de gravité

On appelle centre de gravité g du nuage des individus, le point dont les coordonnées sont les valeurs moyennes des variables. Il est donné par

$$\boxed{g = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_p)^t \in \mathbb{R}^p.}$$

Forme matricielle : $g = X^t D 1_n$.

Preuve.

On a

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

$$= \begin{bmatrix} \sum_{i=1}^{n} p_i x_{i1} \\ \sum_{i=1}^{n} p_i x_{i2} \\ \vdots \\ \sum_{i=1}^{n} p_i x_{ip} \end{bmatrix} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} = g. \ \blacksquare$$

1.1.5 Matrice de covariance-corrélation

Définition 1.1 (Matrice de covariance) *La matrice de covariance est une matrice carrée symétrique d'ordre p notée par V.*

$$V = \begin{bmatrix} s_1^2 & & ... & s_{1p} \\ & s_2^2 & ... & \\ & & . & \\ s_{p1} & & ... & s_p^2 \end{bmatrix}.$$

Forme matricielle : $V = Y^t D Y = X^t D X - g g^t$.

Preuve.

Comme on a : $Y = X - Ig^t$, alors

$$\begin{aligned}
V &= Y^t DY \\
&= (X - Ig^t)^t D(X - Ig^t) \\
&= X^t DX - (X^t DI)\, g^t - g\,(I^t DX) + gI^t DIg^t \\
&= X^t DX - gg^t - gg^t + gg^t, \text{ car } I^t DI = \sum_{i=1}^{n} p_i = 1. \\
&= X^t DX - gg^t. \ \blacksquare
\end{aligned}$$

Définition 1.2 (Matrice de corrèlation) *La matrice de corrélation est la matrice regroupant tous les coefficients de corrélation entre les p variables prises deux à deux, on la note par R*

$$R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ . & . & & . \\ . & & \dots & . \\ r_{p1} & & \dots & 1 \end{bmatrix}.$$

Forme matricielle : $R = D_{1/S} S D_{1/S}$.

Preuve.

On a

$$\begin{aligned}
D_{1/S} S D_{1/S} &= \begin{bmatrix} 1/s_1 & & \dots & 0 \\ 0 & 1/s_2 & \dots & \\ & & . & \\ 0 & & \dots & 1/s_p \end{bmatrix} \begin{bmatrix} s_1^2 & & \dots & s_{1p} \\ s_{12} & s_2^2 & \dots & \\ & & . & \\ s_{p1} & & \dots & s_p^2 \end{bmatrix} \begin{bmatrix} 1/s_1 & & \dots & 0 \\ 0 & 1/s_2 & \dots & \\ & & . & \\ 0 & & \dots & 1/s_p \end{bmatrix} \\
&= \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ . & . & & . \\ . & & \dots & . \\ r_{p1} & & \dots & 1 \end{bmatrix} = R. \ \blacksquare
\end{aligned}$$

1.2 Nuage des individus

Chaque individu e_i est un point de l'espace vectoriel $\mathbb{R}^p$ (appelé espace des individus) dont chaque dimension correspond à une variable. L'ensemble des n points représentant les individus, constitue un nuage dans $\mathbb{R}^p$ appelé nuage des individus.

Ressemblance

Comment mesurer la distance entre deux individus ? La distance entre deux individus de l'espace se calcule facilement par la formule de Pythagore (le carré de la distance est la somme des carrés des différences des coordonnées), car les dimensions sont de même nature i.e ce sont des longueurs que l'on mesure avec la même unité.
La distance utilisée dans l'espace est la distance euclidienne classique :

$$\boxed{d^2\left(e_i, e_{i'}\right) = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 .}$$

Remarque 1.1 *Cette définition suppose que les dimensions sont de même nature, i.e que les mesures sont faites dans la même unité.*

Metrique

La métrique M est une matrice carrée symétrique d'ordre p définie positive. La distance entre deux individus e_i et $e_{i'}$ dans l'espace $\mathbb{R}^p$ est donnée par

$$\boxed{d^2\left(e_i, e_{i'}\right) = \left(e_i - e_{i'}\right)^t M \left(e_i - e_{i'}\right) .}$$

Les métriques généralement utilisée en analyse des données pour le nuage des individus sont :

1. $M = D_{1/S^2}$ utilisée pour le tableau initiale X :

$$D_{1/S^2} = \begin{bmatrix} 1/s_1^2 & & ... & 0 \\ 0 & 1/s_2^2 & ... & \\ & & . & \\ 0 & & ... & 1/s_p^2 \end{bmatrix}.$$

2. $M = I_p$ utilisée pour le tableau standards Z.

Inertie

On appelle inertie totale du nuage de points, la moyenne pondérée des carrés des distances des points au centre de gravité :

$$I_g := \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) = \sum_{i=1}^{n} p_i \left\| e_i - g \right\|_M^2 .$$

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

$$I_a = \sum_{i=1}^{n} p_i \left\| e_i - a \right\|_M^2 .$$

On a la relation de Huyghens :

$$I_a = I_g + \|g - a\|^2.$$

Il existe une autre définition de l'inertie :

$$I_g = tr(MV) = tr\left(VM\right).$$

-Si $M = I_p$: $I_g = tr(I_p V) = tr(V) = \sum_{j=1}^{p} S_j^2$.
-Si $M = D_{1/S^2}$: $I_g = tr(D_{1/S^2} V) = tr(D_{1/S} V D_{1/S}) = tr(R) = p$.

1.3 Nuage des variables

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

1.3.1 Liaison entre deux variables

Chaque variable x_j est en fait une liste de n valeurs numériques. Pour étudier l'approximation des variables entre elles, il faut munir cet espace d'une métrique i.e trouver une matrice symétrique d'ordre n définie positive. Sans hésitation le choix se porte sur la matrice diagonale des poids.

Par conséquent ? on a : $\langle x_j, x_{j'} \rangle_D = {x_j}^t D x_{j'} = \sum_{i=1}^{n} p_i x_{ij} x_{ij'}$.

Dans le cas où les variables x_j et $x_{j'}$ sont centrées, le produit scalaire correspond à la covariance i.e $\langle x_j, x_{j'} \rangle_D = s_{jj'}$ et $\|x_j\|_D = s_j$.

L'angle entre deux variables est donné par :

$$\cos(x_j, x_{j'}) = \cos\theta_{jj'} = \frac{\langle x_j, x_{j'} \rangle_D}{\|x_j\|_D \|x_{j'}\|_D} = \frac{s_{jj'}}{s_j s_{j'}}.$$

Sachant que le produit scalaire est défini comme suit :

$$\langle x_j, x_{j'} \rangle_D = \|x_j\|_D \|x_{j'}\|_D \cos(x_j, x_{j'}).$$

Donc

$$r_{ij} = \frac{\langle x_j, x_{j'} \rangle_D}{\|x_j\|_D \|x_{j'}\|_D} = \cos(x_j, x_{j'}) = \cos\theta_{jj'}.$$

On remarque que le coeficient de corrélation linéaire n'est autre que le cosinus de l'angle entre les deux variables x_j et $x_{j'}$.

1.4 Analyse en composantes principales

Il existe plusieurs méthodes adaptées à différents types de données. Parmi eux, on a l'analyse en composantes principales notée ACP. Cette dernière traite des tableaux contenants des individus et des variables quantitatives. Dans le cas oú les valeurs sont qualitatives, on utilise une autre méthode appelée analyse factorielle des correspondances notée AFC.

1.4.1 Principe de l'ACP

L'objectif de l'ACP est de fournir un outil de visualisation des données. Elle permet d'explorer les liaisons entre les variables et la ressemblance entre les individus en réduisant leurs dimensions afin de construire une représentation graphique plus claire dans un espace F_k de $\mathbb{R}^p$ tel que k soit le plus petit possible, i.e on cherche à définir k nouvelles variables qu'on appel les variables initiales contenant le plus d'information possible.

Construction de F_k

Le choix de l'espace de projection s'effectue selon le critère suivant : "déformer le moins possible les distances en les projections". Le sous espace de dimension k recherché est tel que la moyenne des carrés des distances entre projections soit la plus grande possible. En d'autre terme, il faut que l'inertie du nuage projeté sur le sous espace soit maximale.
On peut également déduire de nouvelles variables par combinaison linéaire, ce qui revient à projeter les individus sur de nouveaux axes de F_k.

- On définit P la matrice de projection (operateur) M-orthogonale sur le sous espace F_k tel que :

  1. $P^2 = P$ (P-idempotente).
  2. $P^t M = MP$ (P est M-symetrique).

- On note par f_i la projection de l'individu e_i sur le sous espace F_k.

L'ensemble des individus forme le tableau initial X. Ceci dit l'ensemble des individus projetés

forme le tableau projeté X_{proj}. La relation entre e_i et f_i s'écrit comme suit :

$$f_i = Pe_i.$$

Donc $f_i^t = e_i^t P^t$. Le nuage projeté associé au tableau sera donnée par :

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

- Le centre de gravité projeté est

$$g_{proj} = Pg.$$

Preuve.

On a : $g_{proj} = X_{proj}^t DI$

$= (XP^t)^t DI$

$= P(X^t DI)$

$= Pg.$ ■

- La matrice de variance du tableau projeté est définit par

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

Preuve.

On a : $V_{proj} = X_{proj}^t DX_{proj} - g_{proj} g_{proj}^t$

$= PX^t DXP^t - Pgg^t P^t$

$= P(X^t DX - gg^t)P^t$

$= PVP^t.$ ■

- L'inertie du nuage projeté vaut donc :

$$I_{proj} = tr\left(VMP\right).$$

Preuve.

On a : $I_{proj} = tr\left(V_{proj} MP\right)$

$= tr\left(PVP^t MP\right)$

$= tr\left(PVMPP\right)$, car P est M symétrique

$= tr\left(PVMP\right)$, car P est idempotente

$= tr\left(MPPV\right)$, car $tr\left(AB\right) = tr\left(BA\right)$

$= tr\left(MP^2V\right)$

$= tr\left(MPV\right)$. ■

Projection des individus sur ce nouveau sous espace

Déterminer l'espace de projection F_k revient à trouver l'opérateur de projection M-orthogonal P de rang k, maximisant l'inertie. On construit F_k de proche en proche en cherchant d'abord le sous espace Δ_1 de dimension 1 et d'inertie maximale, puis le sous espace Δ_2 de dimension 1, d'inertie maximale et M-orthogonal à Δ_1, ect... La somme directe de ces sous espaces est F_k :

$$F_k = \Delta_1 \oplus ... \oplus \Delta_k.$$

Remarque 1.2 *En vertu d'un résultat d'algébre linéaire, la M-orthogonalité des droites $\Delta_1, \Delta_2, ...$ est garantie par le fait que la matrice VM est M-symétrique et donc ayant des vecteurs propres M-orthogonaux deux à deux.*

Construction de Δ_1 On détermine la droite Δ_1 passant par le centre de gravité g et maximisant l'inertie du nuage projeté sur elle. Soit $a_1 \in \mathbb{R}^p$ un vecteur directeur unitaire. Le projecteur M-orthogonal sur Δ_1 est :

$$P_1 = a_1 \left(a_1^t M a_1\right)^{-1} a_1^t M = \frac{a_1 a_1^t M}{a_1^t M a_1}.$$

L'inertie du nuage projeté sur Δ_1 est donc

$$\begin{aligned} I_{\Delta_1} &= tr\left(VMP_1\right) \\ &= tr\left(VM\frac{a_1a_1^tM}{a_1^tMa_1}\right) \\ &= \left(1/a_1^tMa_1\right) tr\left(a_1^tMVMa_1\right) \\ &= \left(a_1^tMVMa_1/a_1^tMa_1\right). \end{aligned}$$

On note $(a_1^tMVMa_1/a_1^tMa_1)$ par $f(a_1)$.

$f(a_1)$ est une fonction définie sur $\mathbb{R}^p$ (forme quadratique). Elle atteint son maximum lorsque sa dérivée par rapport à a_1 s'annule. En appliquant la règle de dérivation d'une forme quadratique par rapport à un vecteur on a :

$$VMa_1 = \frac{a_1^tMVMa_1}{a_1^tMa_1}a_1.$$

On pose $\frac{a_1^tMVMa_1}{a_1^tMa_1} = \lambda \in \mathbb{R}$. Alors

$$\boxed{VMa_1 = \lambda a_1.}$$

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

Le sous espace F_k de dimension k est engendré par les k vecteurs propres de la matrice VM associée aux k plus grandes valeurs propres.

Le problème est donc de trouver P projecteur M-orthogonal de rang k maximisant I_{proj} ce qui déterminera F_k.

Eléments principaux

L'ACP repose essentiellement sur les trois éléments suivants :

- Les axes qu'elles déterminent : "axes principaux".
- Les formes linéaires associées : "facteurs principaux".
- Les variables associées : "composantes principales".

Axes principaux On cherche la droite de $\mathbb{R}^p$ passant par g maximisant l'inertie du nuage projeté sur elle. Le vecteur propre a_j de la matrice VM est définit par

$$\boxed{VMa_j = \lambda_j a_j,}$$

où a_j est M-normé à 1 et M-orthogonale.

Facteurs principaux Le facteur principal noté u_j associé à l'axe principal a_j est définit par

$$\boxed{u_j = Ma_j.}$$

où u_j est M^{-1}-normé et de norme $1(u^t M^{-1} u = 1)$ et M^{-1}-orthogonale

On a : $VMa_j = \lambda_j a_j \Leftrightarrow MVMa_j = \lambda_j Ma_j \Leftrightarrow MVu_j = \lambda_j u_j$.

Donc les facteurs principaux u_j sont aussi les vecteurs propres de la matrice MV.

Composantes principales Les composantes principales notées c_j sont les variables définies par les facteurs principaux

$$\boxed{c_j = Xu,}$$

où $c_j \in \mathbb{R}^n$ est le vecteur contenant les coordonnées des projections M-orthogonales des n individus.

La variance d'une composante principale est égale à la valeur propre λ_j. On a

$$var(c_j) = \lambda_j.$$

Chapitre 2

Analyse factorielle des correspondances

L'analyse factorielle des correspondances notée AFC est une méthode factorielle de statistique descriptive multidimensionnelle. L'AFC a été introduit de façon complète dans les années 60 par JP BENZECRI, cette technique statistique permet d'analyser la liaison existante entre deux variables qualitatives (v.qs) par traitement des tableaux de données où les valeurs sont positives et homogènes comme les tableaux de contingence. Le but principal de l'AFC est de lire l'information contenue dans un espace multidimensionnel par une réduction de la dimension de cet espace tout en conservant un maximum d'information contenue dans l'espace de départ.

2.1 Effectifs et fréquences

Soit Ω un ensemble de n individus, sur lequel on considère deux v.qs :

- X à p modalités notées x_i pour $i = 1, ..., p$.
- Y à q modalités notées y_j pour $j = 1, ..., q$.

$$\begin{cases} X : & \Omega \longrightarrow x_i \nsubseteq \mathbb{R} \\ & \omega \longmapsto X(\omega) \end{cases} \quad \text{et} \quad \begin{cases} Y : & \Omega \longrightarrow y_j \nsubseteq \mathbb{R} \\ & \omega \longmapsto Y(\omega) \end{cases}$$

Définition 2.1 ***(Variable qualitative)****En statistique, une v.q est une variable catégorielle (facteur) qui prend pour valeur des modalités (catégories, niveaux), par opposition aux variables quantitatives qui mesurent sur chaque individu une quantité peuvent être : Nominales, Ordinales.*

2.1.1 Effectifs

Comme les valeurs des variables ne sont pas numériques, on s'intéresse aux effectifs et fréquences. On commence d'abord par effectifs :

On a n_{ij} le nombre d'individus possédant à la fois la modalité i de la première v.q et la modalité j de la seconde v.q. On définit alors l'effectif partiel d'un couple de modalité (x_i, y_j) par :

$$n_{ij} := Card\left(\{\omega / X(\omega) = x_i \text{ et } Y(\omega) = y_j\}\right).$$

Le nombre total des individus de la population initiale n est :

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

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement marges en lignes et marges en colonnes, elles sont calculées comme suit :

$$\begin{cases} n_{i.} = \sum_{j=1}^{q} n_{ij}. \\ n_{.j} = \sum_{i=1}^{p} n_{ij}. \end{cases}$$

Tableau de contingence

Un tableau de contingence appelé aussi tableau de dépendance ou tableau croisé est un tableau d'effectifs à p lignes et q colonnes (contient l'intersection de la ligne i et de la colonne j des n_{ij} individus), généralement notée par N.

On obtient les effectifs marginaux en croisant les modalités de deux variables qualitatives X et Y définies sur une même population de n individus.

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

TAB. 2.1 – Tableau de contingence de 2 variable qualitatives X(p modalités)et Y(q modalités)

2.1.2 Fréquences

La fréquence d'une valeur donnée, c'est le quotient (division) de l'effectif de chaque valeur n_{ij} par l'effectif total n. Elle indique la proportion de la présence de la valeur dans la liste. Les fréquences sont définies par :

$$f_{ij} = \frac{n_{ij}}{n} = \frac{effectif\ de\ la\ cellule\ (i,j)}{effectif\ total}. \tag{2.1}$$

On a deux termes générales : $f_{i.}$ pour la marge colonne et $f_{.j}$ pour la marge ligne. Elles sont calculées de la manière suivante :

$$\begin{cases} f_{i.} = \sum_{j=1}^{q} f_{ij} = \frac{n_{i.}}{n}. \\ f_{.j} = \sum_{i=1}^{p} f_{ij} = \frac{n_{.j}}{n}. \end{cases}$$

La fréquence totale :

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

Tableau des fréquences

Le tableau des fréquences qu'on note généralement par F est une matrice à p lignes et q colonnes, obtenu en divisant chaque effectif n_{ij} par l'effectif total n, de terme générale f_{ij}

$X \diagdown Y$	y_1	$\cdots$	y_j	$\cdots$	y_q	$marge\ i$
x_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1q}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{iq}	$f_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_p	f_{p1}	$\cdots$	f_{pj}	$\cdots$	f_{pq}	$f_{p\cdot}$
$marge\ j$	$f_{\cdot 1}$	$\cdots$	$f_{\cdot j}$	$\cdots$	$f_{\cdot q}$	1

TAB. 2.2 – Tableau des fréquences de deux variables

2.2 Indépendance

On a vu que l'AFC contient un tableau de contingence ou de fréquence pour étudier les liaisons entre les deux v.qs à l'initiative du tableau. On ne peut plus définir les liaisons par les coefficients de corrélation comme pour l'ACP.

Définition 2.2 *On dit qu'il ya indépendance entre les deux v.qs lorsque, pour tout i et j, on a l'égalité suivante*

$$f_{ij} = f_{i\cdot} \times f_{\cdot j},$$

où $n_{ij} = \dfrac{n_{i\cdot} \times n_{\cdot j}}{n}$.

Définition 2.3 (fréquences conditionnelles) *Les fréquences conditionnelles sont les nombres suivants :*

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

Remarque 2.1 *Dans l'AFC l'indépendance entre deux v.qs X et Y se traduit par une proportionnalité entre les lignes et les colonnes des tableaux de contingences et de fréquences, i.e qui'il y a une indépendance lorsque tous les pourcentages en colonnes $f_{i/j}$ sont égaux à la marge $f_{i\cdot}$. En d'autre terme, la fréquence conditionnelle $f_{j/i}$ est égale à la fréquence marginale $f_{\cdot j}$.*

L'AFC n'a d'intérêt que s'il ya une dépendance entre les deux variables, dans le cas contraire elle n'apporte pas d'information.

Définition 2.4 *On peut dire que sous H_0 (indépendances des modalités i et j), on aboutit a ces deux conclusions :*

-Si f_{ij} est supérieur au produit des marges, les modalités i et j sont associées sous l'hypothèse d'indépendance. On dit alors que les deux modalités i et j s'attirent.
-Si f_{ij} est inférieur au produit des marges, les modalités i et j sont moins associées sous l'hypothèse d'indépendance. On dit alors qu'il y a répulsion entre les deux modalités i et j.

2.3 Profils

En AFC, le tableau de contingence n'est pas analysé directement. Dans une étude il est transformé en profile. On distingue deux types de profils : les profil-ligne et les profil-colonne. Cette transformation découle de l'objectif qui vise à étudier la liaison entre les deux variables.

2.3.1 Profils-lignes

Un profil-ligne est un vecteur de $\mathbb{R}^q$ dont la $j^{éme}$ coordonnée est égale à la fréquence conditionnelle $f_{j/i}$ notée pl_i. Le $i^{éme}$ pl est :

$$pl_i = \left[\frac{f_{i1}}{f_{i\cdot}}, \frac{f_{i2}}{f_{i\cdot}}, ..., \frac{f_{iq}}{f_{i\cdot}}\right]^t, i = 1, ..., p. \tag{2.2}$$

Tableau des profils-lignes

On désigne par D_1 la matrice diagonale (p lignes et q colonnes) des effectifs marginaux de la v.q X avec N la matrice des effectifs. Le tableau est définit comme suit :

$$X_l = D_1^{-1}N,$$

avec

$$D_1 = \begin{bmatrix} n_{1\cdot} & 0 & ... & 0 \\ & & & \\ & & n_{i\cdot} & 0 \\ & & & \\ 0 & & & n_{p\cdot} \end{bmatrix} \text{ et } N = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1q} \\ n_{21} & n_{22} & & n_{2q} \\ \vdots & & \ddots & \vdots \\ n_{p1} & n_{p2} & \cdots & n_{pq} \end{bmatrix}.$$

Nuage des profils-lignes

Chaque profil-ligne est une suite de q valeurs numériques, il peut être représenté par un point de l'espace $\mathbb{R}^q$, dont chacune des dimensions est associée à une modalité de la variable Y. Donc les profils lignes forment un nuage de p points dans $\mathbb{R}^q$, chaque points de nuage est affecté d'un poids égale à sa fréquence marginale.

La matrice des poids est

$$D_l = \frac{D_1}{n}.$$

Le centre de gravité de ce nuage noté g_l est la moyenne pondérée de tous les points sur tous les axes j. Il est donné par :

$$g_l = \frac{N^t 1_p}{n} = (f_{\cdot 1}, f_{\cdot 2}, ..., f_{\cdot q})^t \in \mathbb{R}^q,$$

où 1_p est le vecteur de $\mathbb{R}^p$ dont toutes les composantes sont égale à 1.

Preuve. On a

$$\frac{N^t 1_p}{n} = \frac{1}{n} \begin{bmatrix} n_{11} & n_{21} & \cdots & n_{p1} \\ n_{12} & n_{22} & & n_{p2} \\ \vdots & & \ddots & \vdots \\ n_{1q} & n_{2q} & \cdots & n_{pq} \end{bmatrix} \times \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

$$= \frac{1}{n} \begin{bmatrix} \sum_{i=1}^{p} n_{i1} \\ \vdots \\ \sum_{i=1}^{p} n_{iq} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} n_{\cdot 1} \\ n_{\cdot 2} \\ \vdots \\ n_{\cdot q} \end{bmatrix} = \begin{bmatrix} f_{\cdot 1} \\ \vdots \\ \vdots \\ f_{\cdot q} \end{bmatrix} = g_l. \quad \blacksquare$$

Remarque 2.2 *Le centre de gravité s'interprète comme un profil-moyen. Dans l'étude des lignes, il sert de référence pour étudier dans quelle mesure et de quelle façon une classe*

d'individus diffère de l'ensemble de la population. Ceci se fait par l'étude de l'écart entre le profil de cette classe et le profil moyen. Ainsi l'étude de la dispersion du nuage autour de son barycentre équivaut à l'étude de l'écart entre profils et marge ou encore à l'étude de la liaison entre les deux variables.

Proposition 2.1 *Puisque la somme des coordonnées de chaque profil-ligne vaut 1, alors les p profils-lignes, ainsi que leur centre de gravité il en résulte que le nuage des profils-lignes appartient à un hyperplan W_l de dimension $\mathbb{R}^q$ d'équation*

$$\sum_{j=1}^{q} x_j = 1, x_j \geq 0.$$

Pour le g_l : $\sum_{j=1}^{q} f_{\cdot j} = 1, f_{\cdot j} \geq 0.$
Pour le pl : $\sum_{j=1}^{q} f_{j/i} = \frac{1}{f_{i\cdot}} \sum_{j=1}^{q} f_{ij} = \frac{1}{f_{i\cdot}} f_{i\cdot} = 1.$
On dit alors que le g_l et le pl appartient à W_l.

2.3.2 Profils-colonnes

Un profil-colonne est un vecteur de $\mathbb{R}^p$ dont la $i^{éme}$ coordonnée est égale à la fréquence conditionelle $f_{i/j}$, notée par pc_j. Le $j^{éme}$ pc est :

$$pc_j = \left[\frac{f_{1j}}{f_{\cdot j}}, \frac{f_{2j}}{f_{\cdot j}}, ..., \frac{f_{pj}}{f_{\cdot j}} \right]^t, j = 1, ..., q. \tag{2.3}$$

Tableau des profils-colonnes

On appel tableau des profils-colonnes, la matrice des fréquences conditionnelles $f_{j/i}$ contenant les p lignes et q colonnes. On désigne par D_2 la matrice diagonale des effectifs marginaux de la variable Y. Le tableau est définit comme suit :

$$X_c = D_2^{-1} N,$$

avec $D_2 = \begin{bmatrix} n_{.1} & 0 & & 0 \\ & & & \\ & & n_{.j} & 0 \\ 0 & & & n_{.q} \end{bmatrix}$.

Nuage des profils-colonnes

La construction du nuage des profils-colonnes s'effectue selon une démarche strictement identique à celle du nuage des profils-lignes. Les profils-colonnes forment un nuage de q points dans l'espace $\mathbb{R}^p$, dont chacune des dimensions est associée à une modalité de la variable X. Chaque point du nuage est affecté d'un poids égale à sa fréquence marginale.

La matrice des poids est

$$D_c = \frac{D_2}{n}.$$

Le centre de gravité de ce nuage noté g_c est la moyenne pondérée de tous les points sur tous les axes i. Il est donc donné par :

$$g_c = \frac{N1_q}{n} = (f_{1\cdot}, f_{2\cdot}, ..., f_{p\cdot})^t \in \mathbb{R}^p,$$

où 1_q est le vecteur de $\mathbb{R}^q$dont toutes les composantes sont égale à 1.

Preuve.

On a

$$\frac{N1_q}{n} = \frac{1}{n} \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1q} \\ n_{21} & n_{22} & & n_{2q} \\ \vdots & & \ddots & \vdots \\ n_{p1} & n_{p2} & \cdots & n_{pq} \end{bmatrix} \times \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} f_{1\cdot} \\ \vdots \\ \vdots \\ f_{p\cdot} \end{bmatrix} = g_c. \quad \blacksquare$$

Remarque 2.3 *Ce centre de gravité s'interprète également comme un profil moyen et joue*

le même rôle pour l'étude de la liaison entre les deux variables.

Proposition 2.2 *Pour les q profils-colonnes, ainsi que leur centre de gravité appartient à un hyperplan W_c de dimension $\mathbb{R}^p$, d'équation*

$$\sum_{i=1}^{p} x_i = 1, x_i \geq 0.$$

Pour le g_c : $\sum_{i=1}^{p} f_{\cdot j} = 1$ avec $f_{\cdot j} \geq 0$.
Pour le pc : $\sum_{i=1}^{p} f_{i/j} = \frac{1}{f_{\cdot j}} \sum_{i=1}^{p} f_{ij} = \frac{1}{f_{\cdot j}} f_{\cdot j} = 1$.
On dit alors que le g_c et le pc appartient à W_c.

Remarque 2.4 *On appel profil marginale ligne (resp profil marginale colonne) la quantité $\frac{n_{i\cdot}}{n}$ (resp $\frac{n_{\cdot j}}{n}$). L'écriture matricielle sera alors : $\frac{D_1}{n}$ (resp$\frac{D_2}{n}$).*

Proposition 2.3 *Dans le cas d'indépendance, le nuage des profils colonnes est réduit à un seul point qui est son centre de gravité g_l. Même chose concerant les profils lignes.*

Preuve.

Soient X et Y deux vqs indépendantes.

On a :

$$\begin{aligned} pc_j &= \left[\frac{f_{1j}}{f_{\cdot j}}, \frac{f_{2j}}{f_{\cdot j}}, ..., \frac{f_{pj}}{f_{\cdot j}}\right]^t \\ &= \left[\frac{f_{1\cdot} f_{\cdot j}}{f_{\cdot j}}, \frac{f_{2\cdot} f_{\cdot j}}{f_{\cdot j}}, ..., \frac{f_{p\cdot} f_{\cdot j}}{f_{\cdot j}}\right]^t \\ &= [f_{1\cdot}, f_{2\cdot}, ..., f_{p\cdot}]^t \\ &= g_c. \end{aligned}$$

Tous les profils colonnes son confondues avec g_c. ■

2.3.3 Ressemblance entre profils : Distance du χ^2

Question : Comment mesurer la dispersion de ces nuages de profils ?

Autrement dit, quelle métrique choisir dans chacun des espaces pour obtenir une bonne analyse ?

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

Distance entre deux profils lignes pl_i et $pl_{i'}$:

$$d_{\chi^2}(pl_i, pl_{i\prime}) = \sum_{j=1}^{q} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i\prime j}}{n_{i'\cdot}} \right)^2 = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i\prime j}}{f_{i\prime\cdot}} \right)^2 .$$

La métrique correspondante est une matrice diagonale :

$$\boxed{M_l = nD_2^{-1}.}$$

Distance entre deux profils colonnes pc_j et $pc_{j\prime}$:

Par analogie, on définit la distance entre deux profils colonnes pc_jet $pc_{j\prime}$par

$$d_{\chi^2}(pc_j, pc_{j\prime}) = \sum_{i=1}^{p} \frac{n}{n_{i\cdot}} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij\prime}}{n_{\cdot j\prime}} \right)^2 = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij\prime}}{f_{\cdot j\prime}} \right)^2 .$$

La métrique correspondante est une matrice diagonale

$$\boxed{M_C = nD_1^{-1}.}$$

2.3.4 Statistique du χ^2

Lorsqu' on étudie un tableau de contingence, il est classique de mesurer le caractère significatif de la liaison entre ces deux variables à l'aide de la statistique du χ^2. Appliquer à un tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques.

i.e :$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(effectif\ obs - effectif\ théo)^2}{effectif\ théo}$.

Alors :

$$\boxed{\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i\cdot} \times n_{\cdot j}}{n} \right)^2}{\frac{n_{i\cdot} \times n_{\cdot j}}{n}}.}$$

Remarque 2.5 *Si les deux variables sont indépendantes, alors* $\chi^2 = 0$.

2.3.5 Inerties totale

L'inertie totale du nuage de point est donnée par la formule suivante

$$\boxed{\varphi^2 = \frac{1}{n}\chi^2.}$$

Cette quantité mesure l'écart à l'indépendace.

L'inertie totale des deux nuages de profils lignes $I(pl)$ et colonnes $I(pc)$ est égale au lien entre les deux v.qs X et Y :

$$I(pl) = I(pc) = \varphi^2.$$

Preuve. On a

$$\begin{aligned}
I(pl) &= \sum_{i=1}^{p} f_{i.} d_{\chi^2}(pl_i, g_l) \\
&= \sum_{i=1}^{p} \frac{n_{i.}}{n} \sum_{j=1}^{q} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n} \right)^2 \\
&= \tfrac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n n_{i.}}{n_{.j}} \times \frac{1}{n_{i.}^2} \left(n_{ij} - \frac{n_{.j} \times n_{i.}}{n} \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n}{n_{.j} n_{i.}} \left(n_{ij} - \frac{n_{.j} \times n_{i.}}{n} \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{.j} \times n_{i.}}{n} \right)^2}{\frac{n_{.j} n_{i.}}{n}} \\
&= \frac{1}{n} \chi^2 \\
&= \varphi^2.
\end{aligned}$$

De la même maniere on peut montrer que $I(pc) = \varphi^2$. ■

2.4 ACP des nuages de profils

L'AFC peut être considéreé comme une version de l'ACP adaptée à des données qualitatives.

Elle est réalisée en appliquant l'ACP sur les tableaux de profils.

Tout en apportant les modifications nécessaires un des deux ensembles de modalités est choisi

comme ensemble "d'individus", l'autre comme ensemble de "variables" de l'ACP.

2.4.1 ACP du nuage des pl et des pc

Dans un tableau de contingence, les mots individus et variables n'ont pas la même signification que dans le tableau des données de l'ACP. En effet, dans le tableau de contingence les lignes et les colonnes représentent les modalités de deux caractères. Pour conserver une homogénéïté dans la présentation des deux analyses, les p modalités du caractère X en lignes portent le nom d'individus et les q modalités du caractère Y en colonnes. Donc les profils lignes jouent le rôle des individus et les profils colonnes sont des variables.

1. ACP du nuage des Profils lignes :
   - Tableau des données :$X = X_l = D_1^{-1}N$.
   - Métrique :$M = nD_2^{-1}$.
   - Matrice de poids :$D = D_1/n$.
   - Centre de gravité $g_l = X_l^t D_l 1_p$.
2. ACP du nuage des profils colonnes :
   - Tableau des données : $X = X_c = D_2^{-1}N$.
   - Métrique :$M = nD_1^{-1}$.
   - Matrice de poids :$D = D_2/n$.
   - Centre de gravité $g_c = X_c^t D_c 1_q$.

2.4.2 Propriétés des profils moyens

1-Le vecteur $\overrightarrow{Og_l}$ (l'origine O au centre de gravité g) est orthogonal sens de la métrique du χ^2 à l'hyperplan W_l contenant le nuage des profils, i.e : $\overrightarrow{Og_l} \perp W_l$.

Preuve.

En effet, soit $x = (x_1, ..., x_q)^t$ un élément de W_1 :

$$\begin{aligned}
\text{On a par définition } \langle \overrightarrow{Og_l}, \overrightarrow{g_l}x\rangle_{M_l} &= (x-g_l)^t M_l g_l \\
&= (x-g_l)^t nD_2^{-1} g_l. \\
&= (x-g_l)^t I \text{ car } nD_2^{-1}g_l = I \\
&= x^t I - g_l^t I \\
&= \sum_{j=1}^{q} x_j^t - \sum_{j=1}^{q} f_{.j} \\
&= 1-1 \\
&= 0.
\end{aligned}$$

■

2-Le profil moyen g_l est un axe principal i.e c'est un vecteur propre M-normé de la matrice VM associé à la valeur propre $\lambda = 0$.

Preuve. On a

1-On montre que $\|g_l\|^2 = 1$.

$$\begin{aligned}
\|g_l\|_M^2 &= g_l^t M g_l \\
&= g_l^t nD_2^{-1} g_l \\
&= g_l^t I \text{ car } nD_2^{-1}g_l = I \\
&= 1.
\end{aligned}$$

2-On montre que g_l est un $\overrightarrow{vp}$ de VM associé à la vp $\lambda = 0$:

$$\begin{aligned}
VMg_l &= (X_l^t D_l X_l - g_l g_l^t) M g_l \\
&= \left(X_l^t \tfrac{D_1}{n} X_l - g_l g_l^t\right) nD_2 g_l \\
&= X_l^t \frac{D_1}{n} X_l I - g_l, \text{ car } g_l^t nD_2^{-1} g_l = g_l^t M g_l = \|g_l\|_M^2 = 1 \\
&= \frac{N^t D_1^{-1}}{n} NI - g_l, \text{ car } X_l = D_1^{-1}N \\
&= \frac{N^t I}{n} - g_l, \text{ car } D_1^{-1}NI = I \\
&= g_l - g_l \\
&= 0.
\end{aligned}$$

■

3-Le profil moyen g_l est aussi un $\overrightarrow{vp}$ de la matrice $X^t DXM$ associé à la vp $\lambda = 1$.

Preuve. Comme $VM_l g_l = (X_l^t D_l X_l - g_l g_l^t) M_l g_l = 0$.

Donc $X_l^t D_l X_l M_l g_l = g_l g_l^t M_l g_l$.

Alors $X_l^t D_l X_l M_l g_l = 1 g_l$, car $g_l^t M_l g_l = 1$

Enfin $\lambda = 1$

-Le facteur principal : $M_l g_l = nD_2^{-1} g_l = 1_q$.

-La composante principale correspondante : $X_l 1_q = D_1^{-1} N 1_q$. ■

2.4.3 ACP non centrées et facteur trivial

Le vecteur $\overrightarrow{Og}$ est orthogonal au support de ce nuage donc g est dit facteur principal ou trivial. Autrement dit, g est un vecteur propre associé à VM avec $\lambda = 0$. Du point de vue technique, on peut montrer qu'il n'est pas nécessaire de centrer explicitement le nuage de point avant de l'analyser. En effet, mis à part le premier facteur, l'analyse du nuage par rapport à O sans centrage conduit aux mêmes facteurs que l'analyse du nuage centré.

Proposition 2.4 *Les vecteurs propres de VM sont les même que $X^t DXM$ et avec les mêmes valeurs propres.sauf pour g qui change de valeur propre de $\lambda = 0$ à $\lambda = 1$.*

Preuve. En effet

$$\begin{aligned} \text{1- } X^t DXMg &= VMg + gg^t Mg \\ &= 0 + g \left\| g \right\|_M^2 \\ &= 0 + g \times 1 \\ &= g. \end{aligned}$$

Ce qui implique que $\lambda = 1$. D'autre part on a pour tout vecteur $a \in \mathbb{R}^q$

$$\begin{aligned} \text{2- } VMa = \lambda a \quad \Rightarrow X^t DXMa &= VMa + gg^t Ma \\ &= \lambda a + g \left\langle g, a \right\rangle_M \\ &= \lambda a. \end{aligned}$$

i.e $a \perp g$ car les vecteurs g et a sont deux vecteurs propres de la matrice VM qui est M-symétrique. Alors d'après la (Remarque 1.2) ils sont M-orthogonaux. En d'autre terme $\left\langle g, a \right\rangle_M = g^t Ma = 0$ et par conséquent le vecteur $g \left\langle g, a \right\rangle_M$ est nul. ■

On remarque qu'on a les mêmes valeurs propres donc il est inutile de centrer les tableaux des profils, on efectuera donc des ACP non centrées et on éliminera la valeur propre $\lambda = 1$ associée au facteur principal g car elle est maximale.

Remarque 2.6 *On déduit que les projections M-orthogonales des profils sur l'axe principale g sont toutes confondues avec g.*

2.4.4 Résumé des deux ACP

Les facteurs principaux sont les vecteurs propres de la matrice MX^tDX et les composantes principaux (qui sont centrée en AFC)sont les vecteurs propres de la matrice XMX^tD. Les deux analyses ACP des *pl* et *pc* conduisent aux mêmes valeurs propres et que les facteurs prinsipaux de l'une sont les composantes pricipales de l'autre.

On résume les résultats des deux ACP dans le tableau suivant :

	Facteurs principaux	Composantes principales
ACP des pl	Vecteur propres de $D_2^{-1}N^tD_1^{-1}N$	Vecteur propres de $D_1^{-1}ND_2^{-1}N^t$
ACP des pc	Vecteur propres de $D_1^{-1}ND_2^{-1}N^t$	Vecteur propres de $D_2^{-1}N^tD_1^{-1}N$

TAB. 2.3 – Résumé des résultats des deux ACP des pl et pc

Remarque 2.7 *La symétrie parfaite entre les deux ACP permet de superposer les plans principaux des deux ACP. Cela donne une répresentation graphique simultanée des modalités des deux variables X et Y.*

2.4.5 Formules de transition

Les facteurs sur les lignes et ceux sur les colonnes sont liés par des relations dites de transition. Cette formule présente les relations entre les points représentant d'une part les lignes et d'autre part les colonnes.

Proposition 2.5 *soit $a = (a_1, ..., a_p)^t$ et $b = (b_1, ..., b_p)^t$ sont des composantes principales des pl et pc respectivement. Les deux formules de transition s'écrivent :*

$$a = \frac{1}{\sqrt{\lambda}} D_1^{-1} N b \text{ et } b = \frac{1}{\sqrt{\lambda}} D_2^{-1} N^t a.$$

Remarque 2.8 *Les composantes principales associées aux valeurs propres $\lambda \neq 1$sont centrées, ie :*$\left(\overline{a} = \overline{b} = 0\right)$.

Preuve. On a

$a = (a_1, ..., a_p)^t \in \mathbb{R}^p$ est un vecteur propre de matrice $D_1^{-1}ND_2^{-1}N^t$ associé à vp $\lambda \neq 1$.

Et $\overline{a} = \sum_{i=1}^{p} f_{i.} a_i = \langle g_c, a \rangle = g_c^t a = 0$?
$g_c = X^t D X M g_c$ avec $X = X_c^t = D_2^{-1} N^t$ et $D = \frac{D_2}{n}$,$M = n D_1^{-1}$
Donc $g_c = N D_2^{-1} N^t D_1^{-1} g_c \Rightarrow \overline{a} = g_c^t a = g_c^t D_1^{-1} N D_2^{-1} N^t a = g_c^t \lambda a$ d'apré2.1
$\Rightarrow \overline{a} = g_c^t \lambda a = \lambda \overline{a} \Rightarrow \overline{a} - \lambda \overline{a} = 0 \Rightarrow \overline{a} = 0$ car $\lambda \neq 1$. ■

2.4.6 La décomposition de l'inertie

On définit l'inertie totale par :

$$\phi^2 = \sum_{k=1}^{n} \lambda_k,$$

Avec $n = min(p-1, q-1)$ car il y a au plus $min(p-1, q-1)$ valeurs propres.
Ainsi les pourcentages d'inertie sont donnés par $\dfrac{\lambda_k}{\phi^2}$.

2.4.7 La contribution des modalités aux inerties des axes

Pour interpréter correctement les graphiques, il faut comme en ACP tenir compte, d'une part, de la proximité entre points et plans principaux et d'autre part, du rôle joué par chaque point dans la détermination d'un axe. Comme les données étant qualitatives, on n'utilise pas ici les cercles de corrélations entre caractères et axes principaux, donc l'interpétation des composantes se fait essentiellement en utilisant les contributions des modalités aux inerties des axes factoriels.
On rappelle que la contribution totale est donnée par :

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

-Ainsi on définit la contribution du pl_i est :

$$\boxed{CTR\,(pl_i) = \frac{f_{i.} a_i^2}{\lambda},\ i = 1, ..., p.}$$

-La contribution du profil colonnes pc_j est :

$$\boxed{CTR\,(pc_j) = \frac{f_{.j} b_i^2}{\lambda},\ j = 1, ..., q}$$

En pratique, on considère les modalité ayant les plus forte contribution, lorsqu'elle dépasse son poids

$$CTR\,(pl_i) \succ f_{i.} \text{ et } CTR\,(pc_j) \succ f_{.j}.$$

Remarque 2.9 *Avant de cloturer ce modeste travail, une application est faite sur de différentes taches ménagères réparties dans les couples (*1744 *couples). Les données corespondantes sont résumées dans un tableau de contingence contenant deux vqs X et Y. La $1^{ière}$ vq définie les 13 différentes taches ménagères (lignes). Tandis que la $2^{ième}$ représente leur répartitions dans le couple (colonnes). Les valeurs (coordonnées) sont les fréquences des taches effectuées. Les informations prisent sont extaires partir d'un tableau nommé "housetasks" de la bibiliothèque du logiciel R. Le but de ce travail est de faire une programmation à l'aide de ce logiciel et d'étudier la ressemblance qui peut y avoir entre ces deux vqs. Pour cela, différents packages et fonctions sont mentionnés dans la partie "Annexe B".*

Conclusion

Aujourd'hui, les méthodes d'analyse des données sont employées dans un grand nombre de domaines qu'il est impossible d'énumérer : l'ingéniorat, gestion, économie...etc. Les objectifs des chercheurs en analyse de données est de répondre aux problèmes posés par des tableaux de grandes dimensions.

Parmi ces méthodes, l'analyse factorielle des correspondances "AFC" qui est une méthode puissante pour synthétiser et résumer de vastes tableaux de contingence, pour objet de traiter des informations obtenues dans les domaines les plus divers, dans des situations très complexes, toutes sont destinées à aider l'utilisateur à la lecture et l'interprétation de l'informations d'une manière plus simple.

De plus, l'AFC permet d'obtenir des graphes simples afin de mieux observer le phénomène étudié et d'avoir une meilleure intuition.

Dans le cas de deux ou plusieurs variables qualitatives. Une autre méthode est utilisable appelée méthode d'analyse factorielle ou méthode d'Analyse des Correspondances Multiples notée ACM.

Bibliographie

[1] Arnaud M. L'analyse de données Polycopié de cours ENSIETA-Réf : 1463. Septembre 2004

[2] Bouchier, A.L'analyse des données multivariées à l'aide du logiciel (A.F.C.s.) 2 mars 2010.

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

[4] Bouredji, H. Analyse factorielle des corespondances simple et multiple. Juin 2016, Biskra

[5] Bounkhala, A. Méthodes ACP et AFC en statistiques et leurs applications, Tlemcen.

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

[7] Escofier,B.Pagés,J. (2008). analyse factorielle simples et multiples. Objectifs, méthodes et interprétation.Dunod

[8] Kassambara, A. Practical Guide to Principal Component Methods in R_STHDA-2017 (edition 1) France

[9] Meraghni, D. (2018). Cours de master 2 Modele Lineaire. Université Mohammed Khider. Biskra.

[10] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

Annexe A : Logiciel *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 : Exemple d'application

Tableau de contingence initial 2.4 :

	Femme	En-alt	Mari	conj	Total
Blan	156	14	2	4	176
Re-prin	124	20	5	4	153
Diner	77	11	7	13	108
Pet-deje	82	36	15	7	140
Ranger	53	11	1	57	122
Vaisselle	32	24	4	53	113
Achats	33	23	9	55	120
Officiel	12	46	23	15	96
Conduite	10	51	75	3	139
Finances	13	13	21	66	113
Assurances	8	1	53	77	139
Repara	0	3	160	2	165
Vacances	0	1	6	153	160
Total	600	254	381	509	1744

TAB. 2.4 – Tableau de contingence des taches ménagéres et leurs répartitions dans le couple

Variable X :

Blan : Blanchisserie ; **Re-prin :** Repas principal ; **Pet-deje** : Petit déjeuner ; **Repara :** Réparations ; Dîner ; **Ranger**, **Vaisselle**, **Achats**, **Officiel**, **Conduit**, **Finances**,

Variable Y :

- **Femme :** travaux effectués par la femme seulement.

- **En-alt :** travaux effectués alternativement.

- **Mari :** travaux effectués par l'homme seulement.

- **conj :** travaux effectués ensemble.

Voici quelques packages et fonctions utilisés pour la programmation.

Packages :

library(FactoMineR) # Calculer et importer les analyses correspondantes, dédié à l'analyse de données

library(ade4) # ADE4 est un package spécialisé en analyses multivariées

Fonctions :

data, sum, numeric, barplot, round, dudi.coa, chisq.test, range,

Programme :

X=data(housetasks) # Importer le tableau de contingence.

n=sum(X) # Effectif total des couples.

marl=numeric(13)

for(i in 1 :13) {marl[i]<-sum(X[i,])} # Boucle pour calculer $n_{i.}$.

marl # Effectifs marginales $n_{i.}$.

176	153	108	140	122	113	120	96	139	113	139	165	160

marc=numeric(4)

for(j in 1 :4) {marc[j]<-sum(X[,j])} # Boucle pour calculer $n_{.j}$.

marc

600	254	381	509

Effectifs marginales $n_{.j}$.

op=par(mfrow=c(1,2))

barplot(marl,names.arg=(row.names(X)),las=2,col="yellow")2.1

barplot(marc,names.arg=(row.names(t(X))),las=2,col="green")

par(op)

x11()

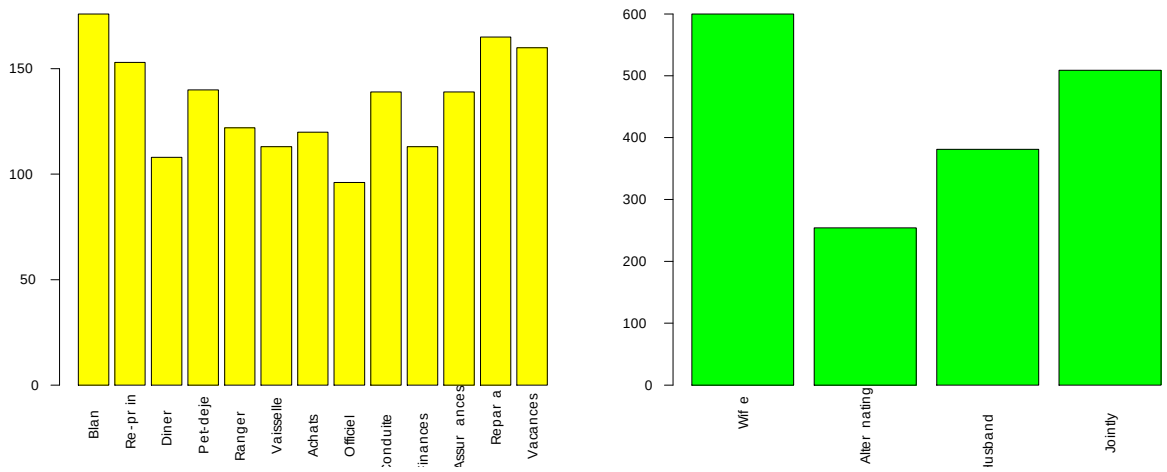

FIG. 2.1 – Histogrammes des effectifs marginaux $n_{i\cdot}$ (à gauche) et $n_{\cdot j}$ (à droite).

F=X/n # Tableau des frequences.2.5

round(F,3)

	Femme	En-alt	Mari	conj	Total
Blan	0.089	0.008	0.001	0.002	0.101
Re-prin	0.071	0.011	0.003	0.002	0.088
Diner	0.044	0.006	0.004	0.007	0.062
Pet-deje	0.047	0.021	0.009	0.004	0.080
Ranger	0.030	0.006	0.001	0.033	0.070
Vaisselle	0.018	0.014	0.002	0.030	0.065
Achats	0.019	0.013	0.005	0.032	0.069
Officiel	0.007	0.026	0.013	0.009	0.055
Conduite	0.006	0.029	0.043	0.002	0.080
Finances	0.007	0.007	0.012	0.038	0.065
Assurances	0.005	0.001	0.030	0.044	0.080
Repara	0.000	0.002	0.092	0.001	0.095
Vacances	0.000	0.001	0.003	0.088	0.092
Total	0.344	0.146	0.218	0.292	1

TAB. 2.5 – Tableau des fréquances

fmarl=marl/n # Frequence marginale $f_{i\cdot}$.

round(fmarl,3)

0.101	0.088	0.062	0.080	0.070	0.065	0.069	0.055	0.080	0.065	0.080	0.095	0.092

sum(fmarl)

1 # Frequence totale.

fmarc=marc/n # Frequence marginale $f_{.j}$.

round(fmarc,3)

0.344	0.146	0.218	0.292

sum(fmarc)

1

op=par(mfrow=c(1,2))

barplot(fmarl,names.arg=(row.names(F)),las=2,col="yellow",ylim=c(0,1))2.2

barplot(fmarc,names.arg=(row.names(t(F))),las=2,col="green",ylim=c(0,1))

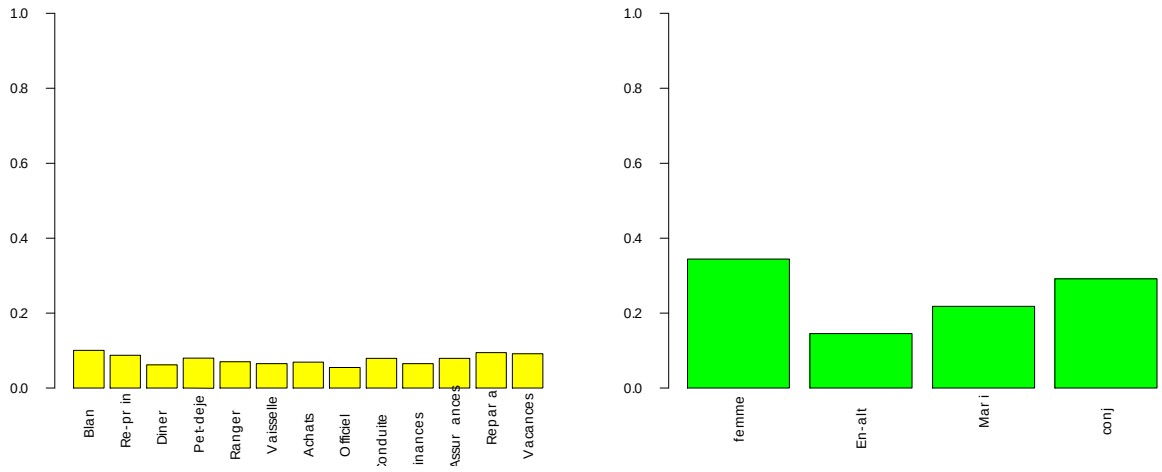

FIG. 2.2 – Histogrammes des frequences marginales $f_{i.}$ (à gauche) et $f_{.j}$ (à droite).

gl=fmarl # Centre de gravité g_l.

gc=fmarc # Centre de gravité g_c.

pl=round(X/marl,3)2.6

rowSums(pl)

	Femme	En-alt	Mari	conj	Total
Blan	0.886	0.080	0.011	0.023	1.000
Re-prin	0.810	0.131	0.033	0.026	1.000
Diner	0.713	0.102	0.065	0.120	1.000
Pet-deje	0.586	0.257	0.107	0.050	0.999
Ranger	0.434	0.090	0.008	0.467	0.999
Vaisselle	0.283	0.212	0.035	0.469	0.999
Achats	0.275	0.192	0.075	0.458	1.000
Officiel	0.125	0.479	0.240	0.156	1.000
Conduite	0.072	0.367	0.540	0.022	1.001
Finances	0.115	0.115	0.186	0.584	1.000
Assurances	0.058	0.007	0.381	0.554	1.000
Repara	0.000	0.018	0.970	0.012	1.000
Vacances	0.000	0.006	0.038	0.956	1.000

TAB. 2.6 – Tableau des profiles-lignes

pc=round(t(t(x)/marc),3)2.7

colSums(pc)

	Femme	En-alt	Mari	conj
Blan	0.260	0.055	0.005	0.007
Re-prin	0.2067	0.078	0.013	0.007
Diner	0.128	0.043	0.018	0.025
Pet-deje	0.136	0.141	0.039	0.013
Ranger	0.088	0.043	0.003	0.112
Vaisselle	0.053	0.094	0.010	0.104
Achats	0.055	0.091	0.024	0.108
Officiel	0.020	0.181	0.060	0.029
Conduite	0.017	0.201	0.197	0.006
Finances	0.022	0.051	0.055	0.130
Assurances	0.013	0.004	0.139	0.151
Repara	0.000	0.012	0.420	0.004
Vacances	0.000	0.004	0.016	0.301
Total	1.000	1.000	0.999	1.001

TAB. 2.7 – Tableau des profils-colonnes

afc=dudi.coa(X,nf=2,scannf=FALSE) # Analyse factorielle.

vp=afc$eig

0.543	0.445	0.127

Valeurs propres.

pvp=(vp/sum(vp))*100

48.692	39.913	11.395

Pourcentage des vps.

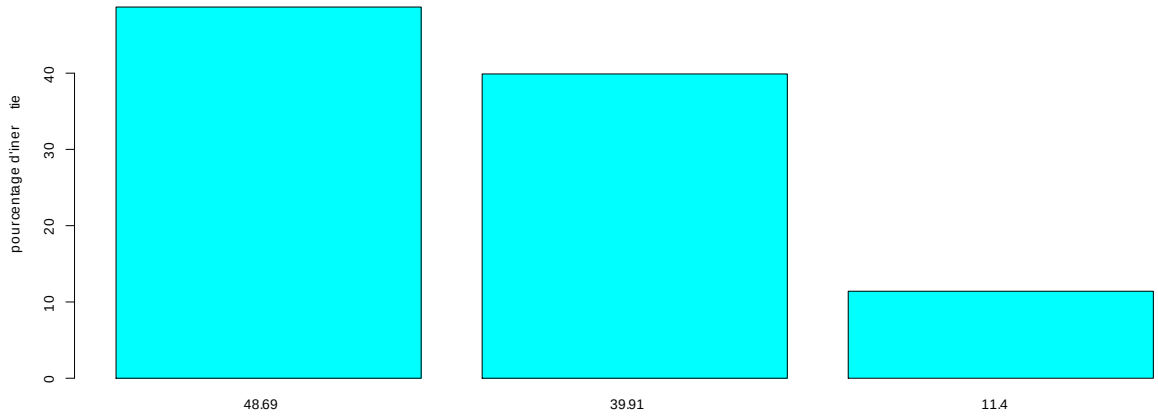

FIG. 2.3 – Pourcentage des valeurs propres.

a<-chisq.test(x) # Test d'independance χ^2.

ki2<-a$statistic

1944.456

ficarré<-ki2/n

1.115 # Inertie totale φ^2.

Comentaire :

Interprétation de l' AFC : la premiere étape consiste à évaluer s'il existe une laison entre les lignes et les colonnes par la statistique du chi-deux où à $\chi^2 = 1944.456$. En la divisant par le nombre d'effectif total $n = 1744$ on trouver l'inertie commune aux deux nuages de profils $\varphi^2 = 1.114$ et qui s'accorde avec la somme des valeurs propres. Les valeurs propres sont utilisées pour déterminer le nombre d'axes, les plus grandes valeurs propres sont, $\lambda_2 = 0.445$. La figure 2.3 montre les pourcentages des valeurs propres. On a :

-Le $1^{ére}$ axe représenter environ 49% d'inertie totale λ_1.

-Le $2^{éme}$ axe représenter environ 40% d'inertie totale λ_2.

-Le $3^{éme}$ axe représenter seulement 11% d'inertie totale λ_3.

Pour cela, on prende que les 2 premiére axes principale(expliquent 89% de l'information disponible).Donc l'analyses est bonne lorsque les premiéres dimensions representent une grande partie de la variabilité.

afc=dudi.coa(x) # On prends comme exemple 2 axes.

round(afc$li,3) # Composantes principales des pl.2.8

ctrl=round(inertia.dudi(afc,row.inertia=TRUE)$row.abs,3) # $CTR(pl)$.

	Coordonées		Contributions	
	Axe1	Axe2	Axe1	Axe2
Blan	0.992	0.495	18.287	5.564
Re-prin	0.876	0.490	12.389	4.736
Diner	0.693	0.453	5.471	1.321
Pet-deje	0.509	0.308	3.825	3.699
Ranger	0.394	−0.434	1.998	2.966
Vaisselle	0.189	−0.442	0.426	2.844
Achats	0.118	−0.403	0.176	2.515
Officiel	−0.227	0.254	0.521	0.796
Conduite	−0.742	0.653	8.078	7.647
Finances	−0.271	−0.618	0.875	5.559
Assurances	−0.647	−0.474	6.147	4.020
Repara	−1.529	0.864	40.730	15.881
Vacances	−0.252	−1.435	1.077	42.454

TAB. 2.8 – Composantes principales et CTR(pl).

range(round(afc$li[,1],3))

−1.529	0.992

Bornes du 1^{er} axe.

range(round(afc$li[,2],3))

−1.435	0.864

Bornes du $2^{ième}$ axe.

round(afc$co,3) # Composantes principales des pc.2.9

ctrc<-round(inertia.dudi(afc,col.inertia=TRUE)$col.abs,3) # $CTR(pc)$.

	Coordonées		Contributions	
	Axe1	Axe2	Axe1	Axe2
Femme	0.838	0.365	44.462	10.312
En-alt	0.062	0.292	0.104	2.783
Mari	−1.161	0.602	54.234	17.787
conj	−0.149	−1.027	1.200	69.118

TAB. 2.9 – Composantes principales et CTR(pc)

range(round(afc$co[,1],3))

−1.161	0.838

Bornes du 1^{er} axe.

range(round(afc$co[,2],3))

−1.027	0.602

Bornes du $2^{ième}$ axe.

plot(afc$co[,1],afc$co[,2],xlim=c(-1.528,0.992),ylim=c(-1.16,0.837),type="n",xlab="48.69%",
ylab="39.91%")

text(afc$co[,1],afc$co[,2],row.names(afc$co),col="red")2.4

text(afc$li[,1],afc$li[,2],row.names(afc$li),col="blue")

abline(h=0,v=0)

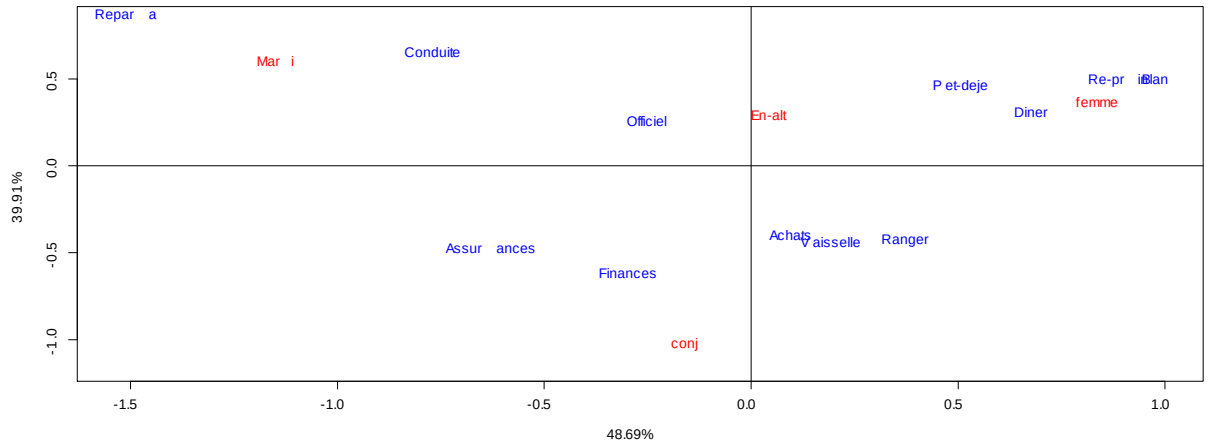

FIG. 2.4 – Représentation simultanée des tâches ménagères réparties dans les couples

Comentaire :

Le graphique 2.4 représente les projections des modalités X et Y sur le premier plan principal, pour montrer les profils lignes et colonnes simultanément, seule la distance entre les points lignes ou points colonnes peut être vraiment interprétée. En remarqe :

-Les taches ménagères (Blan, Re-prin, Diner, Pet-deje) sont plus fréquemment effectées le femme.

-Les taches ménagères (Repara, Conduite) sont associées au mari.

-Les (vacances, Finances, assurances) sont associées au conjoint.

-Les taches (Officiel, achat, vaisselle, ranger) sont fréquemment En-alt.

On note que les travaux ménagers effectués par la femme sont nombreuses en comparaison aux autres travaux domestiques.

Annexe C : Abréviations et Notations

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

Notation	:	signification
$\mathbb{ACP}$	:	Analyse en composantes principales.
AFC	:	Analyse factorielle des correspondances.
n	:	Nombre d'individus.
X, Y	:	Variables qualitatives (vqs).
p, q	:	Nombre de modalités.
$x_i y_j$	:	Modalités (catégories).
$n_{i.}, n_{.j}$	:	Marges lignes et colonnes respectivement.
f_{ij}	:	Fréquence de la ccellule(i, j).
$f_{i.}, f_{.j}$	:	Fréquence de la marge "i" et de la marge "j" respectivement.
H_0	:	Hypothese nulle.
$i.e$	:	C'est à dire.
CTR	:	Contribution.
Tr	:	Trace.
$d_{\chi^2}(a, b)$	:	Distance du chi-deux entre a et b.
$\langle a, b \rangle_M$	:	Produit scalaire entre a et b au sens du métrique M.
g	:	Centre de gravité.
I	:	Inertie.
pl	:	Profil-ligne.
pc	:	Profil-colonne.

w_l : hyperplan de pl

w_c : hyperplan de pc

1_p : vecteur unitaire de taille p

1_q : vecteur unitaire de taille q

I_n : matrice identiti

R : ensemble des nombre reel

vp : valeur propre

$\overrightarrow{vp}$: vecteur propre

vq : variable qualitative

vqs : variables qualitatives