Université Abderrahmane Mira-Bejaia
Faculté des Sciences Économiques, Commerciales et des Sciences de Gestion

Département des Sciences Economiques
Laboratoire Economie et Développement

Polycopié pédagogique

Dossier numéro (à remplir par l'administration) **: SE/MTN103/2023**

Titre

Cours d'analyse de données

Cours destiné aux étudiants de

M2-Economie quantitative.

Dr. Bouznit Mohammed

Année : 2022/2023

SOMMAIRE

Introduction	**1**
Chapitre 1. Quelques rappels de mathématiques et d'analyse bidimensionnelle	**1**
Chapitre 2. Analyse factorielle générale	**12**
Chapitre 3. Analyse en composantes principales (ACP)	**20**
Chapitre 4 . Analyse factorielle des correspondances, et Analyse factorielle des correspondances multiples.	**61**
Références	**76**
Table des métiers	

Introduction générale

Le présent polycopié de cours d'analyse de données est destiné aux étudiants de Master 2-Economie Quantitative. Il vise à présenter les fondements théoriques et pratiques des méthodes descriptives d'analyse de données. Ces méthodes permettent de résumer, représenter graphiquement, réduire les dimensions, regrouper et visualiser l'information pertinente contenue dans de grands ensembles de données multidimensionnelles. Pour une meilleure assimilation du cours, les étudiants doivent avoir des connaissances, plus au moins approfondies, en statistique, en algèbre, et en informatique.

Ce polycopié de cours comporte 4 chapitres. Le premier est consacré aux rappels de mathématiques et d'analyse bidimensionnelle. Le deuxième chapitre présente les fondements théoriques de l'analyse factorielle générale. Les aspects théoriques de la méthode ACP et son application en utilisant le logiciel SPSS sont présentés dans le chapitre 3. Cependant, l'analyse factorielle des correspondances, l'analyse factorielle des correspondances multiples, ainsi que des exemples d'application sur xl-stat seront présentées dans les chapitres 4.

Chapitre 1. Quelques rappels de mathématiques et d'analyse bidimensionnelle

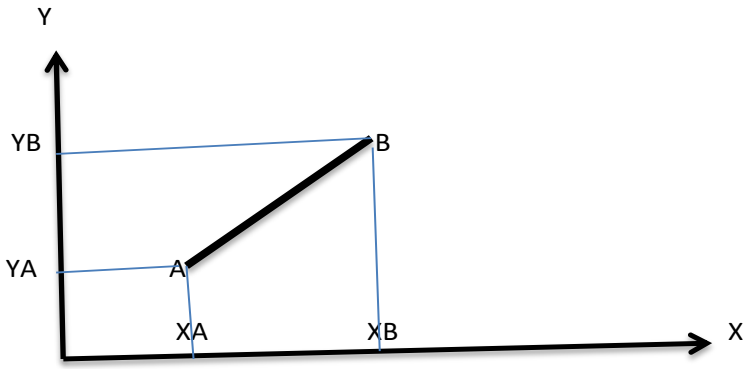
1.1. Produit scalaire dans un plan (deux dimensions)

Soit le vecteur $\vec{U} = \overrightarrow{AB}$, alors la norme de $\vec{U}$, notée $\| \vec{U} \|$, est la longueur de $[AB]$

Remarque 1. 1. *si le vecteur* $\vec{U}$ *a pour coordonnées*(x, y), *alors* $\| \vec{U} \| = \sqrt{x^2 + y^2}$

1.2. Distance entre deux points dans un plan (deux dimensions)

Soient $A(x_A, y_A)$ et $B(x_B, y_B)$ deux points dans un plan



La distance entre A et B, notée $\| \overrightarrow{AB} \|$, est donnée par la formule suivante :

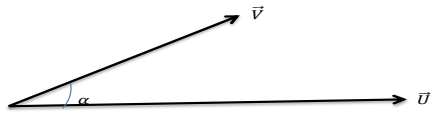
$$\| \overrightarrow{AB} \| = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

Propriété 1

- ✓ $\| \vec{U} \| \in IR^+$
- ✓ $\| \vec{U} \| = 0 \Longrightarrow \vec{U} = \vec{0}$

Propriété 2

Soient $\vec{U}$ et $\vec{V}$ deux vecteurs ;

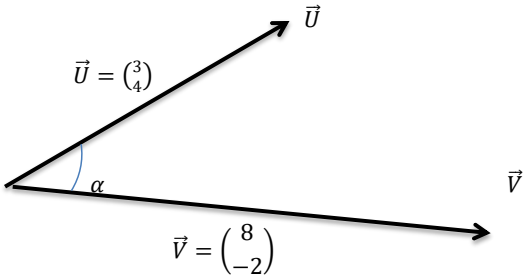


Le produit scalaire $\langle \vec{U}, \vec{V} \rangle = \vec{U}\,\vec{V} = \| \vec{U} \| \| \vec{V} \| Cos(\vec{U}, \vec{V}) \in IR$

$= \| \vec{U} \| \| \vec{V} \| Cos(\alpha)$

Exemple

Soient $\vec{U}$ et $\vec{V}$ deux vecteurs, tel que :



$\langle \vec{U}, \vec{V} \rangle = \vec{U}\,\vec{V} = \| \vec{U} \| \| \vec{V} \| Cos(\vec{U}, \vec{V}) = \| \vec{U} \| \| \vec{V} \| Cos(\alpha)$

où $\|\vec{U}\| = \sqrt{4^2 + 3^2} = 5$ et $\|\vec{V}\| = \sqrt{8^2 + (-2)^2} = \sqrt{68}$

calculer $Cos(\alpha)$? $Cos(\alpha) = \frac{\langle \vec{U}, \vec{V} \rangle}{\|\vec{U}\| \|\vec{V}\|} = \frac{\vec{U}\,\vec{V}}{\|\vec{U}\| \|\vec{V}\|}$

Par ailleurs $\langle \vec{U}, \vec{V} \rangle = \vec{U}\,\vec{V} = \binom{4}{3}\binom{8}{-2}$=(4)(8)+(3)(-2)=26

donc ;

$Cos(\alpha) = \frac{26}{5\sqrt{68}}$=0,63 $\Rightarrow \alpha = Cos^{-1}$(0,63)=50,9°

Propriété 3

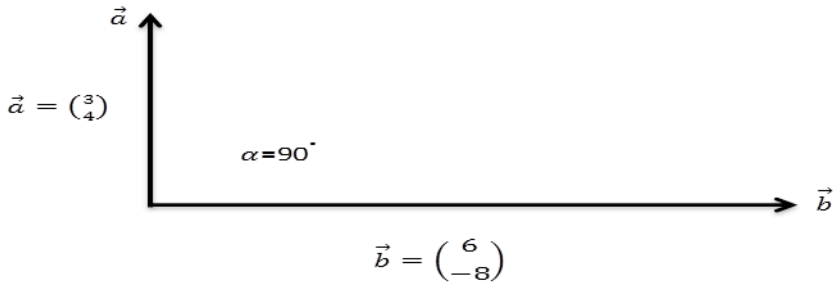
- ✓ Si $\vec{U}\binom{x_u}{y_u}$ **et** $\vec{V}\binom{x_v}{y_v}$**,** alors $\vec{U}\,\vec{V} = \binom{x_u}{y_u}\binom{x_v}{y_v}$=$\boldsymbol{x_u x_v + y_u y_v}$

Propriété 4 *:*

Si $\vec{U} \perp \vec{V}$, alors $\vec{U}\,\vec{V} = 0$

Exemple

Soient $\vec{a}\binom{4}{3}$**,** $\vec{b}\binom{6}{-8}$**,** et $\alpha = 90°$=$\frac{\pi}{2}$



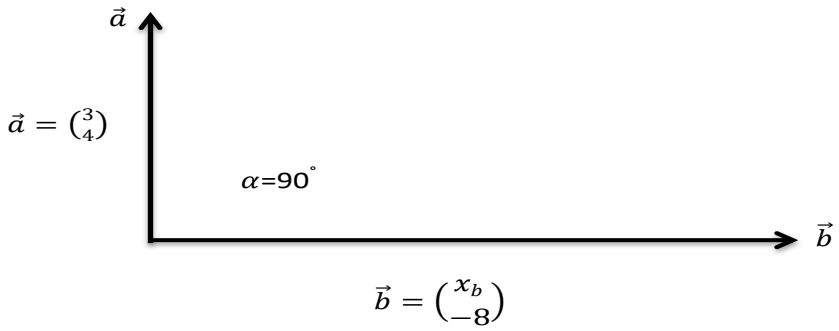
$\langle \vec{a}, \vec{b} \rangle = \vec{a}\,\vec{b} = \|\vec{a}\| \|\vec{b}\| Cos\left(\frac{\pi}{2}\right)$ =0

Par ailleurs ;

$$\langle \vec{a}, \vec{b} \rangle = \vec{a}\,\vec{b} = \binom{4}{3}\binom{6}{-8} = (6)(4) + (3)(-8) = 24 - 24 = 0$$

Donc , $\vec{a}$ *et* $\vec{b}$ sont orthogonaux ($\vec{a} \perp \vec{b}$)

Exemple



Calculer x_b ?

$\vec{a} \perp \vec{b}$, alors $\vec{a}\,\vec{b} = 0 \Rightarrow \binom{4}{3}\binom{x_b}{-8}$=0 $\Rightarrow 4x_b + (3)(-8) = 0 \Rightarrow x_b = 6$

Propriétés

Soient $\vec{U}$, $\overrightarrow{V}$ *et* $\overrightarrow{W}$ trois vecteurs et $(a, b) \in \mathbb{R}^2$

- ✓ $\vec{U}\left(a\overrightarrow{V} + b\overrightarrow{W}\right) = a\vec{U}\,\overrightarrow{V} + b\vec{U}\overrightarrow{W}$
- ✓ $(a\vec{U} + b\overrightarrow{V})\,\overrightarrow{W} = a\vec{U}\,\overrightarrow{W} + b\overrightarrow{V}\overrightarrow{W}$
- ✓ $\vec{U}\,\vec{V} = \overrightarrow{V}\vec{U}$
- ✓ $\vec{U}\vec{U} = \|\vec{U}\|^2$
- ✓ $\vec{U}\,\vec{V} = 0 \Rightarrow \vec{U} \perp \vec{V}$
- ✓ $\vec{0}\vec{U} = 0$

1.3. Calcul matriciel

1.3.1. Les valeurs propres d'une matrice

Soient $A_{(p,p)}$ une matrice carrée et $\lambda \in \mathbb{R}$. On dit que λ valeur propre de $A_{(p,p)}$ si et seulement si :

$|A_{(p,p)} - \lambda I_p|$=0, tel que I_p : La matrice identité

Remarque 1.2.

Le nombre des valeurs propres de la matrice $A_{(p,p)}$ *est égal à p*

Exemple

Soit $A_{(2,2)}=\begin{pmatrix}2 & 0\\ 3 & 1\end{pmatrix}$, calculer les valeurs propres de $A_{(2,2)}$?

Solution

$$\left| A_{(2,2)} - \lambda I_2 \right| = 0 \Rightarrow \left|\begin{pmatrix}2 & 0\\ 3 & 1\end{pmatrix} - \lambda \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right| = 0$$

$$\left|\begin{pmatrix}2-\lambda & 0\\ 3 & 1-\lambda\end{pmatrix}\right| = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$(1) \Rightarrow (2-\lambda)(1-\lambda) - 3(0) = 0$$

$$\Rightarrow (2-\lambda)(1-\lambda) = 0 \Rightarrow \begin{cases} 2-\lambda = 0 \\ 1-\lambda = 0 \end{cases}$$

d'où, les valeurs propres de $A_{(2,2)}$ sont $\lambda_1 = 2$ et , $\lambda_2 = 1$

Remarque 1.3.

Les somme des valeurs propres d'une matrice $A_{(p,p)}$ est égale à la somme des pivots. Donc, on écrit :

$$\sum_{j=1}^{p} \lambda_j = traceA_{(p,p)}$$

Exemple

Soit la matrice A=$\begin{pmatrix}2 & 0\\ 3 & 1\end{pmatrix}$, alors $\lambda_1 = 2$ et , $\lambda_2 = 1$

$$\sum_{j=1}^{2} \lambda_j = 2 + 1 = 3$$

et $traceA_{(p,p)} = 2 + 1 = 3$

1.3.2. Vecteurs propres d'une matrice

Si $\lambda \in \mathbb{R}^*$ est la valeur propre de la matrice $A_{(p,p)}$, alors il existe un vecteur propre non nul U tel que : $AU = \lambda U$. Donc, on dit que U est un vecteur propre de $A_{(p,p)}$.

Exemple

Soit $A_{(2,2)} = \begin{pmatrix} 0 & -2 \\ -4 & 2 \end{pmatrix}$

On montre que 4 est une valeur propre de $A_{(2,2)}$ et on cherche le vecteur propre qui en associe.

Solution

$\lambda = 4$ est une valeur propre de A si et seulement si $\left| A_{(2,2)} - 4I_2 \right| = 0$

$\left| A_{(2,2)} - 4I_2 \right| = \left| \begin{pmatrix} 0 & -2 \\ -4 & 2 \end{pmatrix} - 4\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = \left| \begin{pmatrix} -4 & -2 \\ -4 & -2 \end{pmatrix} \right|$= 8 - 8= 0. donc $\lambda = 4$ est une valeur propre de A

- **Calcul du vecteur propre associé à $\boldsymbol{\lambda = 4}$**

Soit $U = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ vecteur propre de A associé à $\boldsymbol{\lambda = 4}$**.** Donc**,** $AU = \lambda U \Rightarrow (AU - \lambda U) = \underline{0_2}$

$$\Rightarrow (A - \lambda I)U = \underline{0_2}$$

$$\Rightarrow \left[\begin{pmatrix} 0 & -2 \\ -4 & 2 \end{pmatrix} - 4\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} -4 & -2 \\ -4 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \begin{cases} -4u_1 - 2u_2 = 0 \\ -4u_1 - 2u_2 = 0 \end{cases} \Rightarrow \begin{cases} u_2 = -2u_1 \\ u_1 \in \mathbb{R}^* \end{cases}$$

Donc, $U = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ -2u_1 \end{pmatrix} = u_1 \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ où $u_1 \in \mathbb{R}^*$

Cela veut dire, tous les vecteurs non nuls sous forme $\begin{pmatrix} u_1 \\ -2u_1 \end{pmatrix}$ avec $u_1 \in \mathbb{R}^*$ sont des vecteurs propres de la matrice A associé à $\lambda = 4$. D'où, pour $u_1 = 1$, le vecteur propre associé à $\lambda = 4$ est égale à $\begin{pmatrix} 1 \\ -2 \end{pmatrix}$.

Remarque 7: Si $\lambda = 0$, alors le vecteur propres qui en associe n'existe pas

Vecteur propre unitaire

Soit U le vecteur propre de matrice $A_{(p,p)}$ associé à la valeur propre λ. On dit que U est vecteur propre unitaire si et seulement si : $\left\| \vec{U} \right\| = 1$

Autrement dit, soit $U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}$; alors $\| \vec{U} \| = 1 = \sqrt{\sum_{j=1}^{p} u_j^2} = 1$

Remarque 1.4.

Si $\| \vec{U} \| \neq 1$*, alors le vecteur propre U peut être rendu unitaire en divisant les coordonnées* u_j *par* $\| \vec{U} \|$*. On écrit :*

$$U_{unitaire} = \begin{pmatrix} u_1 / \| \vec{U} \| \\ u_2 / \| \vec{U} \| \\ \vdots \\ u_p / \| \vec{U} \| \end{pmatrix} \text{ et } \| \overrightarrow{U_{unitaire}} \| = \sqrt{\sum_{j=1}^{p} u_j^2 / \| \vec{U} \|^2} = 1$$

1.4. Analyse bidimensionnelle

Dans cette section, nous allons rappeler les principaux indicateurs de mesures qui permettent de décrire la relation entre deux variables. En effet, la nature des variables étudiées (qualitatives nominales, qualitatives ordinales, ou quantitatives)

1.4.1. Relation entre deux variables quantitatives

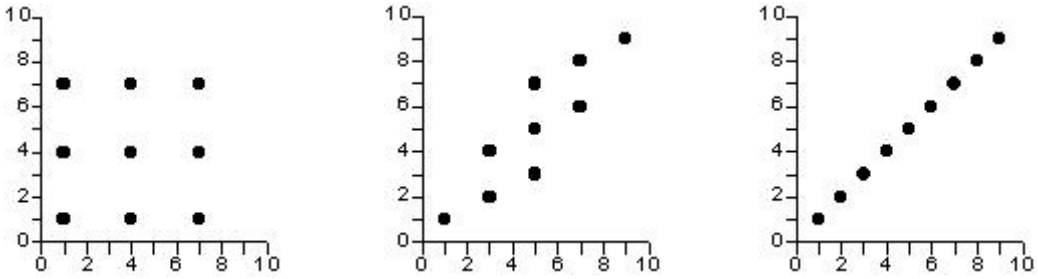
Soient deux variables quantitatives, X et Y, observées sur n individus, alors, nous avons deux copules d'observations (x_i, y_i) tel que $i = 1, \dots n$.

Donc, X et Y peuvent être récrites comme suit :

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \text{ et } Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

En effet, la représentation graphique de différents couples (x_i, y_i) forme un nuage de points qui permet de visualiser la nature de relation, et éventuellement la nature de tendance, entre les deux variables X et Y. Plusieurs formes de liaisons, corrélations, peuvent être identifiées (voir la Figure 1, X en abscisse et Y en ordonnée) :

1. Absence de liaison **2. relation linaire** **3. Forte relation linaire**



4. Relation non linéaire-non-monotone **5. Relation non linaire croissante**

Figure 1. Représentations graphiques des formes de relations entre deux variables quantitatives

1.4.2. Coefficient de corrélation linaire de Bravais-Pearson, noté $r_{X,Y}$

La relation linéaire entre deux variables quantitative X et Y s'identifie à travers le calcul du coefficient de corrélation de **Bravais-Pearson ($r_{X,Y}$).**

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sigma_X \sigma_Y}$$

Tels que :

n : nombre d'observations

$\bar{X}$: La moyenne arithmétique de X

$\bar{Y}$: La moyenne arithmétique de Y

$\sigma_X \; et \; \sigma_Y$: écarts-types de X et Y respectivement

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2} \; ; \sigma_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2}$$

Remarque 1.5.

- ✓ $-1 \leq r_{X,Y} \leq +1$
- ✓ $r_{X,Y} = 1$ *: Forte corrélation positive entre X et Y*
- ✓ $r_{X,Y} = -1$ *: Forte corrélation négative entre X et Y*
- ✓ $r_{X,Y} = 0$ *: absence de corrélation entre X et Y*
- ✓ *Absence de corrélation linaire linéaire ne signifie pas nécessairement l'indépendance entre X et Y*

Ci-après, des représentations graphiques illustrant la relation entre deux variables quantitatives et les coefficients de corrélation de **Bravais-Pearson** qui en découlent**.**

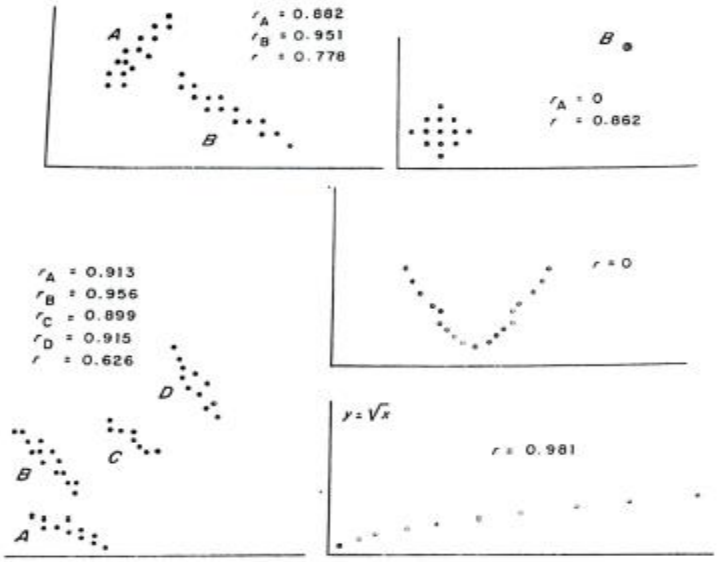
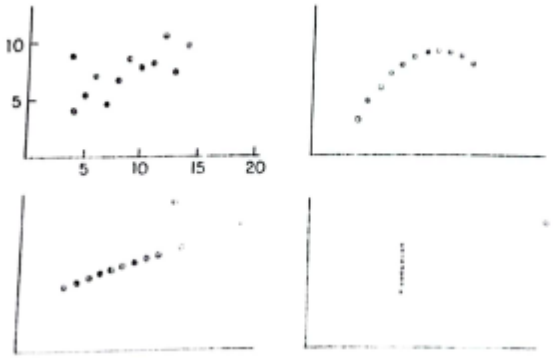


Figure 2. Coefficients de corrélations et corrélation entre deux variables quantitatives[1]

Dans le cas de la présence d'observations extrêmes, le coefficient de corrélation de Bravais-Pearson n'est pas un indice assez fiable pour identifier la nature de relation linéaire entre deux variables quantitatives (i.e. $r_{X,Y}$ n'est pas robuste, il est très sensible aux données aberrantes). Pour plus de détail, la Figure 3, extrait du livre de (Saborta, 2011), montre que les quatre relations mentionnées ci-dessous sont relatives à des couples de variables qui ont les mêmes moyennes, les mêmes variances, et le même coefficient de corrélation

[1] Ces graphiques sont extraits du livre Saborta G. (2011). Probabilités, analyse de données et statistiques. Editions TECHNIP.



$$\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4 = 9$$

$$\bar{Y}_1 = \bar{Y}_2 = \bar{Y}_3 = \bar{Y}_4 = 7{,}5$$

$$r_{X,Y} = 0{,}82$$

Figure 3. Sensibilité du coefficient de corrélation linéaire[2]

Exercice

Déterminer les valeurs propres de la matrice A

$$A = \begin{pmatrix} 3{,}5 & 0 & 0 \\ 0 & 5{,}2 & 0 \\ 0 & 0 & 6{,}9 \end{pmatrix}$$

Les valeurs propres de A

$$\left| A_{(3,3)} - \lambda I_3 \right| = 0 \Longrightarrow \begin{vmatrix} 3{,}5 - \lambda & 0 & 0 \\ 0 & 5{,}2 - \lambda & 0 \\ 0 & 0 & 6{,}9 - \lambda \end{vmatrix} = 0$$

$$\Longrightarrow (3{,}5 - \lambda)(\,5{,}2 - \lambda)(\,6{,}9 - \lambda) = 0$$

Donc, les valeurs propres de A sont : $\lambda_1 = 6{,}9\,;\ \lambda_2 = 5{,}2;\ \lambda_1 = 3{,}5$

Les vecteurs propres

Soit $U_1 = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ vecteur propre de A corréspodnat à λ_1**=6,9**

Nous avons :

[2] idem.

$$(A_{(3,3)} - \lambda I_3)\, U_1 = 0_3 \Longrightarrow \begin{pmatrix} 3{,}5-6{,}9 & 0 & 0 \\ 0 & 5{,}2-6{,}9 & 0 \\ 0 & 0 & 6{,}9-6{,}9 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Longrightarrow \begin{cases} -3{,}4u_1 = 0 \\ 1{,}7u_2 = 0 \\ 0u_3 = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} u_{1=0} \\ u_{2=0} \\ u_3 \in \mathbb{R} \end{cases}$$

D'où $U_1 = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = u_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$,

$U_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ est le vecteur propre de A associé à λ_1**=6,9**

U_1 est-il unitaire ?

$\left\| \overrightarrow{\boldsymbol{U_1}} \right\| = \sqrt{{u_1}^2 + {u_2}^2 + {u_3}^2} = \sqrt{0^2 + 0^2 + 1^2} = 1 \Longrightarrow U_1$ est un vecteur propre normé (unitaire)

Nous procédons de la même manière pour le cas des deux valeurs propres λ_2 et λ_3

$U_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $U_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

Chapitre 2. Analyse factorielle générale (AFG)

Soit $X_{(n,p)}$ une matrice relative à un tableau de données telles que : n et p désignent le nombre d'individus et des variables respectivement.

Donc ;

$$X_{(n,p)} = \begin{pmatrix} x_{11}\, x_{12} \,..\, x_{1j} \,.\, x_{1p} \\ x_{21}\, x_{22} \,..\, x_{2j} \,.\, x_{2P} \\ \vdots \quad \vdots \;\vdots\; \vdots \quad \vdots \\ x_{n1} \quad . \quad ..\, x_{nj} \,.\, x_{np} \end{pmatrix}$$

2.1. Analyse dans $\mathbb{R}^P$

Dans $\mathbb{R}^P$, la matrice de données $X_{(n,p)}$ peut être considérée comme un nuage de n points (les n individus) dans un espace vectoriel de p dimensions (p variables) (i.e. analyse dans $\mathbb{R}^P$).

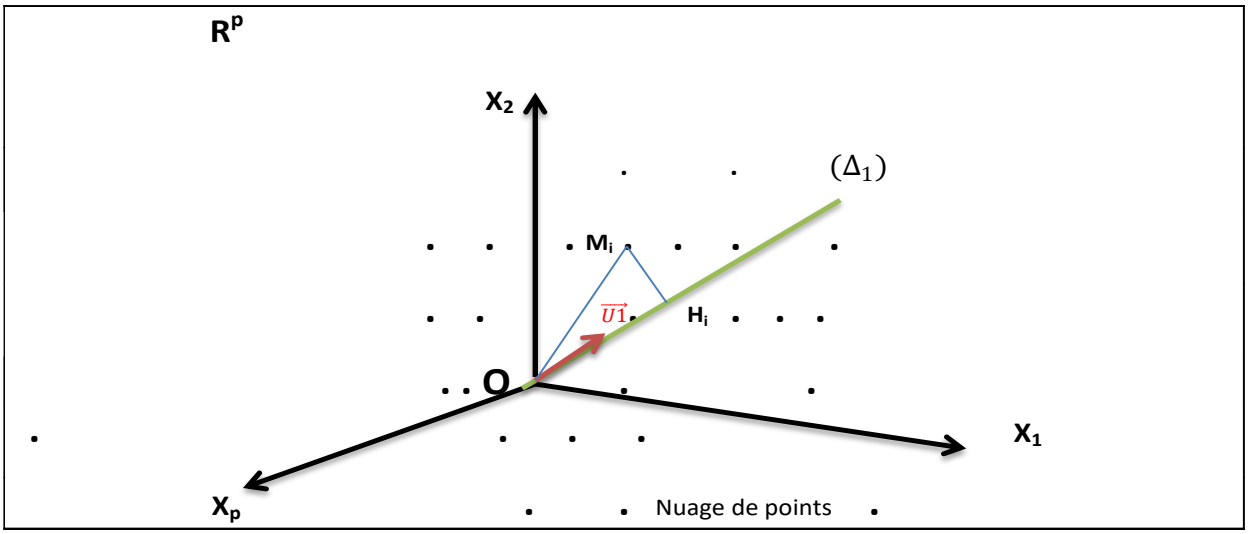


Figure 4. Nuage de points-individus

On cherche (Δ_1) qui soit plus proche de l'ensemble des points au sens des moindres carrées (i.e. Δ_1 est la droite d'ajustement du nuage des individus).

$\overrightarrow{OH_i}$: la projection orthogonale de $\overrightarrow{OM_i}$ sur (Δ_1) ⇒ rendre minimale $Q = \sum_{i=1}^{n} M_i H_i^2$

$[OH_i] \leq [OM_i]$

$\| \overrightarrow{OH_i} \| \leq \| \overrightarrow{OM_i} \|$

Selon le théorème de Pythagore, $OM_i^2 = OH_i^2 + M_iH_i^2$

Donc, pour l'ensemble des individus (n): $\sum_{i=1}^{n} OM_i^2 = \sum_{i=1}^{n} OH_i^2 + \sum_{i=1}^{n} M_iH_i^2$

$\sum_{i=1}^{n} OM_i$ étant fixe (dépend des données réelles), alors :

$Min \sum_{i=1}^{n} M_iH_i^2 \Rightarrow Max \sum_{i=1}^{n} OH_i^2$

Par ailleurs ; $\overrightarrow{OH_i}$ est la projection de $\overrightarrow{OM_i}$ sur $\overrightarrow{U_1}$ ($\overrightarrow{U_1}$ est le vecteur directeur de (Δ_1), $\overrightarrow{U_1}$ est unitaire), alors $\overrightarrow{OH_i} = \overrightarrow{OM_i}\overrightarrow{U_1} = x_iU_1 = \sum_{j=1}^{p} x_{ij}u_j$, (chaque individu (i) est représenté par un vecteur $X_i = (x_{i1}, x_{i2}, \ldots . x_{ij}, \ldots . , x_{ip})$)

Pour les n individus : $\sum_{i=1}^{n} OH_i = XU_1 = \begin{pmatrix} x_{11} & x_{12} & \vdots & x_{1j} & x_{1p} \\ x_{21} & x_{22} & \vdots & \vdots & x_{2P} \\ \vdots & \vdots & \vdots & x_{ij} & \vdots \\ x_{n1} & & & x_{nj} & x_{np} \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{1j} \\ u_{1p} \end{pmatrix} = \begin{pmatrix} \vdots \\ \sum_{j=1}^{p} x_{ij}u_{1j} \\ \vdots \end{pmatrix} = F_1$

On cherche à maximiser $\sum_{i=1}^{n} OH_i^2$

$\sum_{i=1}^{n} OH_i^2 = \| \vec{F}_1 \|^2 = F_1'F_1 = (XU_1)'(XU_1) = U_1'X'XU_1$. Donc, il suffit de trouver le vecteur U_1 qui maximise la quantité $U_1'X'XU_1$, c'est à dire :

$$\begin{cases} Max\ Q = U_1'X'XU_1 \\ S/C \\ U_1'U_1 = 1 \end{cases} \quad \ldots\ldots\ldots (1)$$

Pour optimiser le système (1), nous ferons appel à la méthode de multiplicateur de Lagrange ($L(U_1, \lambda_1)$).

$$\begin{cases} L(U_1, \lambda_1) = U_1'X'XU_1 - \lambda_1(U_1'U_1 - 1) \ldots \ldots (2) \\ S/C \\ U_1'U_1 = 1 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (3) \end{cases}$$

$$\begin{cases} \frac{dL(U_1, \lambda_1)}{dU_1} = 0 \\ U_1'U_1 = 1 \end{cases} \Rightarrow \begin{cases} \frac{dL(U_1, \lambda_1)}{dU_1} = 2X'XU_1 - 2\lambda_1U_1 = 0 \\ U_1'U_1 = 1 \end{cases}$$

$$\begin{cases} \frac{dL(U_1, \lambda_1)}{dU_1} = X'XU_1 = \lambda_1U_1 \ldots \ldots (4) \\ U_1'U_1 = 1 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (5) \end{cases}$$

A partir des équations (4) et (5), U_1 est le vecteur propre de la matrice $X'X$ associé à la valeur propre λ_1 ($X'X$ contient p valeurs propres puisque est une matrice de plein rang). En pré-multipliant (4) par U_1' on aura :

$$U_1'X'XU_1 = \lambda_1 U_1'U_1 \Rightarrow U_1'X'XU_1 = \lambda_1 \ldots\ldots\ldots(6)$$

Par ailleurs, nous cherchons à maximiser $U_1'X'XU_1$, ceci est vérifié quand la valeur propre λ_1 est maximale. Par conséquent, U_1 est le vecteur propre de la matrice $(X'X)$ associé à la plus grande valeur propre ($U_1'U_1 = 1, U_1$ est vecteur propre unitaire- normé à 1)

Remarque 2.1.

Le vecteur propre normé à 1 de la matrice $X'X$, associé à la plus grande valeur propre λ_1, est le vecteur directeur de la droite d'ajustement (Δ_1), appelée également le premier axe principale.

Nous cherchons à nouveau le deuxième vecteur directeur $\overrightarrow{U_2}$ de la droite (Δ_2) , deuxième axe principale, tel que ($\overrightarrow{U_1} \perp \overrightarrow{U_2}$)

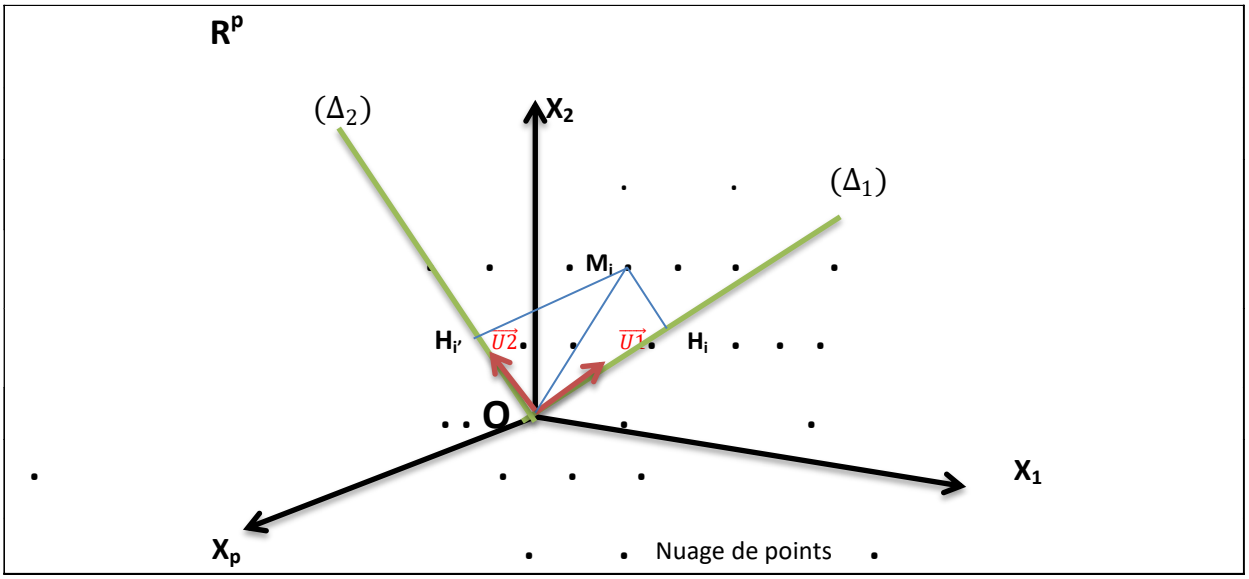


Figure 5. Nuage de points

Autrement dit, nous cherchons le sous espace vectoriel à deux dimensions($\overrightarrow{U_1}, \overrightarrow{U_2}$) qui soit le plus proche du nuage des individus. C'est-à-dire, nous cherchons $\overrightarrow{U_2}$ unitaire ($U_2'U_2 = 1$) et $\overrightarrow{U_1} \perp \overrightarrow{U_2}$ ($U_2'U_1 = 0$)

La projection orthogonale de M_i sur (Δ_2) :

$\overrightarrow{OH_i'} = \overrightarrow{OM_i}\,\overrightarrow{U_2}$ =$x_i U_2$ (avec $X_i = (x_{i1}, x_{i2}, \dots . x_{ij}, \dots . , x_{ip})$ est le vecteur-individus dans $\mathbb{R}^{\boldsymbol{p}}$**)**

$$\sum_{i=1}^{n} OH_i' = XU_2 = F_2$$

Par ailleurs ;

$F_2'F_2 = \sum_{i=1}^{n} OH_i'^2$ =$(XU_2)'(XU_2) = U_2'X'XU_2$

Donc,

$$\begin{cases} Max\ Q = \sum_{i=1}^{n} OH_i'^2 = U_2'X'XU_2 \\ S/C \\ U_2'U_2 = 1 \\ U_2'U_1 = 0 \end{cases} \quad \text{.....(SI)}$$

On utilise la fonction de Lagrange pour optimiser le système d'équation (SI) :

$L(U_2, \lambda_2, \mu) = U_2'X'XU_2 - \lambda_2(U_2'U_2 - 1) - \mu U_2'U_1$

$$\begin{cases} \dfrac{dL(U_2, \lambda_2, \mu)}{dU_2} = 0 \Rightarrow 2X'XU_2 - 2\lambda_2 U_2 - \mu U_1 = 0 \dots \dots \dots \dots \dots (7) \\ U_2'U_2 - 1 = 0 \\ U_2'U_1 = 0 \end{cases}$$

Montrons que $\mu = 0$?

En pré-multipliant l'équation (7) par U_1', on aura :

$U_1'(6) \Leftrightarrow 2U_1'X'XU_2 - 2\lambda_2 U_1'U_2 - \mu U_1'U_1 = 0$

$0 - 0 - \mu = 0 \Rightarrow \mu = 0$

d'où ; $(6) \Leftrightarrow 2X'XU_2 - 2\lambda_2 U_2 = 0$

$\Rightarrow X'XU_2 = \lambda_2 U_2$ ……(8)

En pré-multipliant (8) par U_2', on aura :

$$U_2'X'XU_2 = \lambda_2 U_2'U_2 \Rightarrow U_2'X'XU_2 = \lambda_2$$

Donc, U_2 est le vecteur propre de la matrice $X'X$ associé à la valeur propre λ_2. La quantité $U_2'X'XU_2$ est optimale lorsque λ_2 est la deuxième plus grande valeur propre de la matrice $(X'X)$.

Propriétés

- La matrice $(X'X)$ possède U_j vecteurs propres orthonormés associes aux valeurs propres λ_j rangées par ordre décroissant.
- le vecteur propre U_2 normé de la matrice $(X'X)$, associé à la deuxième valeur propre λ_2, est le vecteur directeur porté par le deuxième droite d'ajustement (Δ_2) (i.e. U_2 deuxième axe principal)
- Le sous espace vectoriel, à deux dimensions qui ajuste au mieux le nuage des individus au sens des moindres carrées, est généré par deux vecteurs propres $\overrightarrow{U_1}$ et $\overrightarrow{U_2}$ de la matrice $(X'X)$ correspondants à la plus grande, la 1iére, valeur propre λ_1, et à la deuxième valeur propre λ_2 respectivement.
- De façon générale, le sous espace de faible dimensions se détermine par :
  $U_\alpha'X'XU_\alpha = \lambda_\alpha$ tel que $\alpha = 1 \ldots . p$
  U_α vecteur propre de la matrice $X'X$ associé à la valeur propre λ_α (les valeurs propres sont rangées par un ordre décroissant). $\Longrightarrow U_\alpha$ le vecteur directeur, axe principal α, porté par la droite d'ajustement α.

Remarque 2.2.

Trois critères sont souvent utilisés pour choisir le nombre de dimensions du sous espace :

- *La part de l'inertie : On retient un nombre K de dimensions dont la part de l'inertie est maximale*
- *La règle de Kaiser : nous retenons que les valeurs propre qui ont une valeur supérieure à leurs moyennes ($\bar{\lambda} = \sum_{\alpha=1}^{p} \frac{\lambda_\alpha}{p}$) (pour une ACP normée, nous retenons les axes factoriels associent aux valeurs propres supérieures à 1)*
- *La règle du coude : les valeurs propres sont rangées par ordre décroissant. Donc, on retient les valeurs propres qui se trouvent avant le coude du graphique des valeurs propres en fonction de leurs indices j.*

2.2. Analyse dans $\mathbb{R}^n$

L'analyse dans $\mathbb{R}^n$ consiste à analyser la corrélation entre (p) variables. Pour ce faire, les (p) variables seront représentées graphiquement dans un espace vectoriel de (n) dimensions (n individus).

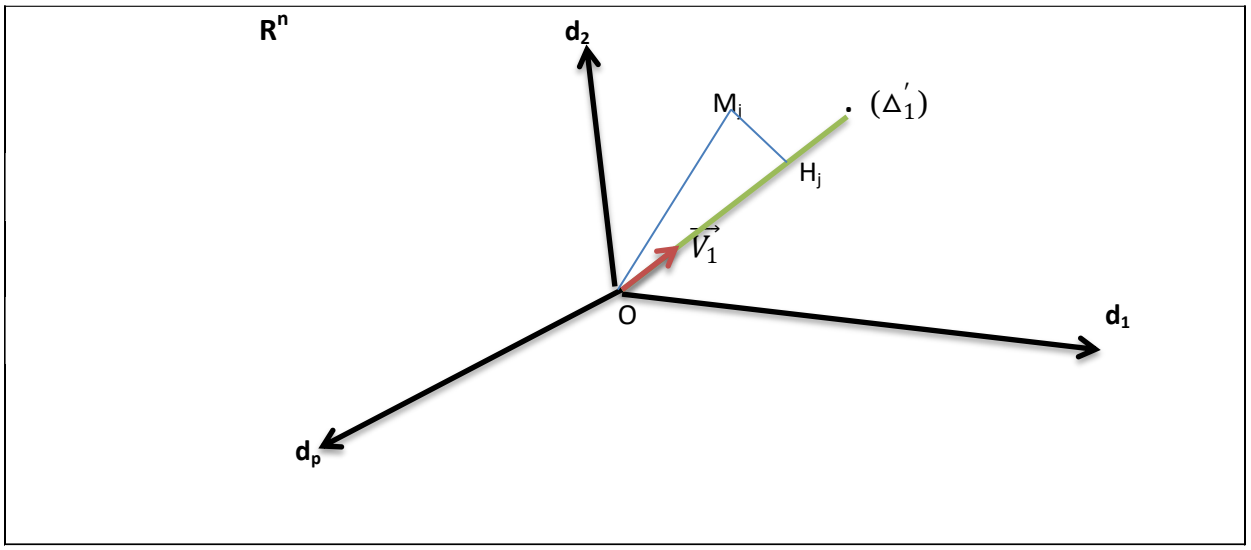


Figure 6. Nuage de points-variables

Remarque 2.3.

- ✓ *Dans la pratique, le nuage de points ne peut être représenté graphiquement que dans un plan (Sous espace vectoriel à deux dimensions). Le passage d'un espace vectoriel de n dimensions à un sous espace vectoriel à deux dimensions se fait en minimisant la perte d'information au sens des moindres carrées.*
- ✓ *La démarche que nous avions utilisée dans le cas d'analyse dans $\mathbb{R}^p$ sera appliquée à nouveau dans le cas d'analyse dans $\mathbb{R}^n$*

La projection orthogonale de $[OM_j]$ sur la droite d'ajustement (Δ_1') est $[OH_j]$; alors $\| \overrightarrow{OM_j} \| \leq \| \overrightarrow{OH_j} \|$

Soit $\overrightarrow{V_1}$ le vecteur directeur de la droite d'ajustement (Δ_1') $\Rightarrow$ $V_1'V_1 = (v_1, v_2, \dots., v_n)\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_3 \end{pmatrix} = \sum_{i=1}^{n} v_i^2 = 1$ ($\overrightarrow{V_1}$ est unitaire- normé)

Par ailleurs, $OH_j = (X^j)'V_1$ telle que X^j est la coordonnée de la variable j dans $\mathbb{R}^n$

$$(X^j)' = (x_{1j}, x_{2j}, \dots\dots\dots\dots, x_{nj})$$

- ✓ On note par G_1 les coordonnées de toutes les variables (p variables) sur (Δ_1'), alors :

$$G_1 = X'V_1 = \begin{pmatrix} OH_1 \\ OH_2 \\ \vdots \\ OH_j \\ \vdots \\ OH_p \end{pmatrix}$$

Exemple : OH_1 : coordonnée de la première variable (X^1) sur (Δ_1')

Par ailleurs ; selon le théorème de Pythagore, $\sum_{j=1}^{p} OM_j^2 = \sum_{j=1}^{p} OH_j^2 + \sum_{j=1}^{p} M_j H_j{}^2$

$\sum_{j=1}^{p} OM_j^2$ étant fixe, alors Max$\sum_{j=1}^{p} OH_j^2 \Rightarrow$ Min$\sum_{j=1}^{p} M_j H_j{}^2$

En outre, $G_1{}'G_1 = (OH_1, OH_2, \ldots., OH_p) \begin{pmatrix} OH_1 \\ OH_2 \\ \vdots \\ OH_j \\ \vdots \\ OH_p \end{pmatrix} = \sum_{j=1}^{p} OH_j^2$ …….(1)

et $G_1{}'G_1 = (X'V_1)'(X'V_1) = V_1'XX'V_1$……..(2)

de (1) et (2), nous déduisons que $\sum_{j=1}^{p} OH_j^2 = V_1'XX'V_1 = Q$

Donc,

$$\begin{cases} \text{Max Q} = V_1'XX'V_1 \\ S/C \\ V_1'V_1 = 1 \end{cases}$$

$$L(V_1) = V_1'XX'V_1 - \mu_1(V_1'V_1 - 1)$$

$$\frac{dL(V_1)}{dV_1} = 0 \Rightarrow 2\,XX'V_1 - 2\,\mu_1 V_1 = 0$$

$$\Rightarrow \begin{cases} XX'V_1 = \mu_1 V_1 \ldots\ldots.. (3) \\ S/C \\ V_1'V_1 = 1 \ldots\ldots\ldots. (4) \end{cases}$$

De (3) et (4), V_1 est le vecteur propre de la matrice XX' associé à la valeur propre μ_1

En pré-multipliant (3) par V_1', on obtient :

$$V_1'XX'V_1 = \mu_1 V_1'V_1 \Rightarrow V_1'XX'V_1 = \mu_1$$
$$S/C$$
$$V_1'V_1 = 1$$

D'où ; Max Q=$V_1' XX' V_1 \Rightarrow \mu_1$ est maximale

Donc, V_1 est le vecteur propre de la matrice XX' associé à la plus grande valeur propre μ_1(i.e. la première valeur propre μ_1)

Transition de $\mathbb{R}^p$ vers $\mathbb{R}^n$

Dans $\mathbb{R}^p$:

$X'XU_\alpha = \lambda_\alpha U_\alpha$ ……(1)

Dans $\mathbb{R}^n$:

$XX'V_\theta = \mu_\theta V_\theta$ ….(2)

On multiplie à gauche (2) par X' , on obtient : $(X'X)X'V_\theta = \mu_\theta V_\theta$…(3)

De (3), $X'V_\theta$ est vecteur propre de la matrice $(X'X)$ associé à la valeur propre μ_θ

De (2) et (3), $U_\alpha = X'V_\alpha$ et $\lambda_\alpha = \mu_\alpha$ pour $\alpha = 1, \dots \dots p$

U_α est-il unitaire ?

$\| \overrightarrow{U_\alpha} \| = \sqrt{U_\alpha' U_\alpha} = \sqrt{(X'V_\alpha)'(X'V_\alpha)} = \sqrt{V_\alpha' XX' V_\alpha} = \sqrt{\mu_\alpha} = \sqrt{\lambda_\alpha}$ pour $\alpha = 1, \dots \dots p$

Pour que U_α soit unitaire alors : $U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} XV_\alpha$ …(4)

De (4), on déduit **:** $U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} XV_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} G_\alpha \Longrightarrow G_\alpha = \sqrt{\lambda_\alpha} U_\alpha$ donne les coordonnées de la composante principale α

Remarque 2.4.

La diagonalisation de la X'X est suffisante pour déterminer les coordonnées des facteurs principaux, ainsi que celles des composantes principales.

Chapitre 3. Analyse en composantes principales (ACP)

- La méthode ACP a été conçue par Karl Pearson en 1931, elle est intégrée dans les sciences de mathématique et statistique par Harold Hotelling en 1933
- Le développement des méthodes et moyens de calculs, ainsi que les logiciels informatiques a impulsé l'utilisation de la méthode l'ACP.

Remarque 3.1.

- *La méthode ACP s'applique sur des données quantitatives*
- *On distingue deux types d'ACP : ACP-normée et ACP normée.*

3.1. Caractéristiques relatives aux tableaux de données quantitatives multidimensionnelles

a. Identification des variables et individus

Soit le tableau des données numériques $X_{(n,p)}$ de n individus (en lignes) et p variables (en colonnes) :

$$X_{(n,p)} = \begin{pmatrix} x_{11}\ x_{12}\ldots x_{1j}\ldots x_{1p} \\ x_{21}\ x_{22}\ldots x_{2j}\ldots x_{2P} \\ \vdots\ \ \vdots\ \vdots\ \vdots\ \vdots\ \vdots \\ x_{i1}\ x_{i2}\vdots\vdots x_{ij}\vdots\vdots x_{ip} \\ x_{n1}\ \ \vdots\ \ \vdots\ \vdots x_{nj}\vdots\vdots x_{np} \end{pmatrix}$$

Où x_{ij} est la valeur prise par la variable j sur l'individu i. De ce fait, l'identification matricielle des variables et des individus se fait de la manière suivante :

Variables : $X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$

Individus : $i = (x_{i1}\ x_{i2}\ \ldots\ldots x_{ip})$

b. Matrice des poids

Le calcul des caractéristiques de l'échantillon est étroitement lié au poids de chaque individu. Notons par p_i le poids correspondant à l'individu (i), alors : la matrice des poids (D) est une matrice diagonale de taille n, elle est définie comme suit :

$$D_{(n,n)} = \begin{pmatrix} P_1\ 0\ \ldots\ 0 \\ 0\ P_2\ldots\ 0 \\ \vdots\ \ \vdots\ \ldots\ \vdots \\ \vdots\ \ \vdots\ \vdots\ \vdots \\ 0\ \ 0\ \vdots P_n \\ \ \ \ \ \ \end{pmatrix}$$

Sous l'hypothèse selon laquelle tous les individus ont le même poids $1/n$ (les données sont tirées avec des probabilités égales = $1/n$), la matrice des poids sera récrite comme suit :

$D = \begin{pmatrix} 1/n & 0 & \cdots & 0 \\ 0 & 1/n & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/n \end{pmatrix} = \frac{1}{n} I_n$ où $I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ est la matrice identité

c. Centre de gravité (point moyen)

Le centre de gravité, noté g, du nuage de points est le vecteur des moyennes arithmétiques des P variables du tableau de données.

$g' = (\bar{X}_1, \bar{X}_2, \ldots \ldots \bar{X}_P) \Rightarrow g' = X'DI$ où $I = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ est le vecteur le R^n comportant que de 1

Propriétés

✓ Lorsque les individus qui ont le même poids, le centre de gravité (g') est donné par la formule suivante :

$$g' = X'DI = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{i1} & x_{n1} \\ x_{21} & x_{22} & \cdots & x_{i2} & x_{n2} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_{n1} & \vdots & \cdots & x_{nj} & x_{np} \end{pmatrix} \begin{pmatrix} \frac{1}{n} & 0 & \cdots & 0 \\ 0 & \frac{1}{n} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \Rightarrow \bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

✓ Soit $\hat{X}_{(n,p)} = X - Ig'$ le tableau centré associé à $X_{(n,p)}$, alors :

- $\hat{X}_{(n,p)} = X_{(n,p)} - Ig' = X_{(n,p)} - X'DI = (I - II'D)X$
- La matrice de variance-covariance (V)

$V = X'DX - gg' = \hat{X}'D\hat{X}$; avec : $X'DX = \sum_{i=1}^{n} p_i \, ii' = \sum_{i=1}^{n} \frac{1}{n} ii'$

✓ Soit $\tilde{X}_{(n,p)} = \frac{x_{ij} - \bar{X}_j}{s_{X^j}}$ avec s_{X^j}:l'écart type de la variable X^j, alors :

- $\tilde{X}_{(n,p)} = \hat{X} D_{1/s}$ est le tableau des données centrées et réduites associé à $X_{(n,p)}$ (avec $D_{1/s}$ la matrice diagonale où la diagonale principale contient les inverses des écarts types des p variables), donc : $D_{1/s} = \begin{pmatrix} 1/s_1 & 0 & \cdots & 0 \\ 0 & 1/s_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/s_j \end{pmatrix}$
- La matrice des coefficients de corrélation linéaire entre les p variables est la suivante : $C = D_{1/s} V D_{1/s} = \tilde{X}' D \tilde{X}$ (D est la matrice des poids)

Pour $P = 1/n \Longrightarrow D = \begin{pmatrix} 1/n & 0 & \cdots & 0 \\ 0 & 1/n & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/n \end{pmatrix}$

d. L'espace des individus

La présentation graphique des n individus se fait dans un espace vectoriel (F) à p dimensions (i.e. chaque individu est représenté par p coordonnées), ce qui donne la forme d'un nuage de points avec g' son centre de gravité. Par conséquent, la distance euclidienne entre deux individus (i, i') de cet espace vectoriel est donnée par la formule quadratique suivante :

$d^2(i,i') = (i-i')'(i-i')$ est le produit scalaire usuel, donc :

$d^2(i,i') = \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$ est la formule de la distance euclidienne classique entre deux points individus.

Cette formule permet de calculer la distance euclidienne classique. De ce fait, deux individus sont proches si leurs p coordonnées ont des valeurs numériques presque égales.

e. Espace des variables (p variables)-analyse dans $\mathbb{R}^n$

Dans un espace des variables de n diminuons ($\mathbb{R}^n$) , chaque variables est décrite par un vecteur de n données numériques.

Soit Cos $\theta_{j,k}$, l'angle entre deux variables centrées $(\hat{X}^j, \hat{X}^k)$, alors :

$$\text{Cos } \theta_{j,k} = \frac{\langle \hat{X}^j, \hat{X}^k \rangle}{\|\hat{X}^j\|\|\hat{X}^k\|} = \frac{(\hat{X}^{j\prime}\hat{X}^k)}{\sqrt{\hat{X}^{j\prime}\hat{X}^j}\sqrt{\hat{X}^{k\prime}\hat{X}^k}} = \frac{\sum_{i=1}^{n}\hat{x}_i^j\hat{x}_i^k}{\sqrt{\sum_{i=1}^{n}(\hat{x}_i^j)^2}\sqrt{\sum_{i=1}^{n}(\hat{x}_i^k)^2}} = \frac{Cov(\hat{X}^j,\hat{X}^k)}{s_{\hat{X}^j}s_{\hat{X}^k}} = r_{(\hat{X}^j,\hat{X}^k)}$$

Avec $r_{(\hat{X}^j,\hat{X}^k)}$: le coefficient de corrélation entre les deux variables centrées $\hat{X}^j, et\ \hat{X}^k$

f. Inertie

L'inertie totale d'un nuage de points se définit comme étant la moyenne pondérée des carrées des distances des points au centre de gravité

$$I_t = \sum_{i=1}^{n} p_i(i-g)'M(i-g) = \sum_{i=1}^{2} p_i\|i-g\|^2$$

Par ailleurs, $I_t = Trace\ MV = Trace\ VM$ (i.e. l'inertie totale est égale à le trace la matrice (MV ou VM). C'est pour cela qu'on distingue deux cas :

- Si $M = I, \Rightarrow \quad I_t = \sum_{j=1}^{p} V(X^j)$
- Si $M = D_{1/s^2} \Rightarrow Trace\,(MV) = Trace\,(D_{1/s^2}V) = Trace\,(D_{1/s}VD_{1/s}) = trace\,(C) = p$

3.2. Principe de la méthode ACP

La méthode ACP consiste à ajuster le nuage des n individus, observés sur p variables quantitatives, par un sous espace orthonormé de faible dimensions (généralement on choisit un sous espace à deux dimensions) toute en minimisant la perte d'information (i.e. le sous espace retenu permet visualiser la projection la plus fidèle possible des proximités entre les n individus observés sur les p variables). Pour ce faire, nous traçons une droite d'ajustement qui passe par le centre de gravité, elle n'est pas astreinte qu'il passe par l'origine comme le cas de l'analyse factorielle générale, et qu'elle soit plus proche de l'ensemble des points. Dans la Figure 7 ci-dessous,(Δ1) et la droite d'ajustement du nuage des individus (i.e. Δ1 est le premier axe du sous espace) .

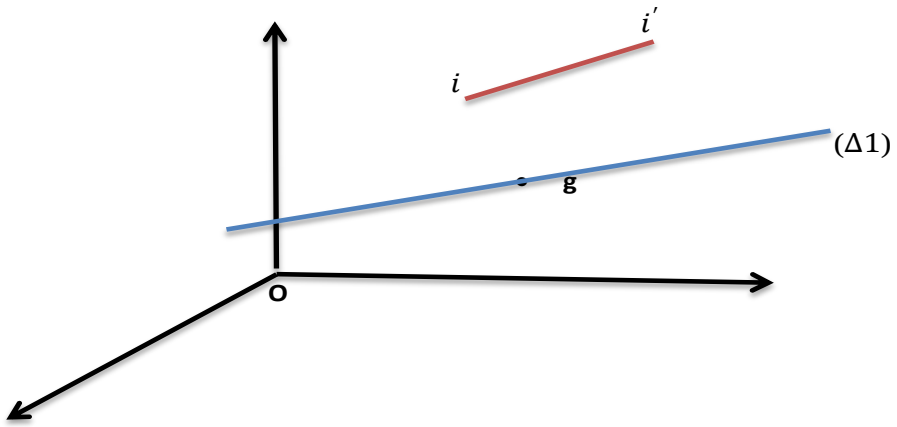


Figure 7. Nuage des points dans $\mathbb{R}^P$

La Figure 8 montre la projection orthogonale de deux individus (i, i') sur (Δ1), et h_i, et $h_{i'}$ représentent les valeurs de projection de deux points-individus (i, i').

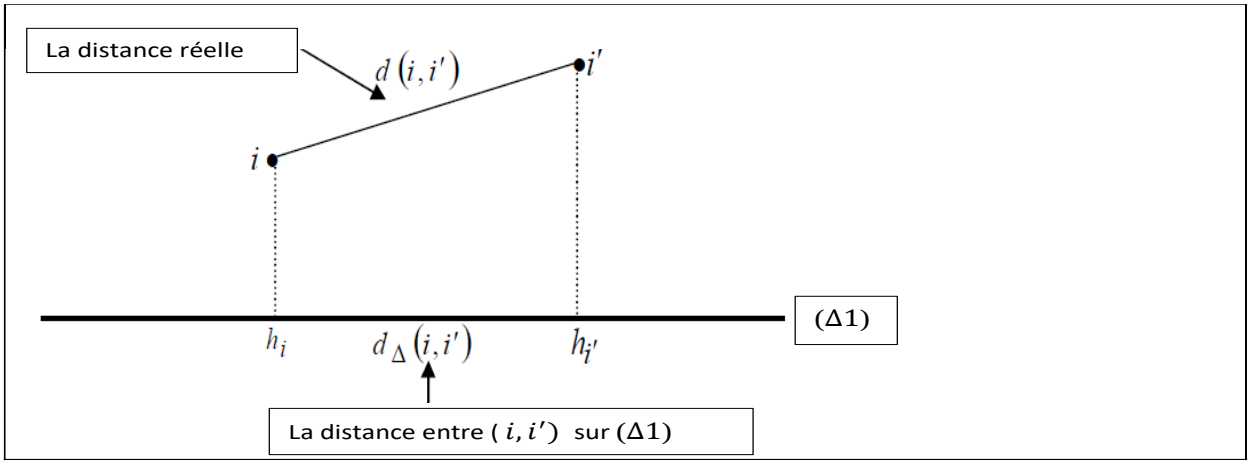


Figure 8. Distance projetée entre deux individus

Donc : $d^2_{\Delta 1}(i, i') \leq d(i, i'), \forall\ (i, i')$

La méthode ACP consiste à maximiser la distance projetée $\sum_{i,i'=1}^{n} d_{\Delta 1}^2(i,i') =$ pour que le sous espace de faible dimensions ajuste aux mieux le nuage des points-individus

En outre la formule classique qui permet de calculer la distance est toujours vérifiée. C'est-à-dire :

$$\sum_{i,i'=1}^{n} d_{\Delta 1}^2\left(i,i'\right) = = \sum_{i=1}^{n}\sum_{i'=1}^{n}\left(h_i - h_{i'}\right)^2$$

$$= n\sum_{i=1}^{n} h_i^2 - 2\sum_{i=1}^{n}\sum_{i'=1}^{n} h_i h_{i'} + n\sum_{i'=1}^{n} h_{i'=1}^2$$

$$= \left[n\sum_{i=1}^{n} h_i^2 - 2\sum_{i=1}^{n} h_i \sum_{i'=1}^{n} h_{i'} + n\sum_{i'=1}^{n} h_{i'}^2\right]$$

Sous l'hypothèse que deux individus (i, i') sont très voisins, alors les valeurs des variables sont égales ou presque égales. Donc, pour les n individus nous avons :

$\sum_{i=1}^{n} h_i^2 = \sum_{i'=1}^{n} h_{i'=1}^2$ et $\sum_{i=1}^{n} \frac{1}{n} h_i = \sum_{i'=1}^{n} \frac{1}{n} h_{i'} = \bar{h}$

Tel que $\sum_{i=1}^{n} \frac{1}{n} h_i = \sum_{i'=1}^{n} \frac{1}{n} h_{i'} = \bar{h}$ est la moyenne de projection (i. e. $\bar{h}$ correspond à la projection du centre de gravité g sur ($\Delta 1$))

D'où, $\sum_{i,i'=1}^{n} d_{\Delta 1}^2\left(i,i'\right) = n\sum_{i=1}^{n} h_i^2 - 2n^2 \sum_{i=1}^{n} \frac{1}{n} h_i \sum_{i'=1}^{n} \frac{1}{n} h_{i'} + n\sum_{i=1}^{n} h_i^2$

$$= n\sum_{i=1}^{n} h_i^2 - 2n^2 \bar{h}.\bar{h} + n\sum_{i'=1}^{n} h_i^2$$

$$= 2n(\sum_{i=1}^{n} h_i^2 - n\bar{h}^2)$$

$$= 2n\sum_{i=1}^{n}(h_i - \bar{h})^2$$

$$= 2n\sum_{i=1}^{n}(h_i - g)^2$$

$$\sum_{i,i'=1}^{n} d_{\Delta 1}^2\left(i,i'\right) = 2n\sum_{i=1}^{n} d_{\Delta 1}^2(i,g)$$

Donc ; $\text{Max}\{\sum_{i,i'=1}^{n} d_{\Delta 1}^2(i,i')\} \Leftrightarrow \text{Max}\{2n\sum_{i=1}^{n} d_{\Delta 1}^2(i,g)\} \Leftrightarrow \text{Max}\{\sum_{i=1}^{n}(i-g)^2\}$

Le centre de gravité étant fixe, la première droite d'ajustement ($\Delta 1$) du sous espace vectoriel sera tracée en maximisant la somme carrées des distances au centre de gravité. Pour ce faire, le centre de gravité de nuage de points sera pris comme origine de l'espace vectoriel et ce via la transformation des données en les rendant centrées par rapport à leurs moyennes respectives, comme suit :

$\hat{X}_{(n,p)} = X_{(n,p)} - g'$. Donc, le tableau des données s'écrit comme suit:

$$\hat{X}_{(n,p)} = \begin{pmatrix} x_{11}-\bar{X}_1 & x_{12}-\bar{X}_2 & \cdots & x_{1j}-\bar{X}_j & \cdots & x_{1p}-\bar{X}_p \\ x_{21}-\bar{X}_1 & x_{22}-\bar{X}_2 & \cdots & x_{2j}-\bar{X}_j & \cdots & x_{2P}-\bar{X}_p \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1}-\bar{X}_1 & \cdots & & x_{nj}-\bar{X}_j & & x_{np}-\bar{X}_p \end{pmatrix}$$

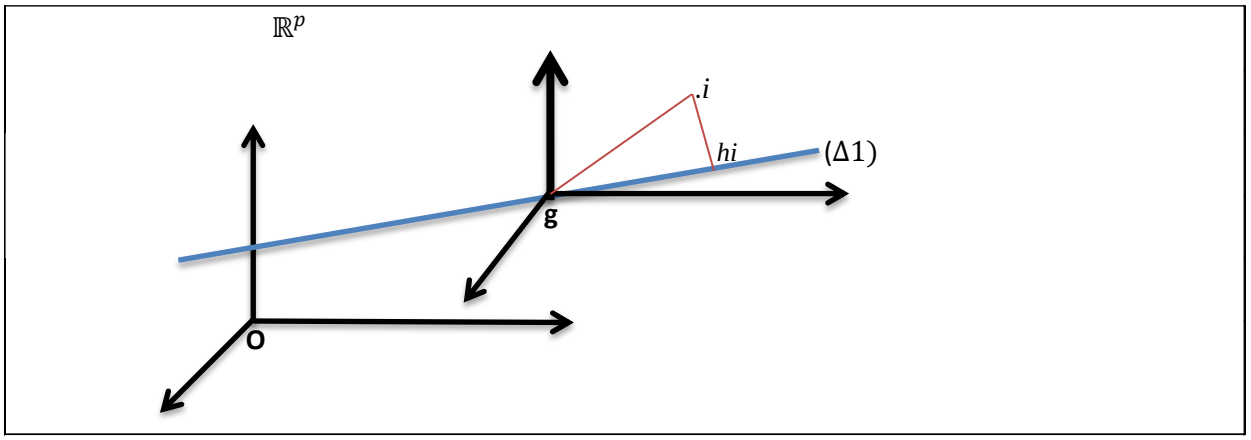


Figure 9. Changement d'origine

Par conséquent, le sous espace qui ajuste au mieux le nuage de points sera identifié en utilisant l'analyse factorielle générale (AFG) du tableau des données centrées $\hat{X}_{(n,p)}$.

D'où ;

$$\sum_{i=1}^{n} d^2(g,i) = \sum_{i=1}^{n} d^2(g,h_i) + \sum_{i=1}^{n} d^2(i,h_i)$$

$\sum_{i=1}^{n} d^2(g,i)$ Étant fixe, alors $\text{Min}\{\sum_{i=1}^{n} d^2(i,h_i)\} \Longrightarrow Max\{\sum_{i=1}^{n} d^2(g,h_i)\}$

Donc, le sous espace à deux dimensions sera identifié comme il a été démontré dans l'analyse factorielle générale (voir Chap. 2).

Remarque 3.2.

Selon les transformations faites sur les variables initiales, nous distinguons deux types d'ACP, il s'agit d'ACP non-normée et ACP normée

3.2.1. ACP non-normée (ACP simple)

ACP non-normée s'applique sur la matrice des données centrées pré-multipliées par une matrice des poids (les données ne sont pas pondérées, donc on prend la matrice $D_{1/\sqrt{n}}$) $\Longrightarrow$ le tableau des données Z s'écrit comme suit :

$$Z = D_{1/\sqrt{n}}\hat{X} = \begin{pmatrix} 1/\sqrt{n} & 0 & 0 \\ 0 & 1/\sqrt{n} & 0 \\ 0 & 0 & 1/\sqrt{n} \end{pmatrix} \begin{pmatrix} x_{11}-\bar{X}_1 & x_{12}-\bar{X}_2 & x_{1j}-\bar{X}_j & x_{1p}-\bar{X}_p \\ x_{21}-\bar{X}_1 & x_{22}-\bar{X}_2 & x_{2j}-\bar{X}_j & x_{2p}-\bar{X}_p \\ & & x_{ij}-\bar{X}_j & \\ x_{n1}-\bar{X}_1 & & x_{nj}-\bar{X}_j & x_{np}-\bar{X}_p \end{pmatrix}$$

$$Z = \begin{pmatrix} 1/\sqrt{n} & 0 & 0 \\ 0 & 1/\sqrt{n} & 0 \\ 0 & 0 & 1/\sqrt{n} \end{pmatrix} \begin{pmatrix} \hat{x}_{11} & \hat{x}_{12} & \hat{x}_{1j} & \hat{x}_{1p} \\ \hat{x}_{21} & \hat{x}_{22} & \hat{x}_{2j} & \hat{x}_{2p} \\ & & \hat{x}_{ij} & \\ \hat{x}_{n1} & & \hat{x}_{nj} & \hat{x}_{np} \end{pmatrix}$$

Tel que les éléments de la matrice Z: $z_{ij} = \frac{1}{\sqrt{n}}(x_{ij} - \bar{X}_j) = \frac{1}{\sqrt{n}}\hat{x}_{ij}$

Donc, la matrice à diagonaliser prend la forme suivante : $A = Z'Z = \hat{X}'D\hat{X} = \frac{1}{n}\hat{X}'\hat{X} = V$ (V la matrice des variances-covariances).

D'où ; $V = \frac{1}{n}\hat{X}'\hat{X} = \begin{pmatrix} V(X_1) & Cov(X_2,X_1) & & Cov(X_p,X_1) \\ Cov(X_1,X_2) & V(X_2) & V(X_3) & \\ Cov(X_1,X_p) & & & V(X_p) \end{pmatrix}$

Trace (V)= $V(X_1) + V(X_2) + V(X_p)$

a. Analyse dans $\mathbb{R}^p$

Détermination des axes principaux et des facteurs principaux du sous espace

Axes principaux

1ier axe principale

La matrice des variances-covariances ($V = \frac{1}{n}\hat{X}'\hat{X} = Z'Z$) sera soumise à l'analyse en composantes principales, ce qui permettra de déterminer les axes principaux du sous espace. On va résoudre le système (I) pour déterminer le premier axe principal.

$$\begin{cases} Max\ \left\{U_1'(\frac{1}{n}\hat{X}'\hat{X})U_1\right\} \\ S/_C\ U_1'U_1 - 1 = 0 \end{cases} \Leftrightarrow \begin{cases} Max\ \{U_1'Z'ZU_1\} \\ S/_C\ U_1'U_1 - 1 = 0 \end{cases} \quad (I)$$

Où U_1est le vecteur directeur porté par la droite d'ajustement (Δ1) ($\|\ \vec{U}_1\| = 1$)

Nous avons, $(\frac{1}{n}\hat{X}'\hat{X})\ U_1 = \lambda_1 U_1 \Leftrightarrow Z'ZU_1 = \lambda_1 U_1$, alors :

- ✓ U_1 est vecteur propre de la matrice $Z'Z$ associé à la plus grande valeur propre λ_1.
- ✓ U_1 est vecteur porté par la droite (Δ1)$\Longrightarrow U_1$ est appelé premier axe principale

2ime axe principale

On va solutionner le système (II) pour déterminer le 2ième axe principal (comme il a été démontré dans l'analyse factorielle générale -voir chap.2)

$$\begin{cases} Max\ \left\{U_2'(\frac{1}{n}\hat{X}'\hat{X})U_2\right\} \\ S/_C\ U_2'U_2-1=0 \\ U_2 \perp U_1 \end{cases} \Leftrightarrow \begin{cases} Max\ \{U_2'Z'ZU_2\} \\ S/_C\ U_2'U_2-1=0 \\ U_2'U_1=0 \end{cases} \quad \text{(II)}$$

Où U_2est le vecteur directeur porté par la droite d'ajustement ($\Delta 2$) (tel que $U_2 \perp U_1$)

($Z'ZU_2=\lambda_2U_2=(\frac{1}{n}\hat{X}'\hat{X})\ U_2=\lambda_2U_2 \Leftrightarrow VU_2=\lambda_2U_2$ alors :

- ✓ U_2 est vecteur propre de la matrice $Z'Z$ associé à la deuxième valeur propre λ_2.
- ✓ U_2 est vecteur porté par la droite ($\Delta 2$)$\Longrightarrow U_2$ est le deuxième axe principal

Facteurs principaux

Les coordonnées des facteurs principaux s'écrivent :

$$F_1=\hat{X}U_1=\begin{pmatrix} \hat{x}_{11} & \hat{x}_{12} & \cdots & \hat{x}_{1j} & \hat{x}_{1p} \\ \hat{x}_{21} & \hat{x}_{22} & \cdots & \hat{x}_{2j} & \hat{x}_{2p} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \hat{x}_{n1} & \cdots & \cdots & \hat{x}_{nj} & \hat{x}_{np} \end{pmatrix}\begin{pmatrix} u_{11} \\ u_{12} \\ \cdot \\ \cdot \\ u_{1p} \end{pmatrix}=\begin{pmatrix} \sum_{j=1}^{p}\hat{x}_{1j}u_{1j} \\ \vdots \\ \sum_{j=1}^{p}\hat{x}_{ij}u_{1j} \\ \vdots \\ \sum_{j=1}^{p}\hat{x}_{nj}u_{1j} \end{pmatrix} \quad 1^{er}\text{ facteur principal}$$

$$F_2=\hat{X}U_2=\begin{pmatrix} \hat{x}_{11} & \hat{x}_{12} & \cdots & \hat{x}_{1j} & \hat{x}_{1p} \\ \hat{x}_{21} & \hat{x}_{22} & \cdots & \hat{x}_{2j} & \hat{x}_{2p} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \hat{x}_{n1} & \cdots & \cdots & \hat{x}_{nj} & \hat{x}_{np} \end{pmatrix}\begin{pmatrix} u_{21} \\ u_{22} \\ \cdot \\ \cdot \\ u_{2p} \end{pmatrix}=\begin{pmatrix} \sum_{j=1}^{p}\hat{x}_{1j}u_{2j} \\ \vdots \\ \sum_{j=1}^{p}\hat{x}_{ij}u_{2j} \\ \vdots \\ \sum_{j=1}^{p}\hat{x}_{nj}u_{2j} \end{pmatrix} \quad 2^{\text{ième}}\text{ facteur principal}$$

Donc, la formule générale qui permet de calculer les coordonnées des facteurs principaux s'écrit :

$$F_\alpha=\hat{X}U_\alpha=\begin{pmatrix} \sum_{j=1}^{p}\hat{x}_{1j}u_{\alpha j} \\ \vdots \\ \sum_{j=1}^{p}\hat{x}_{ij}u_{\alpha j} \\ \vdots \\ \sum_{j=1}^{p}\hat{x}_{nj}u_{\alpha j} \end{pmatrix}\ \alpha^{ième}\text{ facteur principal } (\alpha=1,\ldots.p)\ (U_\alpha \text{ le vecteur propre de la matrice } V$$

associé à la valeur propre λ_α)

b. Analyse dans $\mathbb{R}^n$

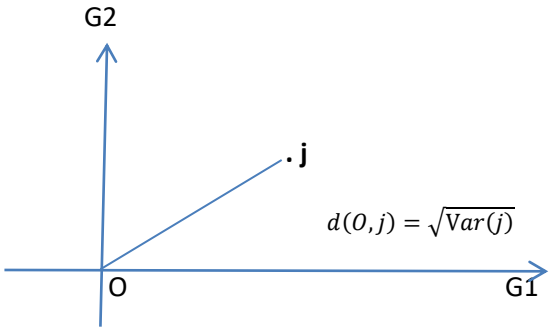
Le nuage des points-variables est un espace vectoriel à n dimensions (les variables sont représentées en fonction des individus). Comme il a été exposé ci-dessus, l'analyse dans $\mathbb{R}^n$ consiste à projeter le nuage des points-variables sur un sous espace de faible dimensions tout en conservant le mieux possible les angles, les corrélations, entre les variables.

Donc, l'analyse matrice X des données multidimensionnelles $\Longrightarrow$ l'analyse factorielle du tableau des données $Z=\begin{pmatrix} 1/\sqrt{n} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{n} \end{pmatrix}\begin{pmatrix} \hat{x}_{11} & \hat{x}_{12} & \cdots & \hat{x}_{1j} & \hat{x}_{1p} \\ \hat{x}_{21} & \hat{x}_{22} & \cdots & \hat{x}_{2j} & \hat{x}_{2p} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \hat{x}_{i1} & \hat{x}_{i2} & \cdots & \hat{x}_{ij} & \hat{x}_{ip} \\ \hat{x}_{n1} & \hat{x}_{n2} & \cdots & \hat{x}_{nj} & \hat{x}_{np} \end{pmatrix}=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1j} & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2j} & z_{2p} \\ \vdots & \vdots & \cdots & z_{ij} & \vdots \\ z_{n1} & \vdots & \cdots & z_{nj} & z_{np} \end{pmatrix}$

- ✓ La distance d'une variable par rapport à l'origine des axes :

$$d^2(O,j)=\sum_{i=1}^{n}(z_{ij}-\bar{z}_j)^2=\sum_{i=1}^{n}{z_{ij}}^2=\mathrm{V}ar(j)=\frac{1}{n}\sum_{i=1}^{n}\hat{x}_{ij}=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{X}_j)$$

$$\Longrightarrow d(O,j)=\sqrt{\mathrm{V}ar(j)}$$



- ✓ La distance entre deux variables (j,j') s'écrit comme suit :

$$d^2(j,j')=\sum_{i=1}^{n}(z_{ij}-z_{ij'})^2=\sum_{i=1}^{n}{z_{ij}}^2+\sum_{i=1}^{n}{z_{ij'}}^2-2\sum_{i=1}^{n}z_{ij}\,z_{ij'}$$

Sachant que $z\_ij=\frac{1}{\sqrt{n}}\,\hat{x}_{ij}$; alors $d^2(j,j')=Var(j)+Var(j')-2Cov(j,j')$

D'où :

- $d(j,j')\ \uparrow$, si $Cov(j,j')<0$
- $d(j,j')\ \downarrow$, si $Cov(j,j')>0$
- $d^2(j,j')=Var(j)+Var(j')$, si $Cov(j,j')=0$

Remarque 3.3.

Dans l'analyse dans $\mathbb{R}^{n}$, les distances entre deux variables se calculent par rapport à l'origine

Composantes principale s

Nous avons : $G_\theta = Z'V_\theta$ tel que V_θ est vecteur propre unitaire de la matrice ZZ' associé à la valeur propre μ_θ

Comme la matrice $Z'Z$ a déjà été diagonalisée dans l'analyse $\mathbb{R}^{p}$, alors les coordonnées des composante principale α sont écrites comme suit :

$G_\alpha = Z'V_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z'ZU_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Z'F_\alpha = \sqrt{\lambda_\alpha} U_\alpha$ tel que $\alpha = 1, \dots \dots p$

3.2.2. ACP normée

L'ACP normée s'applique dans le cas d'un tableau de données rectangulaire avec des données quantitatives exprimées dans différentes unités de mesure. Dans ce cas, toutes les variables seront transformées pour les rendre centrées et réduites pour qu'elles soient indépendantes des unités de mesure (i.e. supprimer l'effet unité de mesure). Pour ce faire, chaque variable de la matrice $X_{(n,p)}$ sera transformée en retranchant sa moyenne et en divisant par son écart type. Par conséquent, les distances calculées ne dépendent plus des unités de mesure, ce qui veut dire toutes les variables ont la même importance quelle que soit sa dispersion. Les termes de la matrice $X_{(n,p)}$ s'écrivent comme suit :

$\tilde{x}_{ij} = \frac{x_{ij} - \bar{X}_j}{s_j}$ avec $\bar{X}_j$ la moyenne de la variable X^j, $s_j = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{X}_j)^2}$ et l'écart type de la variable X^j.

Donc, la matrice des données centrées et réduites s'écrit:

$$\tilde{X}_{(n,p)} = \begin{pmatrix} \frac{x_{11}-\bar{X}_1}{s_1} & \frac{x_{12}-\bar{X}_2}{s_2} & \cdots & \frac{x_{ij}-\bar{X}_j}{s_j} & \cdots & \frac{x_{1p}-\bar{X}_p}{s_p} \\ \frac{x_{21}-\bar{X}_1}{s_1} & \frac{x_{22}-\bar{X}_2}{s_2} & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \cdots & \frac{x_{np}-\bar{X}_p}{s_p} \\ \frac{x_{n1}-\bar{X}_1}{s_1} & \frac{x_{n2}-\bar{X}_2}{s_2} & \cdots & \cdots & \cdots & \cdots \end{pmatrix} = \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{ij} & \cdots & \tilde{x}_{1p} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \cdots & \tilde{x}_{np} \\ \tilde{x}_{n1} & \tilde{x}_{n2} & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

$\tilde{X}_{(n,p)}$: Matrice des données centrées et réduites

Donc, les échelles seront corrigées ce qui conduit à définir une nouvelle matrice R qui est le résultat du produit de la métrique $D_{1/\sqrt{n}}$, (la matrice des poids), et la matrice des données centrées et réduites $\tilde{X}_{(n,p)}$.

$$R = D_{1/\sqrt{n}}\tilde{X}_{(n,p)} = \begin{pmatrix} 1/\sqrt{n} & 0 & 0 \\ 0 & 1/\sqrt{n} & 0 \\ 0 & 0 & 1/\sqrt{n} \end{pmatrix} \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \tilde{x}_{1j} & \tilde{x}_{1p} \\ \tilde{x}_{21} & \tilde{x}_{22} & \tilde{x}_{2j} & \tilde{x}_{2p} \\ \tilde{x}_{n1} & \tilde{x}_{n2} & \tilde{x}_{nj} & \tilde{x}_{np} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}}\tilde{x}_{11} & & \frac{1}{\sqrt{n}}\tilde{x}_{1p} \\ & \frac{1}{\sqrt{n}}\tilde{x}_{ij} & \\ \frac{1}{\sqrt{n}}\tilde{x}_{n1} & & \frac{1}{\sqrt{n}}\tilde{x}_{np} \end{pmatrix}$$

$$R = \begin{pmatrix} R_{11} & & & & & R_{1p} \\ & \ddots & & & \ddots & \\ & & R_{ij} & & & \\ & \ddots & & & \ddots & \\ R_{n1} & & & & & R_{np} \end{pmatrix}$$

Donc ; le terme général de la matrice R s'écrit : $R_{ij} = \frac{(x_{ij} - \bar{X}_j)}{s_j\sqrt{n}}$

a. Analyse dans $\mathbb{R}^p$

La distance entre deux individus (i, i') s'écrit comme suit :

$$d^2_{\Delta 1}\left(i, i'\right) = \sum_{i=1}^{p} \frac{\left(x_{ij} - x_{i'j}\right)^2}{\sqrt{n}\, s_j}$$

Dans l'analyse $\mathbb{R}^p$, c'est le tableau des données transformées R qui sera soumis à l'analyse factorielle générale (AFG). Par conséquent, la matrice à diagonaliser est la matrice des corrélations .

$$C = R'R = \frac{1}{n}\tilde{X}'\tilde{X} =$$

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\tilde{x}_{11} & & \frac{1}{\sqrt{n}}\tilde{x}_{1p} \\ & \frac{1}{\sqrt{n}}\tilde{x}_{ij} & \\ \frac{1}{\sqrt{n}}\tilde{x}_{n1} & & \frac{1}{\sqrt{n}}\tilde{x}_{np} \end{pmatrix}' \begin{pmatrix} \frac{1}{\sqrt{n}}\tilde{x}_{11} & & \frac{1}{\sqrt{n}}\tilde{x}_{1p} \\ & \frac{1}{\sqrt{n}}\tilde{x}_{ij} & \\ \frac{1}{\sqrt{n}}\tilde{x}_{n1} & & \frac{1}{\sqrt{n}}\tilde{x}_{np} \end{pmatrix} = \begin{pmatrix} 1 & r(X_1,X_2) & r(X_1,X_j) & r(X_1,X_p) \\ r(X_2,X_1) & 1 & & \\ & & 1 & \\ r(X_p,X_1) & r(X_p,X_2) & & 1 \end{pmatrix}$$

$$Trace\,(C) = p$$

Le terme général de la matrice C : $c_{jj'} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}}\tilde{x}_{ij}\frac{1}{\sqrt{n}}\tilde{x}_{ij'} = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{ij}\,\tilde{x}_{ij'} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{ij} - \bar{X}_j)(x_{ij'} - \bar{X}'_j)}{s_j s_{j'}}$

$c_{jj'}=\frac{Cov(X^j,X^{j'})}{s_j s_{j'}} = r(X^j,\ X^{j'})$ (coefficient de corrélation linéaire entre deux variables (j,j'))

Axes principaux et facteurs principaux

Nous avons :

$$\begin{cases} Max\ \{U_1' R' R U_1\} \\ S/_C\ U_1' U_1 - 1 = 0 \end{cases} \Longrightarrow\ R'RU_1=\lambda_1 U_1$$

D'où U_1 vecteur propre de $R'R$ associé à la valeur propre λ_1 (i.e U_1 est le premier axe principal porté par la première droite d'ajustement ($\Delta 1$)

Les coordonnées des facteurs principaux α s'écrivent :

$$F_\alpha = RU_\alpha \quad \text{tel que } \alpha = 1, \dots p \Longrightarrow F_\alpha = \begin{pmatrix} \frac{1}{\sqrt{n}}\tilde{x}_{11} & \cdots & \frac{1}{\sqrt{n}}\tilde{x}_{1p} \\ \vdots & \frac{1}{\sqrt{n}}\tilde{x}_{ij} & \vdots \\ \frac{1}{\sqrt{n}}\tilde{x}_{n1} & \cdots & \frac{1}{\sqrt{n}}\tilde{x}_{np} \end{pmatrix} \begin{pmatrix} u_{\alpha 1} \\ . \\ u_{\alpha j} \\ . \\ u_{\alpha p} \end{pmatrix} = \begin{pmatrix} . \\ . \\ \sum_{j=1}^{p} u_{\alpha j} \frac{1}{\sqrt{n}}\tilde{x}_{ij} \\ . \\ . \end{pmatrix}$$

$$= \begin{pmatrix} . \\ . \\ \sum_{j=1}^{p} u_{\alpha j} \frac{(x_{ij}-\bar{X}_j)}{s_j\sqrt{n}} \\ . \\ . \end{pmatrix}$$

$\bar{F}_\alpha = 0$ (car les données sont centrés par rapport à leurs moyennes)

$Var(F_\alpha) = \lambda_\alpha$

b. Analyse dans $\mathbb{R}^n$

Dans $\mathbb{R}^n$, la distance entre deux variables $(j\ ,j')$ se calcule par rapport à l'origine du nuage des points-variables.

$$d^2(j,j') = \sum_{i=1}^{n}(R_{ij} - R_{ij'})^2 = \sum_{i=1}^{n} {R_{ij}}^2 + \sum_{i=1}^{n} {R_{ij'}}^2 - 2\sum_{i=1}^{n} R_{ij}\, R_{ij'}$$

Comme $R_{ij} = \frac{(x_{ij}-\bar{X}_j)}{s_j\sqrt{n}}$, alors :

$$d^2(j,j') = \sum_{i=1}^{n}(\frac{x_{ij}-\bar{X}_j}{s_j\sqrt{n}})^2 + \sum_{i=1}^{n}(\frac{x_{ij'}-\bar{X}_{j'}}{s_j\sqrt{n}})^2 - 2\sum_{i=1}^{n}(\frac{x_{ij}-\bar{X}_j}{s_j\sqrt{n}})(\frac{x_{ij'}-\bar{X}_j}{s_j\sqrt{n}})$$

$$=\sum_{i=1}^{n}\frac{(x_{ij}-\bar{X}_j)^2}{ns_j^2}+\sum_{i=1}^{n}\frac{(x_{ij'}-\bar{X}_{j'})^2}{ns_{j'}^2}-2\sum_{i=1}^{n}\frac{(x_{ij}-\bar{X}_j)(x_{ij'}-\bar{X}_{j'})}{ns_j s_{j'}}$$

$$=\sum_{i=1}^{n}\frac{s_j^2}{s_j^2}+\sum_{i=1}^{n}\frac{s_{j'}^2}{s_{j'}^2}-2\frac{Cov(j,j')}{s_j s_{j'}}$$

$$=1+1-2\,r(j,j')\Longrightarrow d^2(j,j')=2[1-r(j,j')]$$

Remarque 3.4.

- $r(j,j')$: *le coefficient de corrélation linéaire* $\Longrightarrow -1\leq r(j,j')\leq +1$,
- *Si* $r(j,j')=-1\Longrightarrow d^2(j,j')=+4\Longrightarrow$ *les deux variables* (j,j') *sont fortement opposées (sont très éloignées)*
- *Si* $r(j,j')=+1\Longrightarrow d^2(j,j')=0\Longrightarrow$ *les deux variables* (j,j') *sont fortement corrélées.*
- *Si* $r(j,j')=0\Longrightarrow d^2(j,j')=2\Longrightarrow$ *les deux variables* (j,j') *sont orthogonales par rapport à l'origine* $(j\perp j')$
- $\forall\,(j,j')$, *alors :* $0\leq d^2(j,j')\leq 4$
- $r(j,j')=Cos(j,j')$

En outre, la distance entre une variable à l'origine s'écrit comme suit :

$$d^2(j,O)=\sum_{i=1}^{n}(R_{ij}-O)^2=\sum_{i=1}^{n}R_{ij}^2=\sum_{i=1}^{n}\frac{(x_{ij}-\bar{X}_j)^2}{ns_j^2}=\sum_{i=1}^{n}\frac{s_j^2}{s_j^2}=1$$

Donc ; le nuage des points variables se présente dans un cercle de corrélation de rayon égale 1 et centré à l'origine (voir la figure 10).

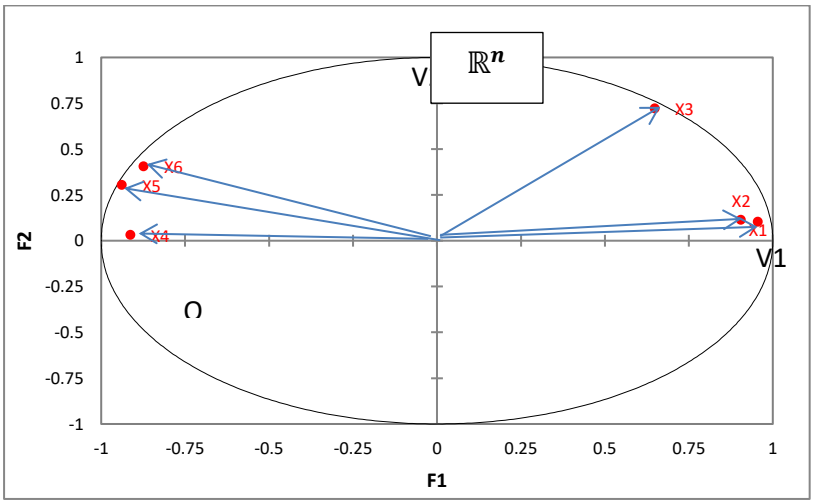


Figure 10. Exemple du cercle de corrélation

Composantes principales

Comme la matrice C a été diagonalisée dans l'analyse $\mathbb{R}^n$, les coordonnées des composantes principales se calculent comme suit :

$$G_\alpha = \sqrt{\lambda_\alpha} U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} R' F_\alpha = \begin{pmatrix} g_{\alpha 1} \\ . \\ . \\ g_{\alpha j} \\ . \\ . \\ g_{\alpha p}. \end{pmatrix} \text{ avec } \alpha = 1 \ldots . p$$

Tels que :

- U_α le vecteur propre α de la matrice C associé à la valeur propre λ_α
- $F_\alpha$le facteur principal α (coordonnées de la protection des points-individus)
- $g_{\alpha j}$=$\sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{X}_j}{s_j \sqrt{n}}\right) \left(\frac{F_{\alpha i}}{\sqrt{\lambda_\alpha}}\right)$ la coordonnée d'un point-variable j sur l'axe α

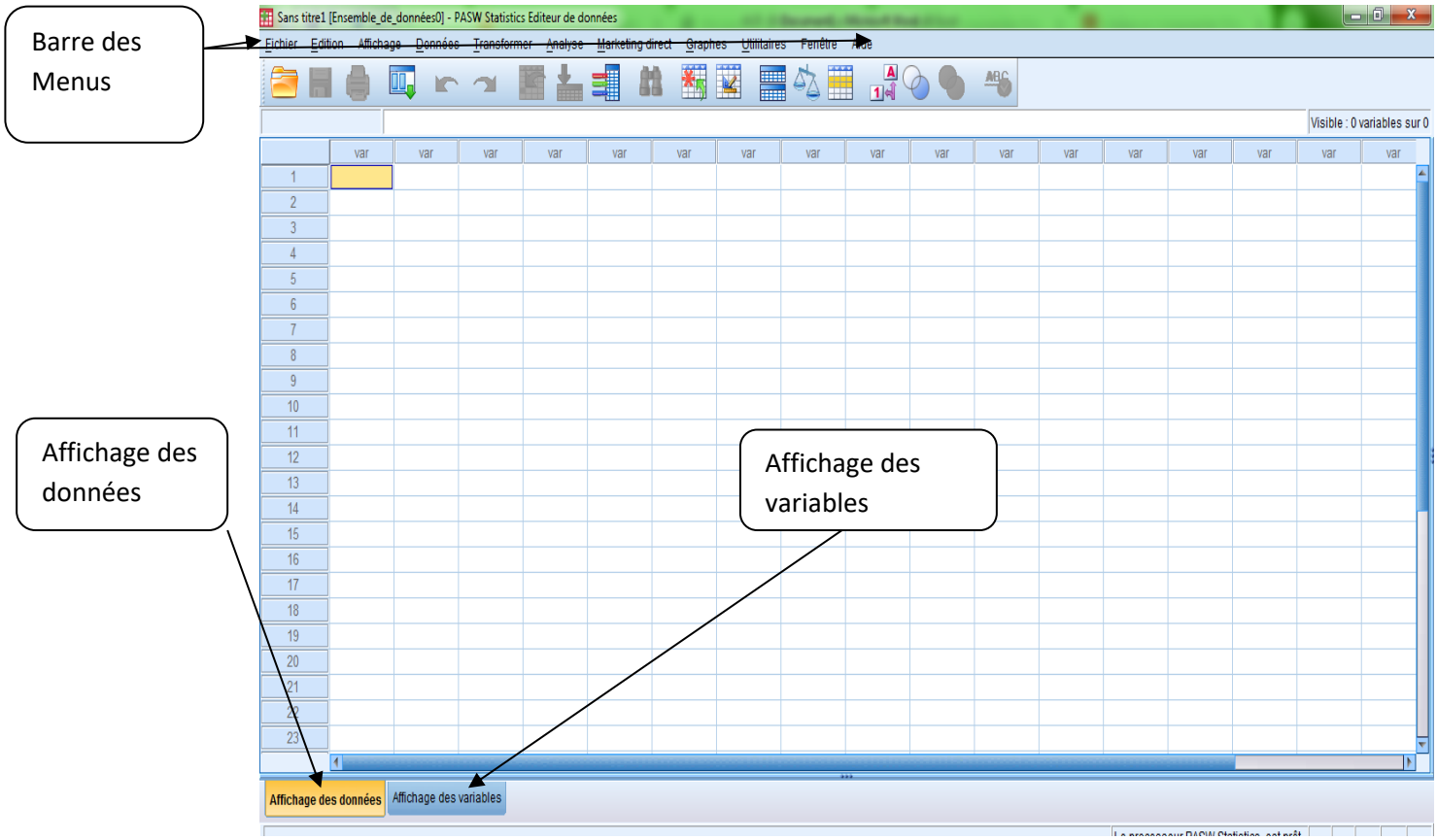
3.3. Application de la méthode « Analyse en Composantes Principales –ACP » sur le logiciel SPSS

Première étape : Créer une base de données SPSS (tableau rectangulaire Individus x Variables)

1. Organisation des données sur Excel

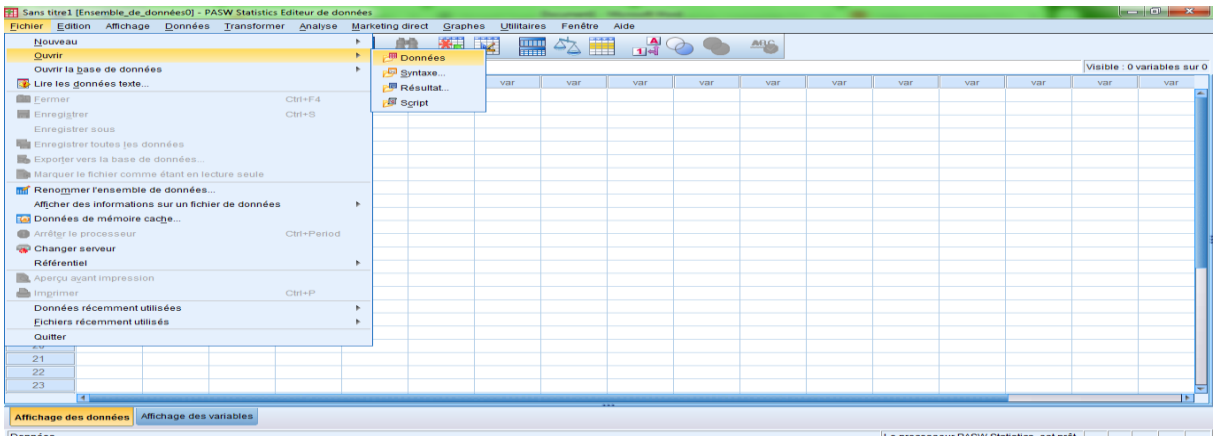
2. Importation des données Excel vers SPSS

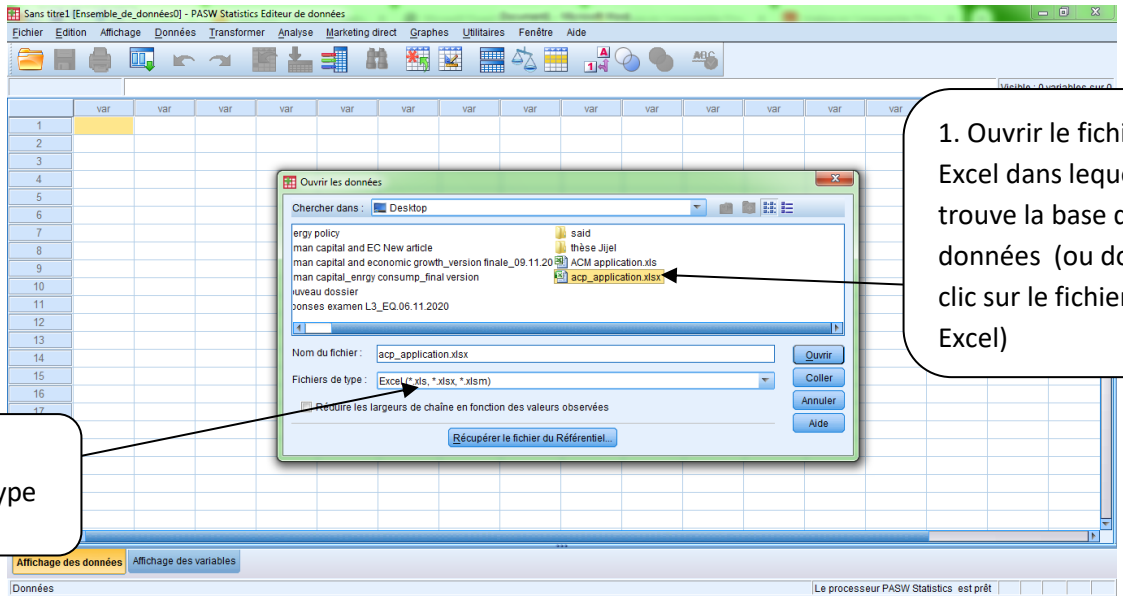
- ✓ Ouvrir le logiciel SPSS



✓ A partir des menus, sélectionnez :

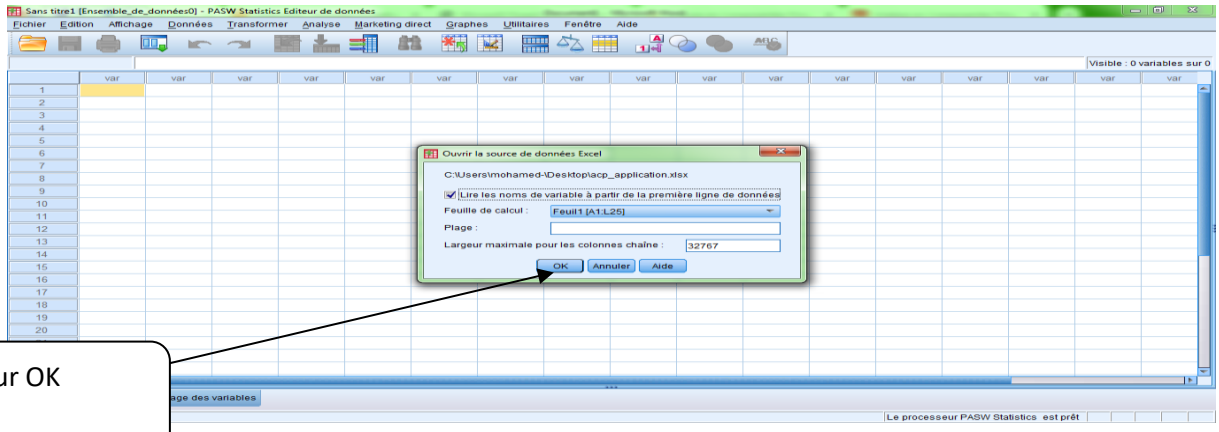
Fichier _Ouvrir_ données



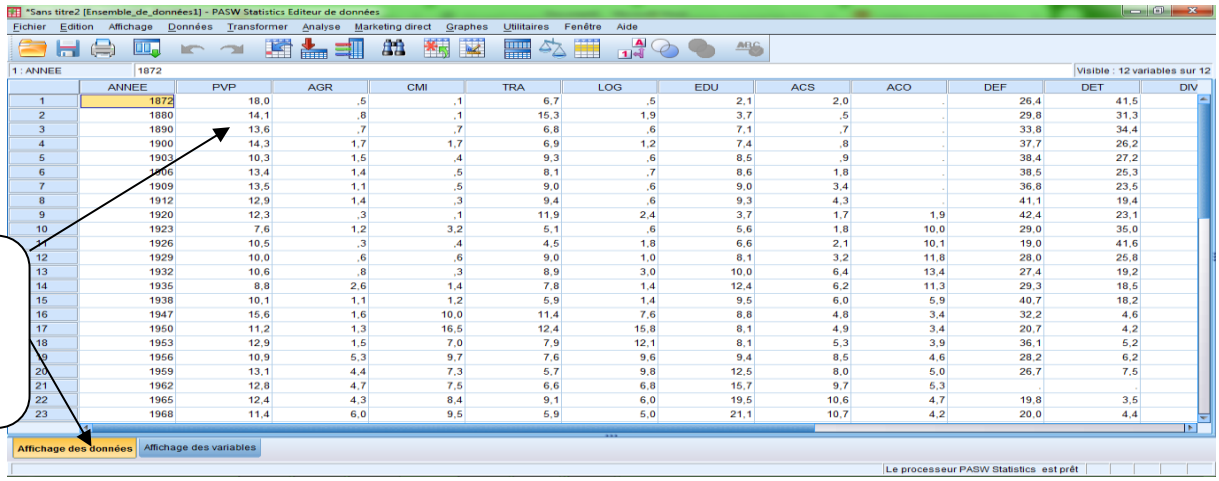


1. Ouvrir le fichier Excel dans lequel se trouve la base de données (ou double clic sur le fichier Excel)

2. Sélectionner Excel comme type du fichier



Cliquer sur OK



	ANNEE	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ACO	DEF	DET	DIV
1	1872	18,0	,5	,1	6,7	,5	2,1	2,0	.	26,4	41,5	
2	1880	14,1	,8	,1	15,3	1,9	3,7	,5	.	29,8	31,3	
3	1890	13,6	,7	,7	6,8	,6	7,1	,7	.	33,8	34,4	
4	1900	14,3	1,7	1,7	6,9	1,2	7,4	,8	.	37,7	26,2	
5	1903	10,3	1,5	,4	9,3	,6	8,5	,9	.	38,4	27,2	
6	1906	13,4	1,4	,5	8,1	,7	8,6	1,8	.	38,5	25,3	
7	1909	13,5	1,1	,5	9,0	,6	9,0	3,4	.	36,8	23,5	
8	1912	12,9	1,4	,3	9,4	,6	9,3	4,3	.	41,1	19,4	
9	1920	12,3	,3	,1	11,9	2,4	3,7	1,7	1,9	42,4	23,1	
10	1923	7,6	1,2	3,2	5,1	,6	5,6	1,8	10,0	29,0	35,0	
11	1926	10,5	,3	,4	4,5	1,8	6,6	2,1	10,1	19,0	41,6	
12	1929	10,0	,6	,6	9,0	1,0	8,1	3,2	11,8	28,0	25,8	
13	1932	10,6	,8	,3	8,9	3,0	10,0	6,4	13,4	27,4	19,2	
14	1935	8,8	2,6	1,4	7,8	1,4	12,4	6,2	11,3	29,3	18,5	
15	1938	10,1	1,1	1,2	5,9	1,4	9,5	6,0	5,9	40,7	18,2	
16	1947	15,6	1,6	10,0	11,4	7,6	8,8	4,8	3,4	32,2	4,6	
17	1950	11,2	1,3	16,5	12,4	15,8	8,1	4,9	3,4	20,7	4,2	
18	1953	12,9	1,5	7,0	7,9	12,1	8,1	5,3	3,9	36,1	5,2	
19	1956	10,9	5,3	9,7	7,6	9,6	9,4	8,5	4,6	28,2	6,2	
20	1959	13,1	4,4	7,3	5,7	9,8	12,5	8,0	5,0	26,7	7,5	
21	1962	12,8	4,7	7,5	6,6	6,8	15,7	9,7	5,3	.	.	
22	1965	12,4	4,3	8,4	9,1	6,0	19,5	10,6	4,7	19,8	3,5	
23	1968	11,4	6,0	9,5	5,9	5,0	21,1	10,7	4,2	20,0	4,4	

Base de données Excel a été importée avec succès vers SPSS

Deuxième étape : Application de l'ACP sur SPSS : Cas des données relatives à la structure fonctionnelle des dépenses de l'Etat français entre 1972-1971 (les données sont exprimées en %).

Définition des variables :

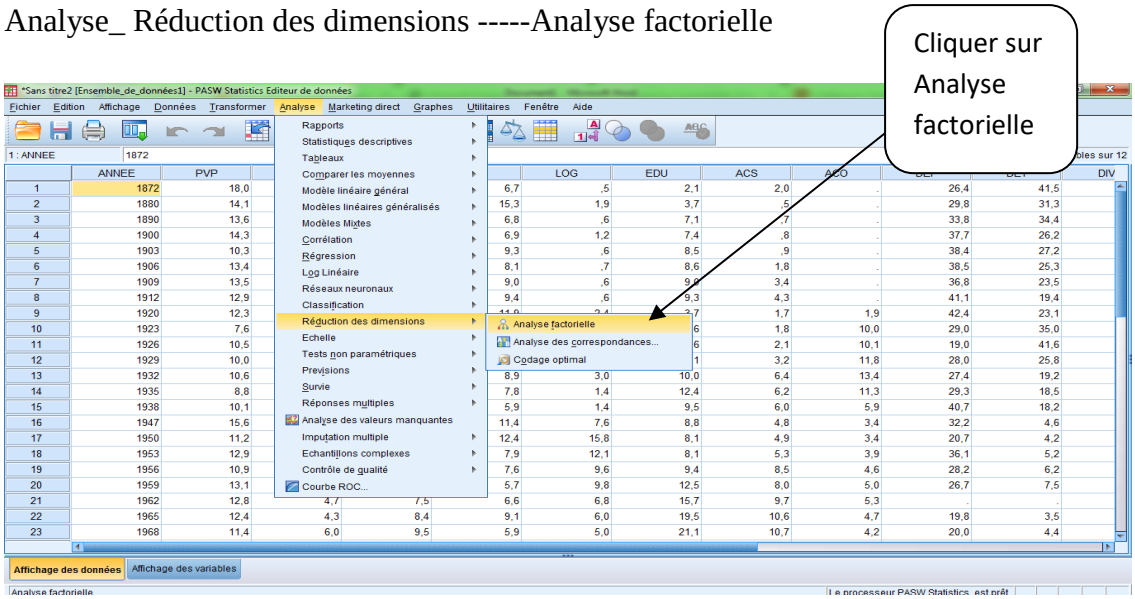
PVP : Pouvoirs publics**. AGR :** Agriculture**. CMI :** Commerce et Industrie**. TRA :** Transport.

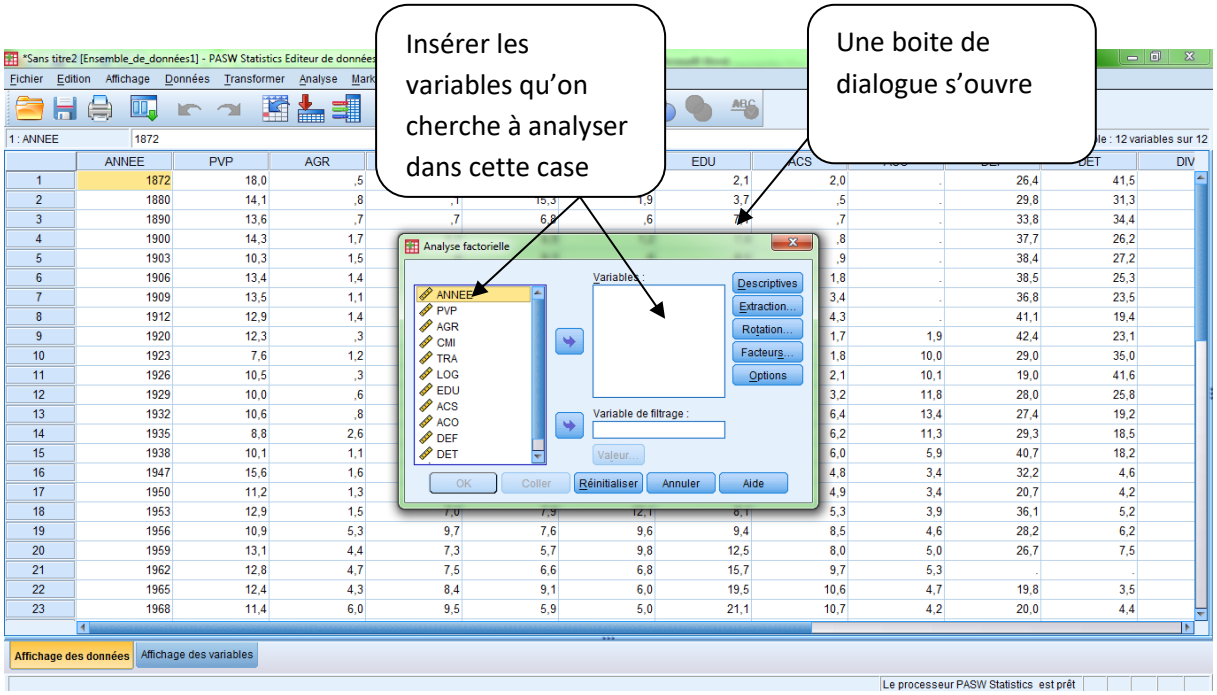
LOG : Logement et Aménagement du Territoire**. EDU :** Education et Culture**. ACS :** Action Sociale**. ACO :** Anciens Combattants**. DEF :** Défense**. DET :** Dette**. DIV :** Divers.

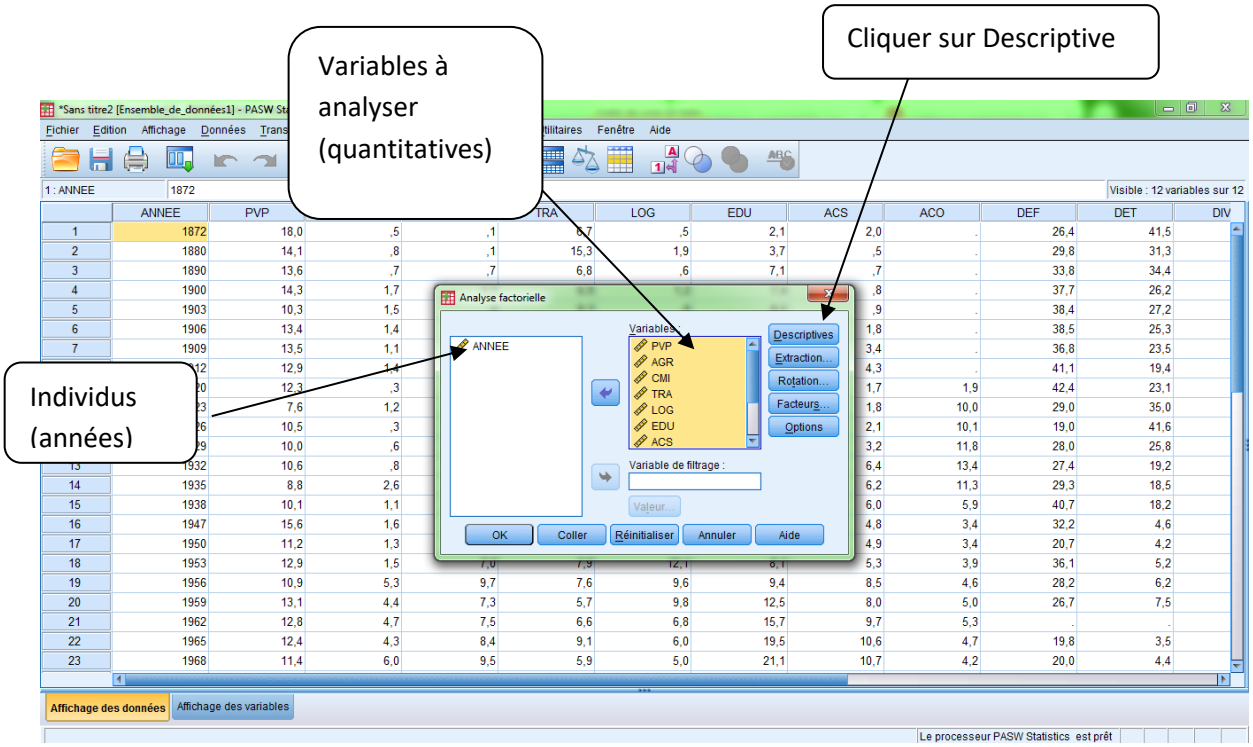
Il est à noter que les logiciels d'analyse de données prennent en considération par défaut les tableaux de données centrées et réduites.

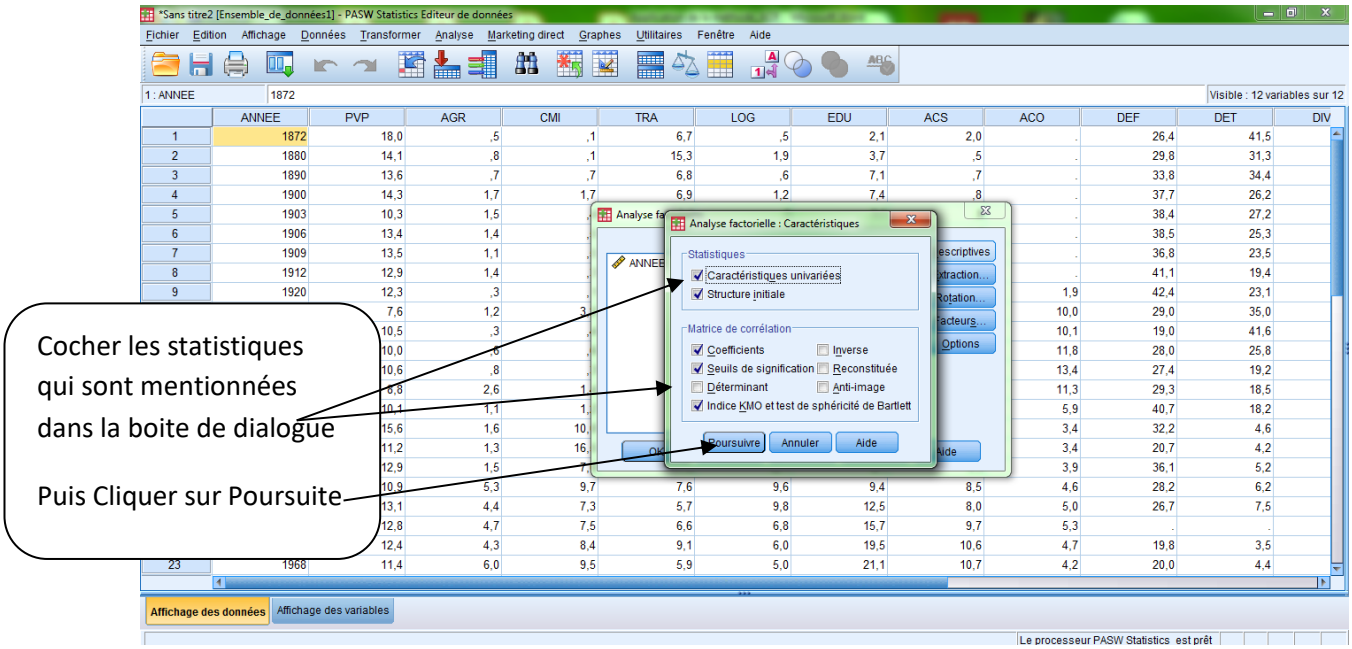
A partir de la barre des menus, sélectionnez :

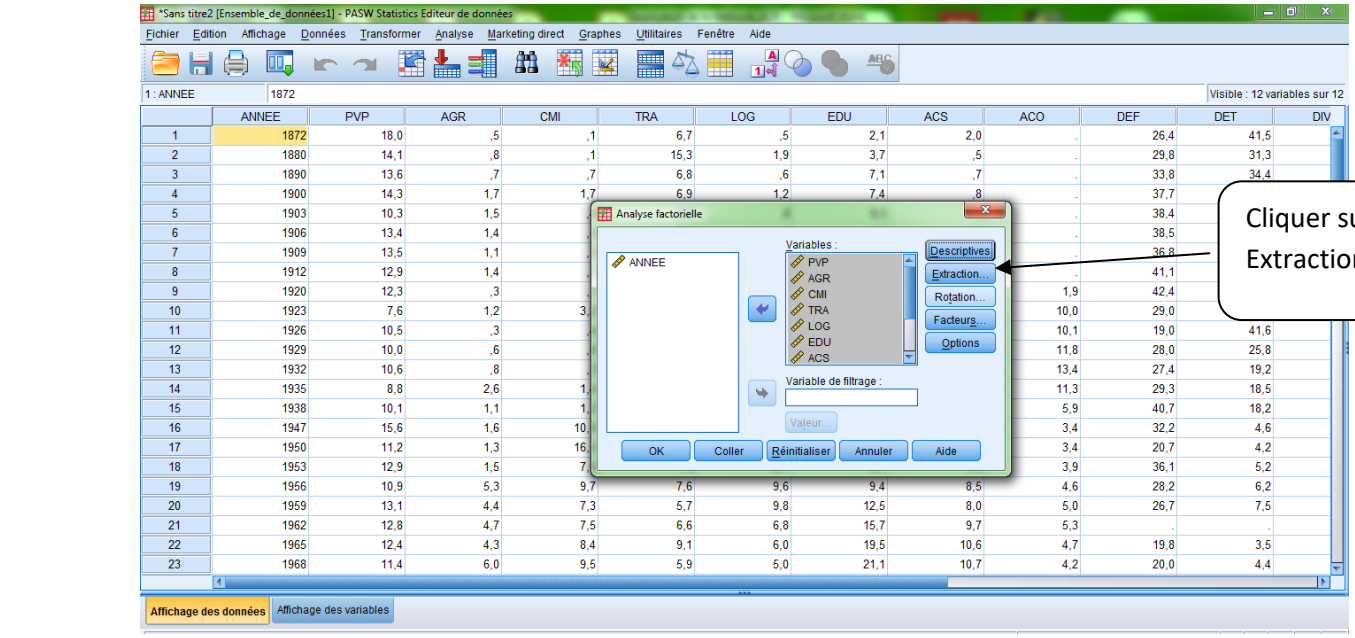
Analyse_ Réduction des dimensions -----Analyse factorielle

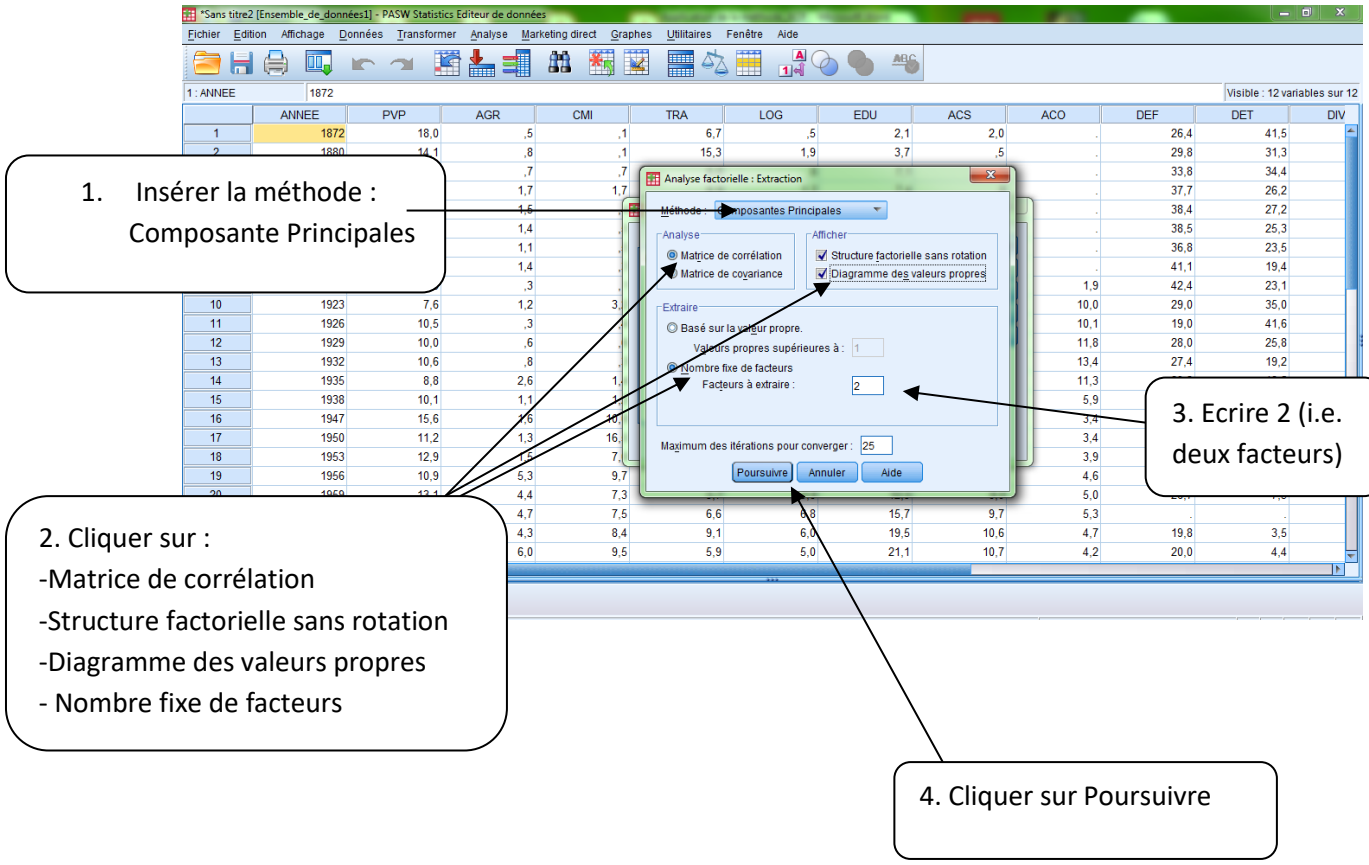


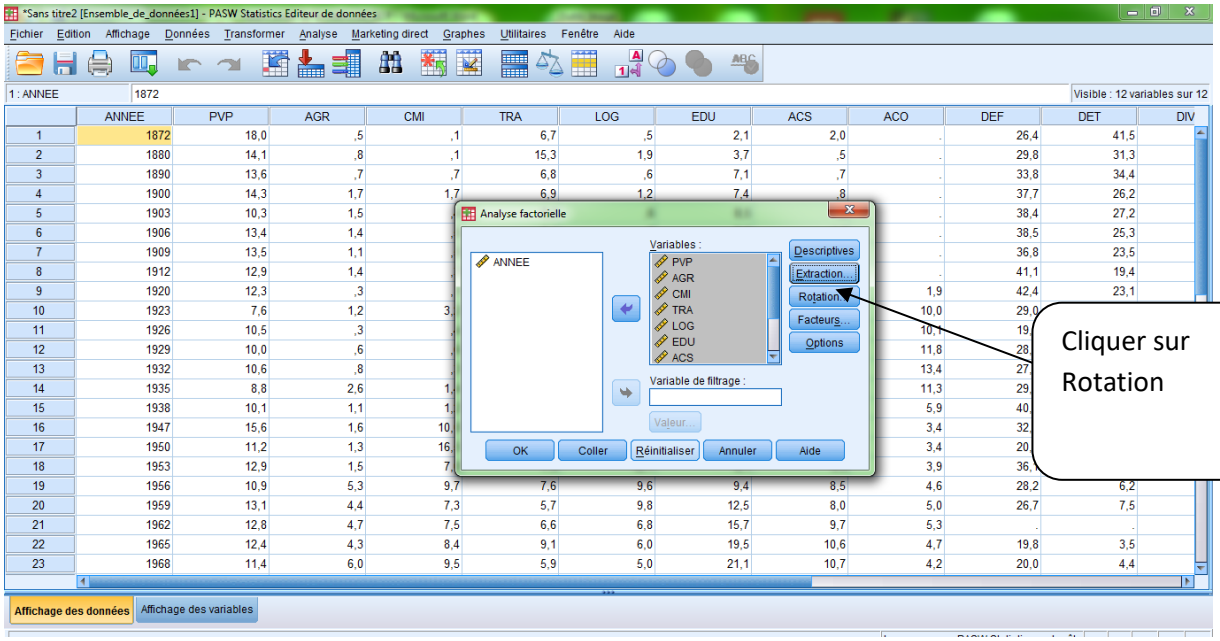

Insérer les variables qu'on cherche à analyser dans cette case
Une boite de dialogue s'ouvre
Analyse factorielle
Variables :
Variable de filtrage :
Valeur...
Descriptives
Extraction...
Rotation...
Facteurs...
Options
OK
Coller
Réinitialiser
Annuler
Aide
Affichage des données
Affichage des variables
Le processeur PASW Statistics est prêt

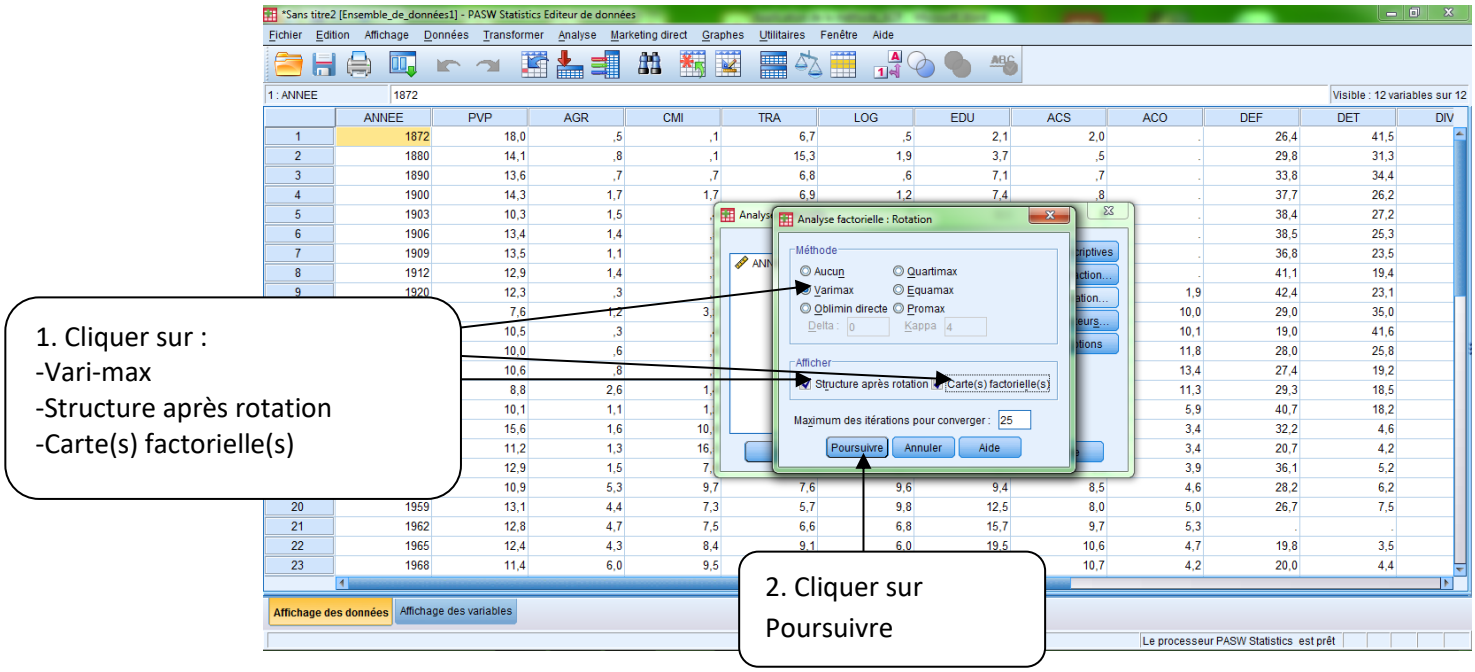

Variables à analyser (quantitatives)
Cliquer sur Descriptive
Individus (années)
Analyse factorielle
Variables :
Variable de filtrage :
Valeur...
Descriptives
Extraction...
Rotation...
Facteurs...
Options
OK
Coller
Réinitialiser
Annuler
Aide
Affichage des données
Affichage des variables
Le processeur PASW Statistics est prêt

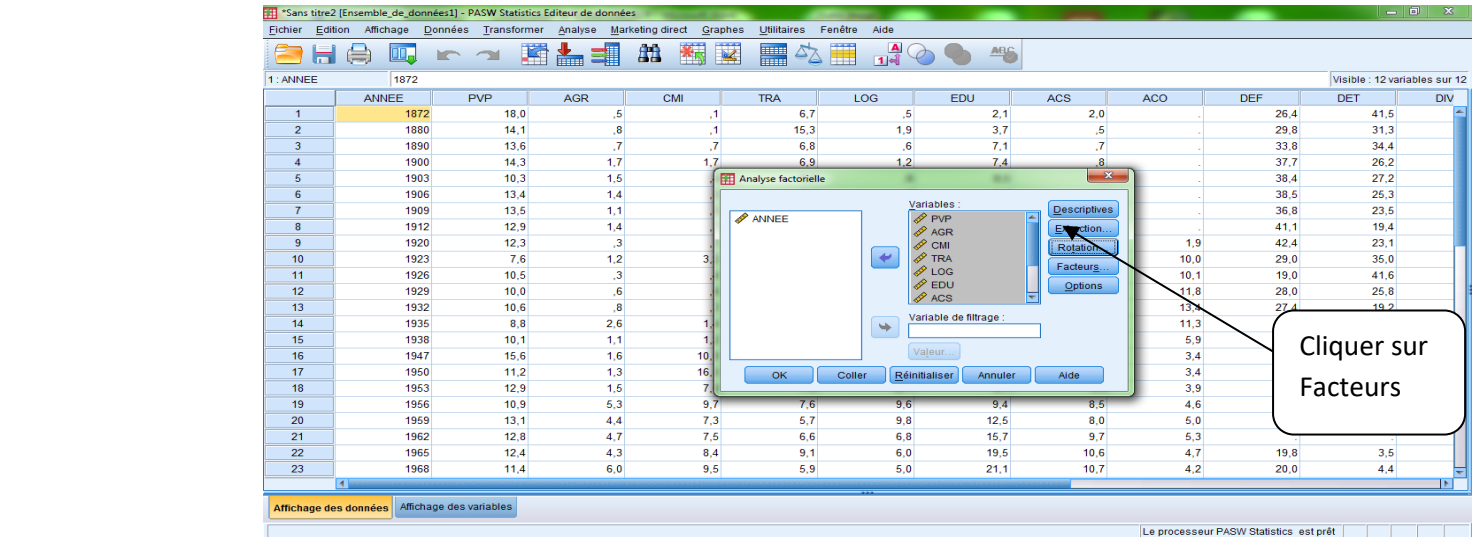

Cocher les statistiques qui sont mentionnées dans la boite de dialogue
Puis Cliquer sur Poursuite
Analyse factorielle : Caractéristiques
Statistiques
Caractéristiques univariées
Structure initiale
Matrice de corrélation
Coefficients
Inverse
Seuils de signification
Reconstituée
Déterminant
Anti-image
Indice KMO et test de sphéricité de Bartlett
Poursuivre
Annuler
Aide

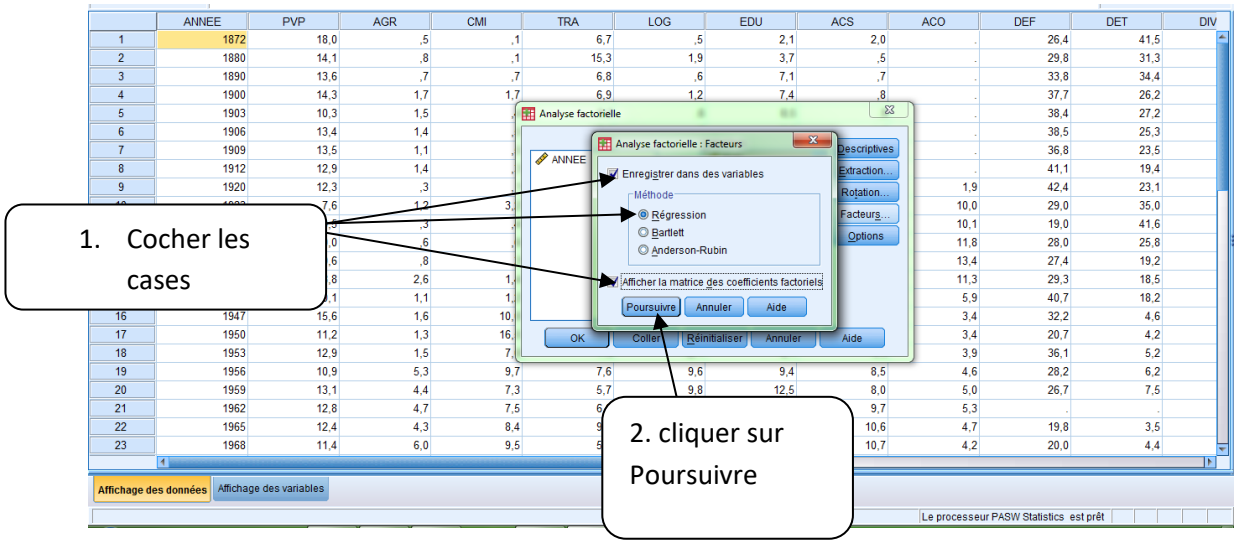

Cliquer sur Extraction
Analyse factorielle
ANNEE
Variables :
PVP
AGR
CMI
TRA
LOG
EDU
ACS
Descriptives...
Extraction...
Rotation...
Facteurs...
Options
Variable de filtrage :
Valeur...
OK
Coller
Réinitialiser
Annuler
Aide

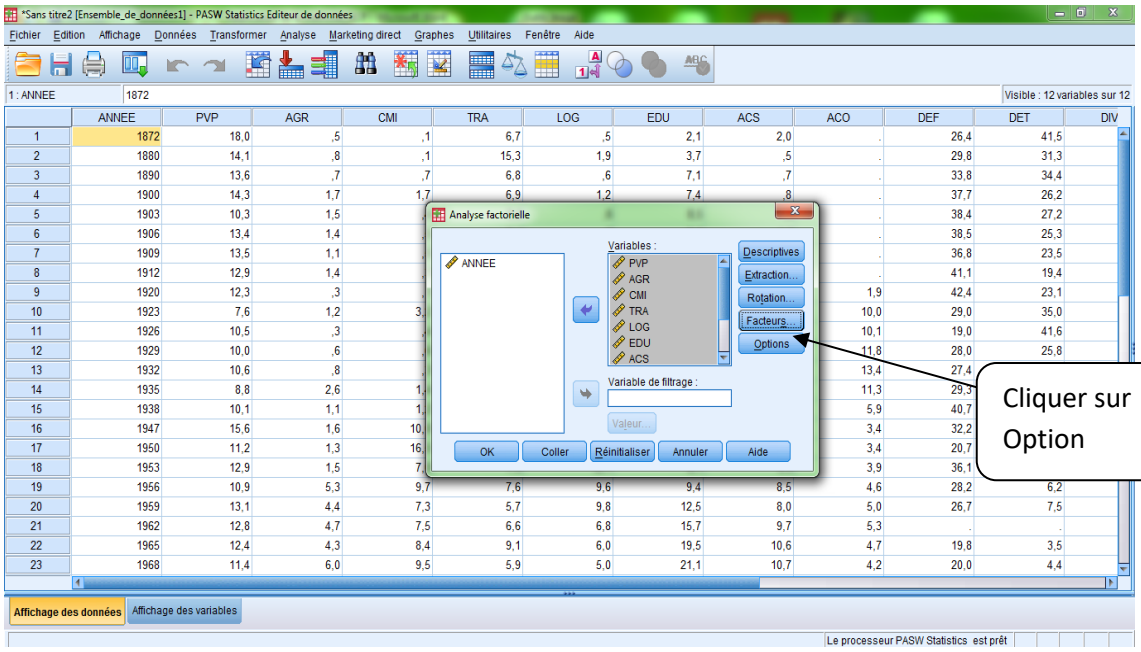

1. Insérer la méthode : Composante Principales
2. Cliquer sur :
-Matrice de corrélation
-Structure factorielle sans rotation
-Diagramme des valeurs propres
- Nombre fixe de facteurs
3. Ecrire 2 (i.e. deux facteurs)
4. Cliquer sur Poursuivre
Analyse factorielle : Extraction
Méthode : Composantes Principales
Analyse
Matrice de corrélation
Matrice de covariance
Afficher
Structure factorielle sans rotation
Diagramme des valeurs propres
Extraire
Basé sur la valeur propre.
Valeurs propres supérieures à : 1
Nombre fixe de facteurs
Facteurs à extraire : 2
Maximum des itérations pour converger : 25
Poursuivre
Annuler
Aide

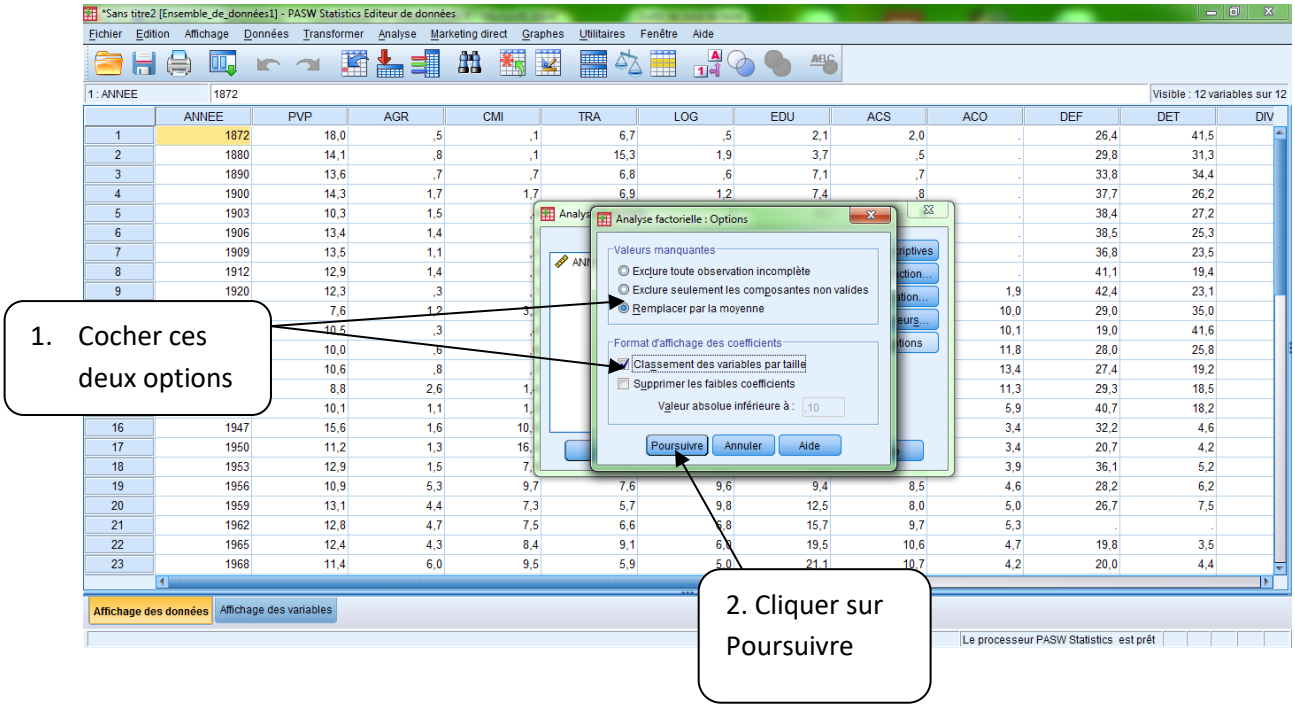

Cliquer sur Rotation
Analyse factorielle
ANNEE
Variables :
PVP
AGR
CMI
TRA
LOG
EDU
ACS
Descriptives
Extraction...
Rotation...
Facteurs...
Options
Variable de filtrage :
Valeur...
OK
Coller
Réinitialiser
Annuler
Aide

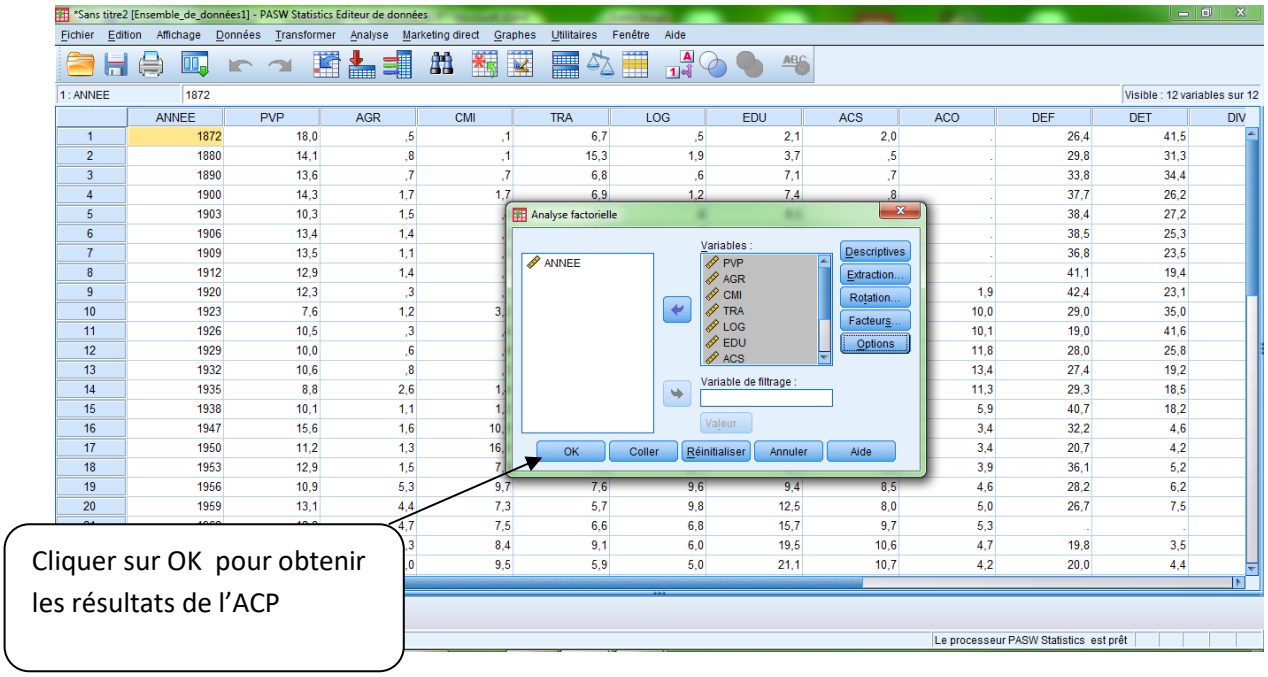

1. Cliquer sur :
-Vari-max
-Structure après rotation
-Carte(s) factorielle(s)
2. Cliquer sur Poursuivre
Analyse factorielle : Rotation
Méthode
Aucun
Quartimax
Varimax
Equamax
Oblimin directe
Promax
Delta : 0
Kappa : 4
Afficher
Structure après rotation
Carte(s) factorielle(s)
Maximum des itérations pour converger : 25
Poursuivre
Annuler
Aide


Cliquer sur Facteurs
Analyse factorielle
ANNEE
Variables :
PVP
AGR
CMI
TRA
LOG
EDU
ACS
Descriptives
Extraction...
Rotation...
Facteurs...
Options
Variable de filtrage :
Valeur...
OK
Coller
Réinitialiser
Annuler
Aide


1. Cocher les cases
2. cliquer sur Poursuivre
Analyse factorielle : Facteurs
Enregistrer dans des variables
Méthode
Régression
Bartlett
Anderson-Rubin
Afficher la matrice des coefficients factoriels
Poursuivre
Annuler
Aide


Analyse factorielle
Variables :
ANNEE
PVP
AGR
CMI
TRA
LOG
EDU
ACS
Descriptives
Extraction...
Rotation...
Facteurs...
Options
Variable de filtrage :
Valeur...
OK
Coller
Réinitialiser
Annuler
Aide
Cliquer sur Option


Analyse factorielle : Options
Valeurs manquantes
Exclure toute observation incomplète
Exclure seulement les composantes non valides
Remplacer par la moyenne
Format d'affichage des coefficients
Classement des variables par taille
Supprimer les faibles coefficients
Valeur absolue inférieure à : ,10
Poursuivre
Annuler
Aide
1. Cocher ces deux options
2. Cliquer sur Poursuivre



Cliquer sur OK pour obtenir les résultats de l'ACP

Principaux Résultats de l'ACP normée

Quant à l'applicabilité de la méthode ACP, le résultat du test de spécificité de Bartlett est consigné dans le tableau ci-dessous, ceci permet d'accepter l'hypothèse H_1 qui stipule au moins l'une des corrélations entre les variables est significativement différente de zéro. Donc, les variables incluses dans l'analyse sont très bien corrélées, et par conséquent les données sont factorisables.

Indice KMO et test de Bartlett

Indice de Kaiser-Meyer-Olkin pour la mesure de la qualité d'échantillonnage.		,233
Test de sphéricité de Bartlett	Khi-deux approx.	274,153
	ddl	55
	Signification	,000

Variance totale expliquée

Composante	Valeurs propres initiales			Sommes extraites du carré des chargements		
	Total	% de la variance	% cumulé	Total	% de la variance	% cumulé
1	4,946	44,961	44,961	4,946	44,961	44,961
2	2,055	18,679	63,640	2,055	18,679	63,640
3	1,289	11,716	75,356			
4	1,017	9,241	84,597			
5	,710	6,450	91,047			
6	,557	5,067	96,114			
7	,204	1,852	97,966			
8	,125	1,139	99,105			
9	,062	,566	99,671			
10	,035	,318	99,989			
11	,001	,011	100,000			

Méthode d'extraction : Analyse en composantes principales.

A partir du tableau variance totale expliquée :

- ✓ La première composante principale explique 44,96% de l'inertie totale
- ✓ La deuxième composante principale explique 18,67% de l'inertie totale
- ✓ La deuxième composante principale explique 11,71% de l'inertie totale
- ✓ Le sous espace à deux dimensions, le premier plan factoriel, explique 63,64% de l'inertie totale

Premièrement, le choix du nombre d'axes factoriels qui permet d'expliquer le maximum d'information contenue dans le tableau de données est d'importance capital pour l'analyse soit plus pertinente. Comme il a été mentionné, l'utilisation du test du coude, *scree test,* ce qui signifie que le nombre d'axes à retenir serait égal au nombre de valeurs propres qui se trouvent avant le premier point d'inflexion du graphique des valeurs propres. Pour notre exemple, le nombre d'axe est égal à 3. Cependant, avec 2 axes, le premier plan factoriel explique le maximum d'inertie totale, par conséquent, l'analyse se base sur le sous espace avec les deux premières dimensions.

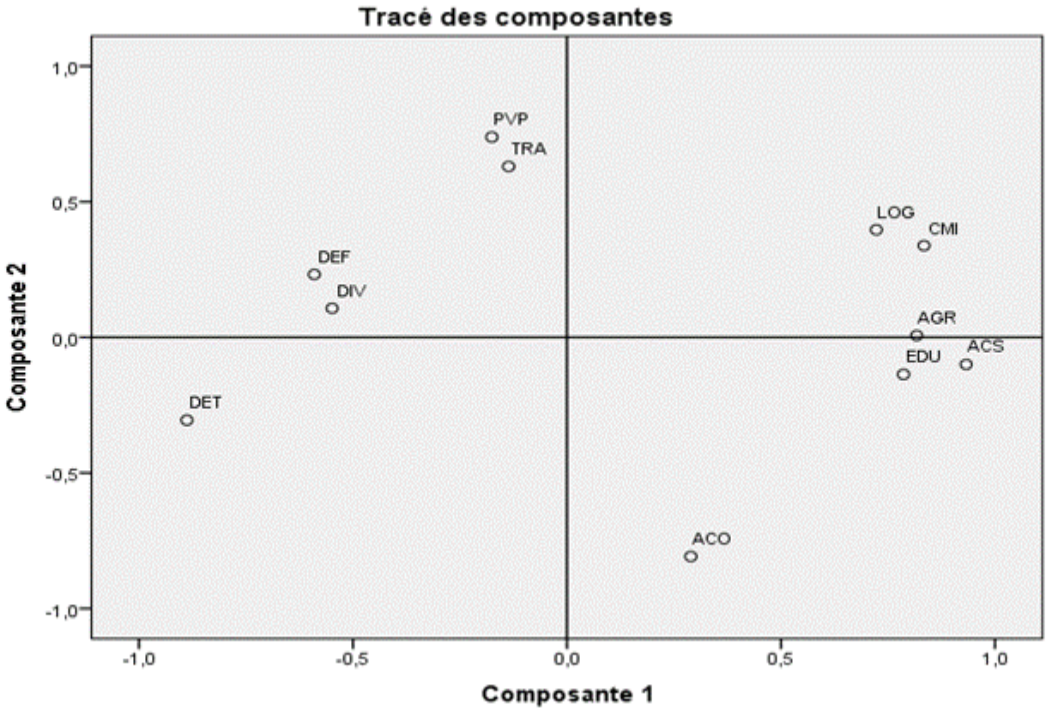
Analyse dans $\mathbb{R}^n$

Dans ce qui, nous nous focalisons sur les principales conclusions.

La représentation graphique du nuage des points-variables dans le premier plan factoriel, fait ressortir :

- ✓ Toutes les variables sont très bien représentées (car elle se rapproche du cercle de corrélation)
- ✓ Une corrélation positive entre : (LOG, CMI, AGR, EDU et ACS), (PVP et TRA), (DEF, DET et DIV)

- ✓ Une opposition entre (LOG, CMI, AGR, EDU, ACS) et (DEF, DIV, DET)
- ✓ Une opposition entre (PVP et TRA) et ACO



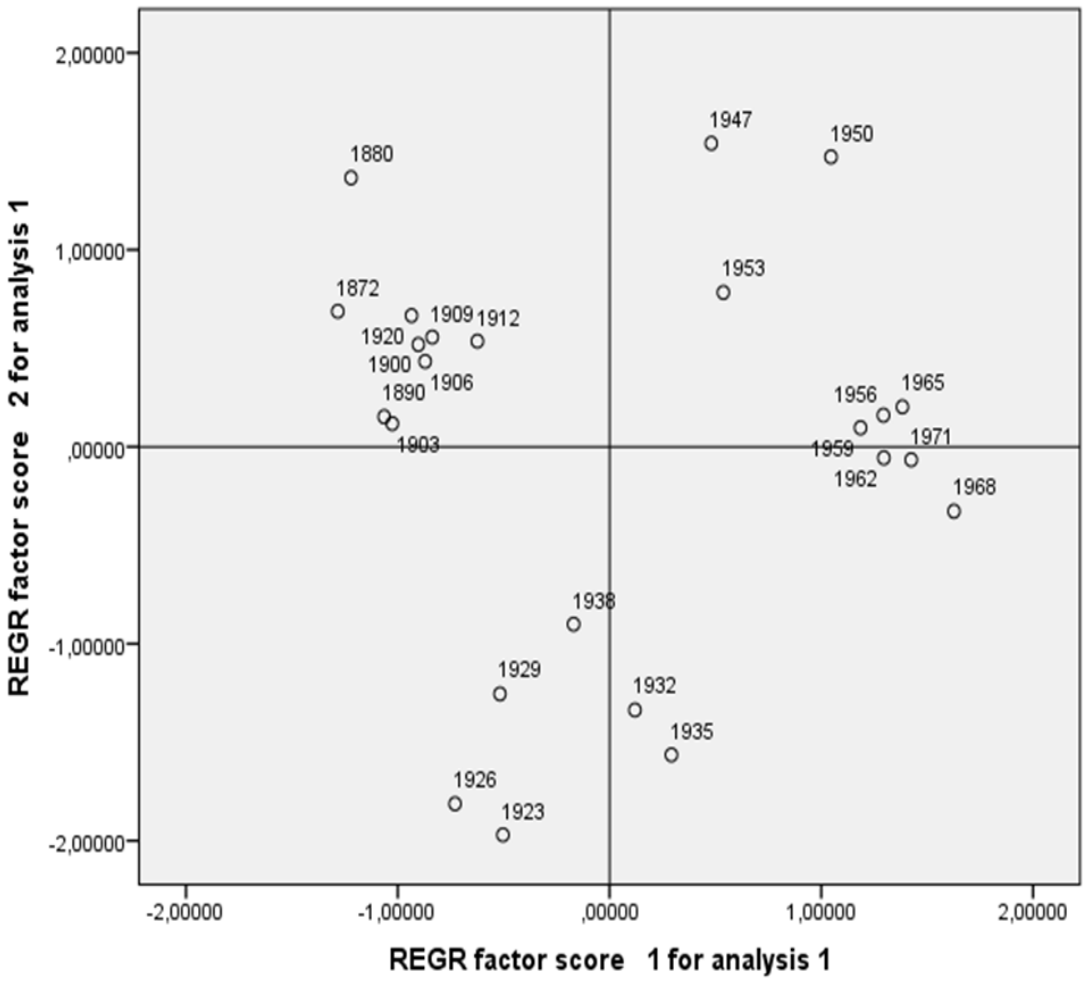
Analyse dans $\mathbb{R}^p$

La représentation graphique des individus, années, en fonction des variables donne ce qu'on appelle le nuage des individus. Le premier plan factoriel monte que les années sont échelonnées selon l'ordre chronologique, et il fait ressortir plusieurs clases d'années. Quatre classes d'années peuvent être distinguées :

- ✓ Classe 1 : 1872, 1880, 1890, 1900, 1903, 1906, 1909, 1912, 1920 (années caractérisées par une forte dense de la défense et les dettes publiques)
- ✓ Classe 2 : 1923 1926 1929 1932 1935 1938 (années juste après la première guerre mondiale, elles sont marquées par les dépenses liées aux anciens combattants)

- ✓ Classe 3 : 1947 1950 1953 (années d'après deuxième guerre mondiale, elles sont marquées par les dépenses destinées aux logements pour financer les programmes publics de reconstruction et au développement du commerce et industrie)
- ✓ Classe 4 : 1956 1959 1962 1965 1968 1971 (années de la relance économique, les tentes glorieuses caractérisées par de forte croissance économique où les dépenses publiques étaient réservées au développement humain)

En faisant le lien avec le nuage des variables,


2,00000
1,00000
,00000
-1,00000
-2,00000
REGR factor score 2 for analysis 1
-2,00000
-1,00000
,00000
1,00000
2,00000
REGR factor score 1 for analysis 1
1880
1872
1909
1912
1920
1900
1906
1890
1903
1947
1950
1953
1956
1965
1959
1971
1962
1968
1938
1929
1932
1935
1926
1923

EXERCICES

Exercice 1

On considère le tableau de données suivant :

Individus /Variables	x_1	x_2
1	-0,5	0
2	0	1
3	-1	2
4	1	-2
5	0,5	-1

1) Calculer le centre de gravité du nuage de points, que remarquez-vous ?
2) Représenter graphiquement les 5 individus.
3) Calculer la matrice Variance-covariance $V = \frac{1}{n} X'X$.
4) Calculer les valeurs propres et les vecteurs propres de la matrice V, et déduire la valeur de l'inertie totale.
5) calculer les coordonnées de la première composante principale (le premier axe factoriel).

Solution

1. Calcul du centre de gravité

La matrice des données $X = \begin{pmatrix} -0.5 & 0 \\ 0 & 1 \\ -1 & 2 \\ 1 & -2 \\ 0.5 & 1 \end{pmatrix}$

Donc, le centre de gravité $\bar{X}_g = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

Remarque : les variables sont centrées par rapport à leurs moyennes

2. Représentation graphique

3. Calcul de la matrice variance/covariance (V)

$$\mathbf{V} = \frac{1}{n} X'X = \frac{1}{5} \begin{pmatrix} -0,5 & 0 & -1 & 1 & 0,5 \\ 0 & 1 & 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} 0,5 & 0 \\ 0 & 1 \\ -1 & 2 \\ 1 & -2 \\ 0,5 & 1 \end{pmatrix} = \begin{pmatrix} 0,5 & -0,9 \\ -0,9 & 2 \end{pmatrix}$$

4.

a. Calcul des valeurs propres de V

$$\left| V_{(2,2)} - \lambda I_2 \right| = 0 \Longrightarrow \begin{vmatrix} 0{,}5 - \lambda & -0{,}9 \\ -0{,}9 & 2 - \lambda \end{vmatrix} = 0$$

$$\Longrightarrow (0{,}5 - \lambda)(\, 2 - \lambda) - (\, -0{,}9\,)(\, -0{,}9) = 0$$

$$\Longrightarrow (0{,}5 - \lambda)(\, 2 - \lambda) - (\, 0{,}81\,) = 0$$

$$\Longrightarrow \lambda^2 - 2{,}5\lambda + 0{,}19 = 0 \ldots\ldots.(1)$$

Solution de l'équation (1)

$\Delta = b^2 - 4ac = (-2{,}5)^2 - 4(1)(0{,}19) = 5{,}46 > 0$

D'où , il existe eux racine

$$\lambda_1 = \frac{-b + \sqrt{\Delta}}{2a} = \frac{2{,}5 + 2{,}33}{2} = 2{,}41$$

$$\lambda_2 = \frac{-b - \sqrt{\Delta}}{2a} = \frac{2{,}5 - 2{,}33}{2} = 0{,}08$$

Dès lors, la valeur de l'inertie totale $I_t = \lambda_1 + \lambda_2 = 2{,}41 + 0{,}08 = 2{,}49$

b. Les vecteurs propres de V correspondants aux valeurs propres $\lambda_1 et\ \lambda_2$

- Cas $\lambda_1 = 2{,}41$

Soit $U_1 = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ vecteur propre de V correspondant à $\lambda_1 = 2{,}41$, alors : $(V_{(2,2)} - \lambda_1 I_2)\, U_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$\Longrightarrow \begin{pmatrix} -1{,}91 & -0{,}9 \\ -0{,}9 & -0{,}41 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Longrightarrow \begin{cases} -1{,}91u_1 - 0{,}9u_2 = 0 \\ -{,}9u_1 - 0{,}41u_2 = 0 \end{cases}$$

$$\Longrightarrow \begin{cases} u_1 = -0{,}47u_2 = 0 \\ u_2 \in \mathbb{R} \end{cases}$$

$$\Longrightarrow U_1 = \begin{pmatrix} -0{,}47u_2 \\ u_2 \end{pmatrix}$$

$$\Longrightarrow U_1 = u_2 \begin{pmatrix} -0{,}47 \\ 1 \end{pmatrix}$$

Pour $u_2 = 1$; $U_1 = \begin{pmatrix} -0,47 \\ 1 \end{pmatrix}$ est un vecteur propre de la matrice V correspondant à la valeur propre $\lambda_1 = 2,41$

U_1 est-il unitaire ?

$\| \overrightarrow{U_1} \| = \sqrt{{u_1}^2 + {u_2}^2} = \sqrt{-0,47^2 + 1^2} = \sqrt{1,22} \neq 1$

Donc, pour rendre U_1 unitaire (normé) on divise ses coordonnées par $\| \overrightarrow{U_1} \|$, comme ainsi :

$U_{1,unitaire} = \begin{pmatrix} -0,47/\sqrt{1,22} \\ 1/\sqrt{1,22} \end{pmatrix} = \begin{pmatrix} -0,42 \\ 0,90 \end{pmatrix}$

- Cas $\lambda_2 = 0,08$

Soit $U_2 = \begin{pmatrix} u'_1 \\ u'_2 \end{pmatrix}$ vecteur propre de V correspondant à $\lambda_2 = 0,08$, alors : $(V_{(2,2)} - \lambda_2 I_2)\, U_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$\Longrightarrow \begin{pmatrix} 0,42 & -0,9 \\ -0,9 & 1,92 \end{pmatrix} \begin{pmatrix} u'_1 \\ u'_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Longrightarrow \begin{cases} 0,42u'_1 - 0,9u'_2 = 0 \ldots\ldots (1) \\ -0,9u'_1 + 1,92u'_2 = 0 \ldots\ldots (2) \end{cases}$$

De (1) $\Longrightarrow 0,42u'_1 = 0,9u'_2$

$\Longrightarrow u'_1 = 2,14u'_2$ **…….(3)**

En remplaçant (3) dans (2) on obtient : $-1,92u'_2 + 1,92u'_2 = 0, \forall\, u'_2 \in \mathbb{R}$

D'où ; $U_2 = \begin{pmatrix} u'_1 \\ u'_2 \end{pmatrix} = \begin{pmatrix} 2,14u'_2 \\ u'_2 \end{pmatrix} = u'_2 \begin{pmatrix} 2,14 \\ 1 \end{pmatrix}$

Pour $u'_2 = 1$, alors $U_2 = \begin{pmatrix} 2,14 \\ 1 \end{pmatrix}$ est un vecteur propre de V correspondant à $\lambda_2 = 0,08$

U_2 est-il unitaire ?

$\| \overrightarrow{U_2} \| = \sqrt{2,14^2 + 1^2} = \sqrt{5,579} = 2,361 \neq 1$

$U_{2,unitaire} = U_{2,normé} = \begin{pmatrix} 2,14/2,361 \\ 1/2,361 \end{pmatrix} = \begin{pmatrix} 0,90 \\ 0,42 \end{pmatrix}$

5. Calcul des coordonnées du premier axe factoriel (F1)

$$F1 = XU_{1,unitaire} = \begin{pmatrix} -0.5 & 0 \\ 0 & 1 \\ -1 & 2 \\ 1 & -2 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} -0{,}42 \\ 0{,}90 \end{pmatrix} = \begin{pmatrix} 0{,}21 \\ 0{,}90 \\ 2{,}22 \\ -2{,}22 \\ -1{,}11 \end{pmatrix}$$

Exercice 2

On étudie les consommations annuelles en 2017, exprimées en DA, de 6 produits alimentaires (les variables), les individus étant 8 catégories socioprofessionnelles (CSP). Les données sont des moyennes par CSP (Tableau n°1):

Tableau n°1 : Données initiales

Individus	PAO	PAA	POT	LEC	RAI	PLP
AGRI	167	1	41	8	6	6
SAAG	162	2	40	12	4	15
PRIN	119	6	39	5	13	41
CSUP	87	11	27	3	18	39
CMOY	103	5	32	4	11	30
EMPL	111	4	34	6	10	28
OUVR	130	3	43	7	7	16
INAC	138	7	53	8	12	20

NB :

Individus	Variables
AGRI : Exploitants agricoles SAAG: Salariés agricoles PRIN : Professions indépendantes CSUP : Cadres supérieurs CMOY: Cadres moyens EMPL: Employés OUVR : Ouvriers INAC : Inactifs	PAO : Pain ordinaire PAA : Autre pain POT: Pommes de terre LEC: Légumes secs RAI: Raisin de tables PLP: Plats préparés

On a réalisé une analyse en composantes principales (ACP) sur le tableau n°1, les résultats obtenus sont les présentés ci-dessous :

Matrice de corrélation (Pearson (n)) :

Variables	PAO	PAA	POT	LEC	RAI	PLP
PAO	**1**	**-0,774**	0,656	**0,889**	**-0,833**	**-0,856**
PAA	**-0,774**	**1**	-0,333	-0,673	**0,959**	**0,771**
POT	0,656	-0,333	**1**	0,603	-0,410	-0,554
LEC	**0,889**	-0,673	0,603	**1**	**-0,824**	**-0,751**
RAI	**-0,833**	**0,959**	-0,410	**-0,824**	**1**	**0,834**
PLP	**-0,856**	**0,771**	-0,554	**-0,751**	**0,834**	**1**

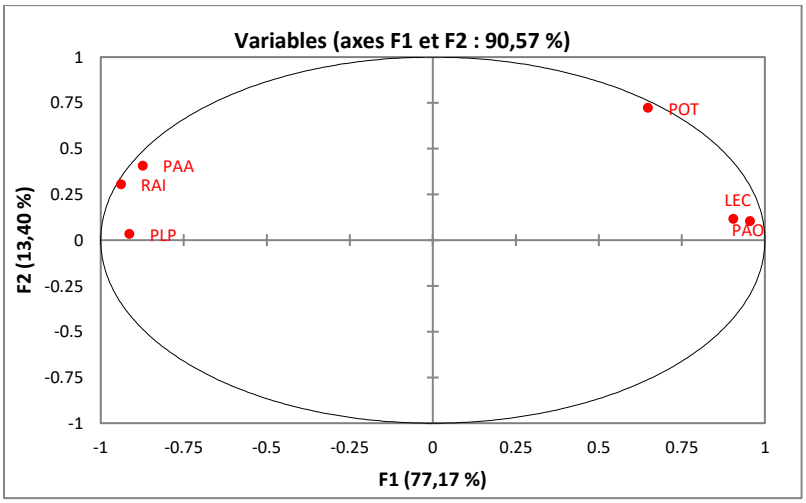
Les valeurs en gras sont différentes de 0 à un niveau de signification alpha=0,05

Valeurs propres :

	F1	F2	F3	F4	F5	F6
Valeur propre	4,630	0,804	0,266	0,202	0,092	0,005
Variabilité (%)	77,170	13,398	4,441	3,368	1,532	0,091
% cumulé	77,170	90,568	95,009	98,377	99,909	100,000

Corrélations entre les variables et les facteurs :

	F1	F2	F3	F4	F5	F6
PAO	0,956	0,103	0,089	-0,072	0,251	-0,017
PAA	-0,873	0,406	0,166	-0,205	-0,045	-0,040
POT	0,649	0,722	-0,176	0,159	-0,042	-0,001
LEC	0,906	0,115	0,396	-0,007	-0,095	0,025
RAI	-0,939	0,305	-0,005	-0,113	0,102	0,053
PLP	-0,913	0,032	0,207	0,342	0,077	-0,009

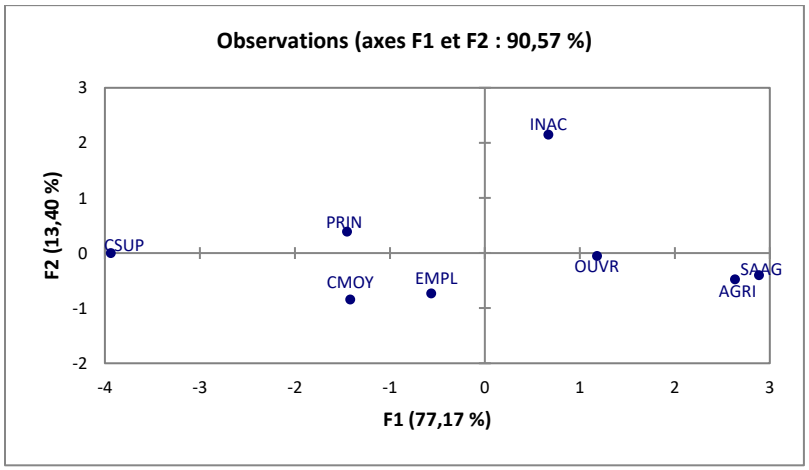


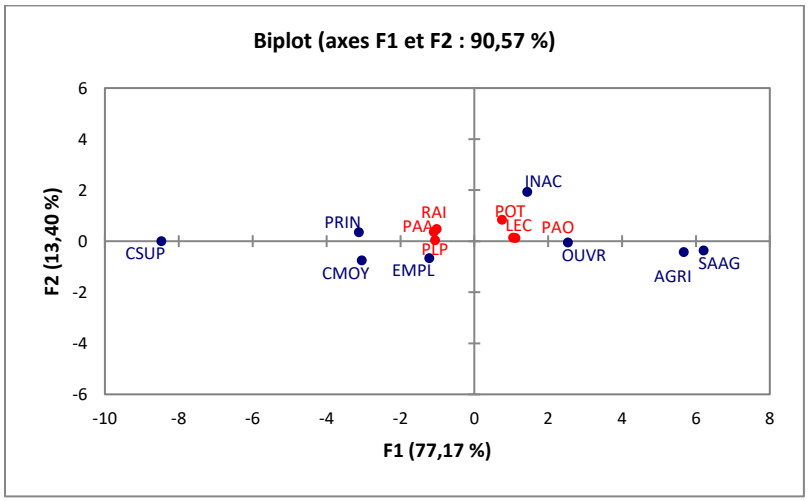
Contributions des variables (%) :

	F1	F2	F3	F4	F5	F6
PAO	19,724	1,310	2,942	2,561	68,451	5,012
PAA	16,458	20,498	10,378	20,736	2,171	29,758
POT	9,083	64,862	11,650	12,475	1,908	0,022
LEC	17,710	1,632	58,986	0,026	9,777	11,869
RAI	19,030	11,567	0,009	6,296	11,246	51,852
PLP	17,995	0,130	16,036	57,906	6,447	1,487

Contributions des observations (%)

	F1	F2	F3	F4	F5	F6
AGRI	18,745	3,626	11,774	16,613	35,257	1,473
SAAG	22,483	2,533	56,989	0,002	1,792	2,516
PRIN	5,689	2,304	1,105	43,311	30,281	1,953
CSUP	41,836	0,000	6,712	32,902	0,122	1,136
CMOY	5,417	11,210	7,737	1,354	2,700	6,583
EMPL	0,862	8,483	0,079	4,280	7,625	63,573
OUVR	3,762	0,051	14,902	0,730	19,599	19,405
INAC	1,206	71,792	0,702	0,807	2,623	3,361





Questions:

1. Pourquoi a-t-on choisi de réaliser une A.C.P
2. Commenter la matrice de corrélations
3. Donner la valeur de l'inertie du nuage de points et son interprétation statistique.
4. Rappeler les règles usuelles de sélection des axes en ACP. Combien d'axes proposez-vous d'interpréter ?
5. Interpréter les 2 premières composantes principales (cercle de corrélation, plan des individus, plan des individus-variables)

Solution

1. On choisit l'ACP car le tableau de données de type Individus x Variables (toutes les variables sont quantitatives)
2. Matrice de corrélation
   - ✓ Forte corrélation négative entre (PAA et PAO) ; (RAI et PAO) ; (PLP et PAO) ; (RAI et LEC)

- ✓ Forte corrélation positive entre (PAO et LEC) ; (PAA et RAI), (PAA et PLP) ; (RAI et PLP)

3. L'inertie du nuage de points (inertie totale)

$$I_t = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 = 5{,}999$$

I_t Représente l'information totale (la variance totale)

4. Les règles usuelles pour sélectionner les principaux axes factorielles (les composantes principales) :
   - ✓ Le test du coude (critère du Kaiser) : on retient les axes correspondants à des valeurs propres supérieures ou égales à 1.
   - ✓ Dans cet exercice, il existe une seule valeur ≥ 1, d'où la composante principale à retenir est celle correspondant à la valeur propre $\lambda_1 = 4{,}63$) (c'est-à-dire la première composante principale F1)

F1 explique 77% de l'inertie totale $\left(\frac{\lambda_1}{I_t} = \frac{4{,}63}{5{,}999} = 0{,}77\right)$

5. Interprétation des résultats (premier plan factoriel F1xF2)

   Le premier plan factoriel explique 90,56% de l'inertie totale avec F1 (77%) et F2 (13,39%).

   L'interprétation es axes factoriel se fait séquentiellement pour chaque axe et chaque nuage de points en regardant les contributions de l'axe en question.

   - ✓ Axe 1 (F1) :

Analyse dans $\mathbb{R}^n$

Les variables qui contribuent le plus à la formation de F1 (les corrélations entre les variables et F1 sont proche de 1) sont : LEC, PAO, PAA, RAI, et PLC.

D'où, l'axe F1 permet d'opposer les individus qui consomment du pain ordinaire (PAO) et légumes secs (LEC) à ceux qui consomment Autre pain (PAA) raisin (RAI) et plats préparés (PLP) $\Longrightarrow$ F1 représente les aliments ordinaires à bon marché et les aliments chers.

Analyse dans $\mathbb{R}^p$

Les individus qui contribuent le plus à la formation du premier axe F1 sont les suivants :

CSUP, SAPG et AGRI

F1 permet d'opposer les habitudes alimentaires des cadres supérieures (CSUP) et les exploitants agricoles(AGRI) et les salariés du secteur agricole (SAA)

Conclusion

L'axe F1 résume l'opposition existante entre les catégories socioprofessionnelles en matière d'alimentation (opposition entre les catégories professionnelles qui consomment des produits de base et celles qui consomment des produits de luxe)

Exercice 3

Une étude sur des fournisseurs de matériel informatique a conduit à apprécier le service, la qualité et le prix de quatre fournisseurs. Pour cela un expert a noté ces entreprises avec des notes allant de -3 à 3. Les résultats sont consignés ci-dessous :

Entp.	Service	Qualité	Prix
E1	-2	3	-1
E2	-1	1	0
E3	2	-1	-1
E4	1	-3	2

1. Calculer le vecteur moyen des individus. Qu'en conclure?
2. Calculer la matrice Variance-Covariance $V = \frac{1}{n} X'X$. Interpréter

On veut faire une ACP non normée;

3. Sur quelle matrice faut-il travailler? Vérifier qu'elle admet une valeur propre nulle. Qu'est-ce que cela implique?
4. On donne λ1 = 7,625. En déduire λ2
5. Calculer les pourcentages d'inertie. Quelle dimension retenez-vous?
6. Soient les vecteurs propres U1 et U2 :

U1	U2
-0,503	-0,643
0,808	-0,114
-0,305	0,757

Calculer les composantes principales F1 et F2.

7. Représenter les individus et les variables dans le plan principal (F1, F2). Interpréter

8. Calculer la corrélation entre les variables initiales et les composantes principales.

Solution

1. Le vecteur moyen des individus (le centre de gravité $\bar{X}_g$)

$$\bar{X}_g=\begin{pmatrix}\overline{Service}\\ \overline{Qualité}\\ \overline{Prix}\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

Les données sont centrées par rapport à leurs moyennes

2. La matrice Var/Cov (**V**)

$$V=\frac{1}{n}X'X=\frac{1}{4}\begin{pmatrix}-2&-1&2&1\\3&1&-1&-3\\-1&0&-1&2\end{pmatrix}\begin{pmatrix}-2&3&-1\\-1&1&0\\2&-1&-1\\1&-3&2\end{pmatrix}=\begin{pmatrix}2,5&-3&0,5\\-3&5&-2\\0,5&-2&1,5\end{pmatrix}$$

Interprétation

Les valeurs qui se trouvent sur la première diagonale de la matrice V sont des variances

$$\begin{cases}V(X1)=V(Service)=5/2\\V(X2)=V(qualité)=5\\V(X3)=V(prix)=3/2\end{cases}$$

Les valeurs qui se trouvent au-dessous (au-dessus) la première diagonale sont des covariance :

$$\begin{cases}Cov(X1,X2)=Cov(X2,X1)=-3\\Cov(X1,X3)=Cov(X3,X1)=1/2\\Cov(X2,X3)=Cov(X3,X2)=-2\end{cases}$$

Conclusion

Il existe une relation positive entre X1 et X3 et une relation négative entre X1etX2 ; X2 et X3

3. On doit travailler sur la matrice Var/Cov (V) car est une ACP non normée.
   - ✓ Vérification que la matrice V admet une valeur propre nulle

     Soit λ une valeur propre de de V, alors : $\left|V_{(3,3)}-\lambda I_3\right|=0$

Pour $\lambda=0\Longrightarrow\left|V_{(3,3)}-0I_3\right|=\begin{pmatrix}2,5&-3&0,5\\-3&5&-2\\0,5&-2&1,5\end{pmatrix}=1/2\begin{pmatrix}5&-6&1\\-6&10&-4\\1&-4&3\end{pmatrix}$

$$\Longrightarrow |V_{(3,3)}| = 1/2 \begin{vmatrix} 5 & -6 & 1 \\ -6 & 10 & -4 \\ 1 & -4 & 3 \end{vmatrix}$$

$$\Longrightarrow |V_{(3,3)}| = 1/2 \left[5 \begin{vmatrix} 10 & -4 \\ -4 & 3 \end{vmatrix} - (-6) \begin{vmatrix} -6 & -4 \\ 1 & 3 \end{vmatrix} + (1) \begin{vmatrix} -6 & 10 \\ 1 & -4 \end{vmatrix}\right] = 0$$

D'où $\lambda = 0$ est une valeur propre de de V $\Longrightarrow$le plan factoriel (F1xF2) explique 100% de l'inertie totale (information totale).

4. Nous avons λ_1= 7,625, on cherche la valeur de λ_2

   La variance totale= l'information totale= l'inertie totale

   $$=V(X1)+V(X2)+V(X3)=\sum_{j=1}^{3} \lambda_j = \lambda_1 + \lambda_2 + \lambda_3$$

   $$=2,5+5+1,5=7,625+\lambda_2 + 0 \Longrightarrow \lambda_2 = 9 - 7,625$$

   $$\Longrightarrow \lambda_2 = 1,375$$

5. Les pourcentages d'inerties

   $\lambda_1 \longrightarrow F1 \longrightarrow I_{F1} = \frac{7,625}{9} = \frac{\lambda_1}{I_t} = 84,74\%$

   $\lambda_2 \longrightarrow F2 \longrightarrow I_{F2} = \frac{\lambda_2}{I_t} = \frac{1,375}{9} = 15,27\%$

   $\lambda_3 \longrightarrow F3 \longrightarrow I_{F3} = \frac{\lambda_3}{I_t} = \frac{0}{9} = 0\%$

   Remarque I_t : l'inertie totale

   Les dimensions retenues (les facteurs principaux retenus) sont F1 et F2

6. Soient $U_1 = \begin{pmatrix} -0,503 \\ 0,808 \\ -0,305 \end{pmatrix}$ et $U_2 = \begin{pmatrix} -0,643 \\ -0,114 \\ 0,757 \end{pmatrix}$ deux vecteurs propres correspondants à λ_1 et λ_2 respectivement.

   Calcul des coordonnées des composantes principales F1 et F2 :

   $$F1 = XU_1 = \begin{pmatrix} -2 & 3 & -1 \\ -1 & 1 & 0 \\ 2 & -1 & -1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} -0,503 \\ 0,808 \\ -0,305 \end{pmatrix} = \begin{pmatrix} 3,737 \\ 1,311 \\ -1,509 \\ -3,538 \end{pmatrix}$$

   $$F2 = XU_2 = \begin{pmatrix} -2 & 3 & -1 \\ -1 & 1 & 0 \\ 2 & -1 & -1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} -0,643 \\ -0,114 \\ 0,757 \end{pmatrix} = \begin{pmatrix} 0,185 \\ 0,528 \\ -1,928 \\ 1,214 \end{pmatrix}$$

Calcul coordonnées des variables

$$G1 = \sqrt{\lambda_1}U_1 = \sqrt{7,625}\begin{pmatrix}-0,503\\0,808\\-0,305\end{pmatrix} = \begin{pmatrix}-0,503\\0,808\\-0,305\end{pmatrix}$$

$$G2 = \sqrt{\lambda_2}U_2 = \sqrt{1,375}\begin{pmatrix}-0,643\\-0,114\\0,757\end{pmatrix} = \begin{pmatrix}-0,755\\-0,134\\0,889\end{pmatrix}$$

7. Représentation graphique

   Interprétation

   Sur F1 :

   On peut distinguer deux groupes d'entreprises : le premier groupe (E1 et E2) et le deuxième groupe (E3 et E4)

   Les entreprises du premier groupe fournissent des produits de qualité

   Les entreprises du deuxième groupe offrent un bon service aux clients.

Exercice 4

Ci-dessous les résultats d'une analyse en composantes principales (ACP) d'un tableau représentant six types de productions agronomiques pendant huit années consécutives.

	X1(Riz)	X2(Patates)	X3(Agrumes)	X4(Bananes)	X5(Melons)	X6(Tomates)
2001	210	95	80	60	32	92
2002	207	88	98	64	33	98
2003	150	82	109	72	31	106
2004	123	85	119	79	28	112
2005	92	96	144	72	60	123
2006	68	104	169	71	66	133
2007	93	105	152	59	54	164
2008	95	119	165	75	70	98
Moyennes	129,8	95,5	129,5	69,0	46,8	115 ,8
Ecart types	54,3	14,0	32,8	7,2	17, 5	23,9

Matrice de corrélation (Pearson (n)) :

Variables	X1(Riz)	X2(Patates)	X3(Agrumes)	X4(Bananes)	X5(Melons)	X6(Tomates)
X1(Riz)	**1**	-0,543	**-0,947**	-0,435	**-0,781**	-0,655
X2(Patates)	-0,543	**1**	**0,726**	-0,063	**0,869**	0,248
X3(Agrumes)	**-0,947**	**0,726**	**1**	0,331	**0,903**	0,588
X4(Bananes)	-0,435	-0,063	0,331	**1**	0,127	-0,247
X5(Melons)	**-0,781**	**0,869**	**0,903**	0,127	**1**	0,405
X6(Tomates)	-0,655	0,248	0,588	-0,247	0,405	**1**

Les valeurs en gras sont différentes de 0 à un niveau de signification alpha=0,05

Test de sphéricité de Bartlett :

Khi² (Valeur observée)	44,103
Khi² (Valeur critique)	24,996
DDL	15
p-value	0,000
alpha	0,05

	F1	F2	F3	F4	F5	F6
Valeur propre	3,772	1,251	0,871	0,081	0,021	0,004
Variabilité (%)	62,866	20,857	14,521	1,350	0,344	0,061
% cumulé	62,866	83,724	98,244	99,594	99,939	100,000

Composantes principales

Observation	F1	F2	F3	F4	F5	F6
2001	-2,576	-1,023	-1,054	-0,089	-0,222	-0,025
2002	-2,309	-0,466	-0,405	0,202	0,284	0,052
2003	-1,599	0,713	0,696	0,134	0,006	-0,096
2004	-0,948	1,567	1,089	-0,426	-0,018	0,042

2005	1,118	0,384	0,302	0,461	-0,172	0,090
2006	2,342	0,074	0,223	0,233	0,050	-0,081
2007	1,733	-2,182	0,901	-0,277	0,024	0,014
2008	2,239	0,933	-1,753	-0,238	0,048	0,003

Coefficients de corrélation entre variables et composantes principales :

	F1	F2	F3	F4	F5	F6
X1(Riz)	-0,937	-0,188	-0,281	0,006	0,089	0,031
X2(Patates)	0,784	-0,195	-0,559	-0,185	-0,015	0,008
X3(Agrumes)	0,988	0,098	0,046	0,014	0,110	-0,023
X4(Bananes)	0,235	0,962	0,125	-0,058	0,006	0,026
X5(Melons)	0,932	-0,046	-0,296	0,200	-0,021	0,024
X6(Tomates)	0,616	-0,491	0,612	-0,058	0,007	0,029

Contributions des variables (%) :

	F1	F2	F3	F4	F5	F6
X1(Riz)	23,252	2,828	9,032	0,051	37,934	26,903
X2(Patates)	16,304	3,048	35,873	42,044	1,134	1,598
X3(Agrumes)	25,864	0,763	0,239	0,234	58,432	14,469
X4(Bananes)	1,461	73,924	1,804	4,116	0,159	18,536
X5(Melons)	23,044	0,167	10,063	49,394	2,092	15,239
X6(Tomates)	10,075	19,270	42,989	4,161	0,250	23,255

Questions :

1) Analyser la matrice de corrélation
2) Quelle est le pourcentage de l'inertie du premier plan factoriel ?
3) Quels sont les individus et les variables qui contribuent le plus à la construction du 1er axe factoriel ?, donner une signification à cet axe
4) Quelle est la variable qui contribue le plus à la construction du 2^{ième} axe factoriel ? donner une signification à cet axe.

5) Représenter les individus et les variables sur le 1er axe factoriel ? les années 2003 et 2006 sont-elles bien représentées ?
6) Donner une présentation graphique des années sur le primer plan factoriel (F1 et F2) ?, interpréter la ?

Chapitre 4 : Analyse factorielle des correspondances, et Analyse factorielle des correspondances multiples.

4.1. Analyse factorielle des correspondances (AFC)

4.1.1. Généralités

- ✓ L'AFC est une méthode qui s'applique aux tableaux de contingence (dites aussi tableaux de dépendances, ou tableaux croisés).
- ✓ Un tableau de contingence montre la répartition d'un échantillon, ou une population, selon deux variables qualitatives.
- ✓ L'AFC est une méthode descriptive d'analyse de données, développée par Benzécri en 1965, qui permet d'étudier les relations entre deux variables qualitatives via la comparaison des profils-lignes et profils-colonnes.
- ✓ Soient deux variables qualitatives X et Y avec k et p modalités respectivement. Les deux variables sont observées sur un échantillon de n individus. Le tableau de contingence (tableau croisé) prend la forme suivante :

Tableau 4.1. Tableau de contingence

X/Y	y_1	y_2	……..	y_j	…..	y_p	$Total$
x_1	n_{11}	n_{12}	……	n_{1j}	……	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	……	n_{2j}	…….	n_{2p}	$n_{2.}$
….							….
x_i	n_{i1}		…..	n_{ij}	…..	n_{ip}	$n_{i.}$
…..							…..
x_k	n_{k1}			n_{kj}		n_{kp}	$n_{k.}$
$Total$	$n_{.1}$	$n_{.2}$	……	$n_{.j}$	…..	$n_{.p}$	n

Tels que :

n_{ij} : Effectif commun à la ligne i et la colonne j (la répartition des effectifs communs est appelée distribution conjointe)

$n_{i.} = \sum_{i=1}^{k} n_{ij}$: Effectif marginal de la i^{ième} ligne (ditribtion marginale -ligne)

$n_{.j} = \sum_{i=1}^{k} n_{ij}$: Effectif marginal de la j^{ième} colonne (distribution marginale-colonne)

$n = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j}$: Effectif total

Profils-lignes :

On entend par profil ligne, les fréquences empiriques des modalités de la variable Y, il s'obtient en divisant l'effectif de chaque modalité par le total de ligne correspondant. Ci-dessous le tableau des profils lignes (noté PL).

$PL_i = (n_{ij}/n_{i.})$:i^{ième} profil ligne

Tableau 4.2. Profils-lignes

X/Y	y_1	y_2	……..	y_j	…..	y_p	$Total$
x_1	$n_{11}/n_{1.}$	$n_{12}/n_{1.}$	……	$n_{1j}/n_{1.}$	……	$n_{1p}/n_{1.}$	**1**
x_2	$n_{21}/n_{2.}$	n_{22}	……	$n_{2j}/n_{2.}$	…….	$n_{2p}/n_{2.}$	1
….							….
x_i	$n_{i1}/n_{i.}$		…..	$n_{ij}/n_{i.}$	…..	$n_{ip}/n_{i.}$	1
…..							…..
x_k	$n_{k1}/n_{k.}$			$n_{kj}/n_{k.}$		$n_{kp}/n_{k.}$	1
Profils-moyens	$n_{.1}/\ n$	$n_{.2}/n$	……	$n_{.j}/N$		$n_{.p}/n$	1

$PL_i = \left(\frac{n_{ij}}{n_{i.}}\right) * 100$: i^{ième} profil ligne

Tableau 4. 3. Profils-lignes en pourcentages

X/Y	y_1	y_2	……..	y_j	…..	y_p	$Total$
x_1	$(\frac{n_{11}}{n_{1.}})100$	$(\frac{n_{12}}{n_{1.}})100$	……	$(\frac{n_{1j}}{n_{1.}})100$	……	$(\frac{n_{1p}}{n_{1.}})100$	**100**
x_2	$(\frac{n_{21}}{n_{2.}})100$	$(\frac{n_{21}}{n_{2.}})100$	……	$(\frac{n_{21}}{n_{2.}})100$	…….	$(\frac{n_{21}}{n_{2.}})100$	100
….							….
x_i	$(\frac{n_{i1}}{n_{i.}})100$		…..	$(\frac{n_{ij}}{n_{i.}})100$	…..	$(\frac{n_{ip}}{n_{2.}})100$	100
…..							…..
x_k	$(\frac{n_{k1}}{n_{2.}})100$			$(\frac{n_k}{n_{2.}})100$		$(\frac{n_{21}}{n_{2.}})100$	100

Profils-Colonnes

On entend par profil-colonne, les fréquences empiriques des modalités de la variable X, il s'obtient en divisant l'effectif de chaque modalité par le total de colonne correspondant. Ci-dessous le tableau des profils-colonnes (noté PC).

$PC_j = (n_{ij}/n_{.j})$: j^{ième} profil colonne

Tableau 4.4. Profils-Colonnes

X/Y	y_1	y_2	……..	y_j	…..	y_p	Profils moyens
x_1	$n_{11}/n_{.1}$	$n_{12}/n_{.2}$	……	$n_{1j}/n_{.j}$	……	$n_{1p}/n_{.p}$	$n_{1.}/n$
x_2	$n_{21}/n_{.1}$	$n_{22}/n_{.2}$	……	$n_{2j}/n_{.j}$	…….	$n_{2p}/n_{.p}$	$n_{2.}/n$
….							….
x_i	$n_{i1}/n_{.1}$		…..	$n_{ij}/n_{.j}$	…..	$n_{ip}/n_{.p}$	$n_{i.}/n$
…..							…..
x_k	$n_{k1}/n_{.1}$			$n_{kj}/n_{.j}$		$n_{kp}/n_{.p}$	$n_{k.}/n$
$Total$	1	1	……	1	…..	1	1

Tableau 4.5. Profils-Colonnes en pourcentages

X/Y	y_1	y_2	……..	y_j	…..	y_p	Profils moyens
x_1	$\left(\frac{n_{11}}{n_{.1}}\right)100$		……		……		$\frac{n_{1.}}{n}100$
x_2	$(\frac{n_{21}}{n_{.1}})100$		……		…….		
….							
x_i	$\left(\frac{n_{i1}}{n_{.1}}\right)100$		…..	$\left(\frac{n_{ij}}{n_{.j}}\right)100$	…..		
…..							
x_k	$\left(\frac{n_{k1}}{n_{.1}}\right)100$						$(\frac{n_{k.}}{n})100$
$Total$	100	100	……	100	…..	100	100

Remarque 4.1.

- *Le test de Khi-deux est utilisé pour tester l'indépendance entre X et Y*

Donc, deux variables X et Y sont indépendants $\Rightarrow \forall\,(i,j), \frac{n_{ij}}{n} \cong \frac{n_{i.}}{n}\frac{n_{.j}}{n} \Leftrightarrow \forall\,(i,j), \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$

$$\Leftrightarrow \forall\,(i,j), \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

Cela signifie que l'égalité des profils-lignes et profils colonnes est une condition nécessaire et suffisante pour que les deux variables X et Y soient indépendantes.

4.2. Fondements théoriques de la méthode AFC

L'application de l'AFC se fait en plusieurs étapes :

- Transformation sur le tableau des données de contingence pour obtenir le tableau des profils-lignes, ou le tableau des profils-colonnes.
- Présentation graphique du nuage des profils-lignes et le nuage des profils-colonnes.

- Etudier la dépendance, voir l'indépendance, entre les deux variables X et Y se fait à travers l'utilisation de la méthode ACP sur le tableau des données des profils-lignes, ou bien sur les le tableau des profils-colonnes.

a. Formulation matricielle du tableau des profils-lignes & tableau des profils-colonnes

Soit N le tableau de contingence de k lignes et p colonnes

$$N=\begin{pmatrix} n_{11} & n_{12} & & n_{1j} & n_{1p} \\ n_{21} & n_{22} & & n_{2j} & n_{2p} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ n_{k1} & \vdots & & n_{kj} & n_{kp} \end{pmatrix}$$

On définit une matrice diagonale des totaux marginaux lignes (D_{tl}), et une matrice diagonale des totaux marginaux colonnes (D_{tc})

$$D_{tl}=\begin{pmatrix} n_{1.} & 0 & \cdots & 0 \\ 0 & n_{2.} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & n_{k.} \end{pmatrix} \text{ et } D_{tc}=\begin{pmatrix} n_{.1} & 0 & \cdots & 0 \\ 0 & n_{.2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & n_{.p} \end{pmatrix}$$

Donc,

La formulation matricielle du tableau des profils-lignes s'écrit :

$$T_{pl}=D_{tl}^{-1}N=\begin{pmatrix} 1/n_{1.} & 0 & \cdots & 0 \\ 0 & 1/n_{2.} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/n_{k.} \end{pmatrix}\begin{pmatrix} n_{11} & n_{12} & & n_{1j} & n_{1p} \\ n_{21} & n_{22} & & n_{2j} & n_{2p} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ n_{k1} & \vdots & & n_{kj} & n_{kp} \end{pmatrix}$$

La formulation matricielle du tableau des profils-colonnes s'écrit :

$$T_{pc}=D_{tc}^{-1}N=\begin{pmatrix} 1/n_{.1} & 0 & \cdots & 0 \\ 0 & 1/n_{.2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/n_{.p} \end{pmatrix}\begin{pmatrix} n_{11} & n_{12} & & n_{1j} & n_{1p} \\ n_{21} & n_{22} & & n_{2j} & n_{2p} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ n_{k1} & \vdots & & n_{kj} & n_{kp} \end{pmatrix}$$

b. Le nuage des profils-lignes

Il est à noter que les profils-lignes seront représentés dans $\mathbb{R}^P$ (i.e. le nuage profils-lignes comporte p points). Selon la nature de relation entre X et Y, on distingue deux cas :

- Si les deux variables X et Y sont indépendantes, alors les profils-lignes sont égaux et les p points profils-lignes sont identiques, ce qui donne un seul point)
- Si les deux variables X et Y sont dépendantes, alors nous ferons appel à la méthode ACP pour étudier cette dépendance. De ce fait, les modalités de la variable X sont considérées comme des individus. Chaque individu est affecté d'un poids $f_{i.}=\frac{n_i}{n}$ pour faire apparaitre son importance dans l'échantillon. En outre, l'application de l'ACP nécessite la définition d'une métrique qui permet de mesurer la distance entre deux profils-lignes.

c. Etude de la dépendance entre X et Y

l'application de l'ACP sur le tableau des profils-lignes en considérant les modalités de la variables X les individus, ce qui donne un nuage des profils-lignes de p points, chaque individu affecté un poids $f_{i.} = \frac{n_{i.}}{n}$, et son centre de gravité g s'obtient en utilisant la formule suivante :

$$g_{pl} = \begin{pmatrix} \frac{n_{.1}}{n} \\ . \\ . \\ . \\ \frac{n_{.p}}{n} \end{pmatrix} = \begin{pmatrix} f_{.1} \\ . \\ . \\ . \\ f_{.p} \end{pmatrix}$$ sont les profils-moyens

Le nuage de p points profils-lignes est représenté dans un sous espace vectoriel de (p-1) dimensions.

c. Définition d'une métrique de Khi-deux

La définition d'une métrique qui permet de calculer la distance entre deux profils-lignes est cruciale pour l'application de l'ACP. Pour ce faire, la métrique de Khi-deux est utilisée dans ce genre d'analyse. Cette métrique est définie comme suit :

Soient PL_i et $PL_{i'}$ deux profils-lignes, alors la distance entre eux s'obtient :

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

M est la métrique, matrice diagonale, $M = nD_{tc}^{-1} = n \begin{pmatrix} 1/n_{.1} & 0 & 0 \\ 0 & 1/n_{.2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1/n_{.p} \end{pmatrix} =$

$$\begin{pmatrix} n/n_{.1} & 0 & 0 \\ 0 & n/n_{.2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & n/n_{.p} \end{pmatrix}$$

D'où la matrice à diagonaliser s'écrit :

L'application de l'ACP sur le tableau des profils-lignes nous amène à diagonaliser la matrice des variances covariances (V) multipliée à droite par la métrique M.

Nous avons :

- La métrique de Khi-deux : $M = nD_{tc}^{-1}$,
- Le tableau des données profils-lignes : $T_{pl} = D_{tl}^{-1} N$
- Le tableau des données centrées : $\hat{T}_{pl} = D_{tl}^{-1} N - g_{pl} g'_{pl}$
- La matrice des variances-covariances : $V = \frac{1}{n} N' D_{tl}^{-1} N - g_{pl} g'_{pl}$

Donc ; la matrice à diagonaliser prendra s'écrit :

$$VM = N' D_{tl}^{-1} N D_{tc}^{-1} - n g_{pl} g'_{pl} D_{tc}^{-1}$$

Remarque 4.2.

g_{Pl} : le centre de gravité du nuage des profils-lignes (i.e. profils moyens) est un vecteur propre de VM associé à la valeur propre 0 et 1. $\Longrightarrow$ Diagonalsier VM $\Leftrightarrow$Diagonaliser $S = N' D_{tl}^{-1} N D_{tc}^{-1}$

d. Composantes principales

Les cordonnées des profils-lignes sur les axes principaux sont données par :

$F_{i,\alpha} = n(N' D_{tl}^{-1} N D_{tc}^{-1})' D_{tc}^{-1} U_\alpha = n S' D_{tc}^{-1}$ (pour $\lambda_\alpha \neq 1, \alpha = 1, \dots p-1$)

Les composantes principales $G_{i,\alpha}$ sont centrées au centre de gravité et leurs variances égales à λ_α.

Remarque 4.2

La même démarche sera suivie pour le cas du nuage profils-colonnes. L'application de l'ACP sur le tableau des profils –colonnes nous amène à diagonaliser la matrice des variances covariances (V) multipliée à droite par la métrique M.

Nous avons :

- *La métrique de Khi-deux : $M = n D_{tc}^{-1}$,*
- *Le tableau des données profils-lignes : $T_{pc} = D_{tc}^{-1} N$*

Donc ; la matrice à diagonaliser prendra la forme matricielle suivante :

$$VM = N D_{tc}^{-1} N' D_{tl}^{-1} - n g_{pc} g'_{pc} D_{tl}^{-1}$$

Comme g_{pc} le centre de gravité du nuage des profils-colonnes (profils moyens) est un vecteur propre de VM associé à la valeur propre 0 et 1. $\Longrightarrow$ Diagonalsier VM $\Leftrightarrow$ Diagonaliser $R = N D_{tc}^{-1} N' D_{tl}^{-1}$

Les composantes principales de profils-colonnes s'écrivent :

$$G_{j,\alpha} = n D_{tc}^{-1} N' D_{tl}^{-1} V_\alpha$$

Comme la matrice R a été diagonalisée, alors nous utilisons la formule de transition pour calculer les coordonnées des composantes principales :

$$G_{j,\alpha} = n D_{tc}^{-1} N' {}^{D_{tl}^{-1} V_\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{k} \frac{n_{ij}}{n_{.j}} G_{i,\alpha} \text{ (pour } \lambda_\alpha \neq 1, \alpha = 1, \dots p-1)$$

Exemple d'application de l'AFC: (cet exemple est tiré du help xl-stat. Il est disponible sur le site https://help.xlstat.com/fr/6380-correspondence-analysis-ca-contingency-table. Visité le 15/06/2023)

Les données du tableau de contingence sont issues d'une enquête auprès d'un échantillon de 1375 personnes en leurs interrogeant sur leurs avis à propos un film regardé ainsi que leurs tranches d'âges.

Tableau 4.6. Données de l'enquête sur le film regardé ainsi que les tranches d'âge

X/Y	MAUVAIS	MOYEN	BON	TRÈS BON	total
16-24	69	49	48	41	207
25-34	148	45	14	22	229
35-44	170	65	12	29	276
45-54	159	57	12	28	256
55-64	122	26	6	18	172
65-74	106	21	5	23	155
75+	40	7	1	14	62
total	814	270	98	175	1357

Source : https://help.xlstat.com/fr/6380-correspondence-analysis-ca-contingency-table

Donc, à travers l'application de l'analyse des correspondances du tableau ci-dessus, nous tentons d'étudier d'éventuelles relations entre les deux variables étudiées, (elles sont mentionnées en lignes et en colonnes), ainsi que les similitudes qui peuvent exister entre les modalités de chaque variable. Les résultats de l'ZFC en utilisant le logiciel XL-stat sont présentés ci-dessous :

- Test d'indépendance entre les lignes et les colonnes :

  L'AFC ne peut être appliquée que dans le cas où les lignes et colonnes sont dépendantes. C'est la raison pour laquelle qu'on a fait appel au test de khi-deux. Les résultats du test sont cosignés dans le tableau ci-dessous :

Test d'indépendance entre les lignes et les colonnes :

Khi² (Valeu	148,268
Khi² (Valeu	28,869
DDL	18
p-value	< 0,0001
alpha	0,05

Comme la probabilité de rejeter H_0 à tort est inférieure à 1%, allers on accepte H_1, ce qui veut dire il existe un lien entre les lignes et les colonnes du tableau de contingence.

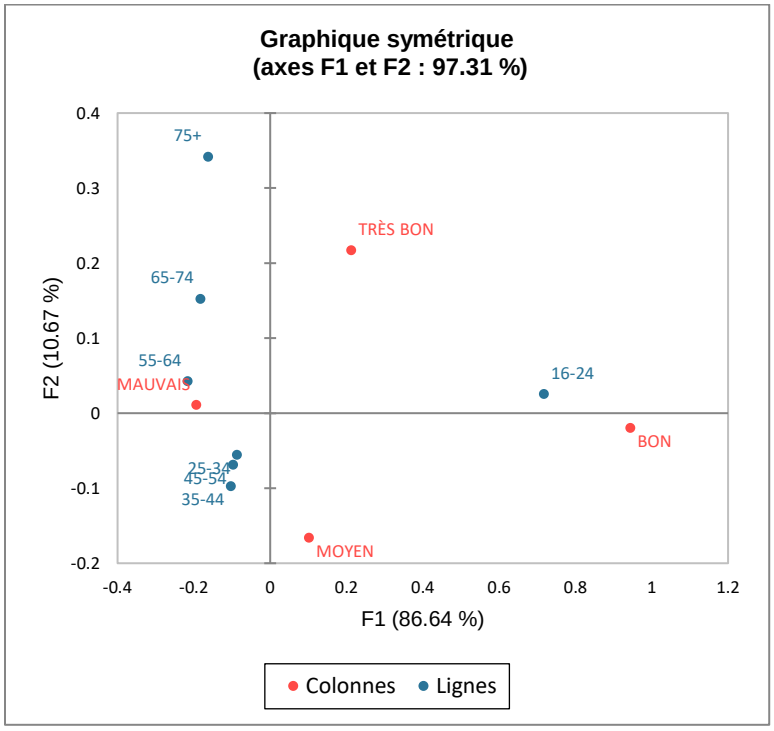
Valeurs propres et pourcentages d'inertie :

	F1	F2	F3
Valeur pro	0,095	0,012	0,003
Inertie (%)	86,640	10,674	2,685
% cumulé	86,640	97,315	100,000

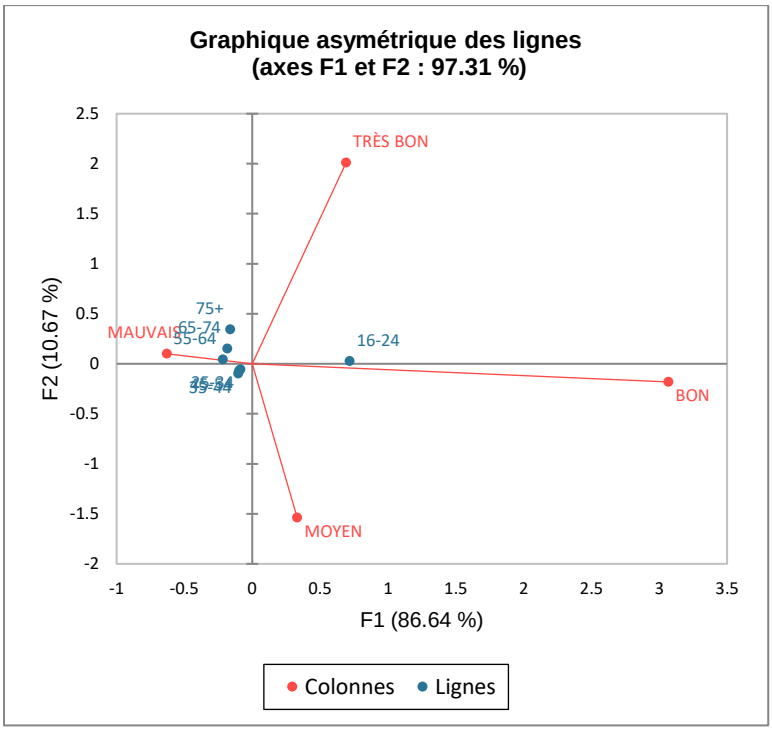
Résultats relatifs aux Profils-lignes

Profils (lignes) :

	MAUVAIS	MOYEN	BON	TRÈS BON	Somme
16-24	0,333	0,237	0,232	0,198	1
25-34	0,646	0,197	0,061	0,096	1
35-44	0,616	0,236	0,043	0,105	1
45-54	0,621	0,223	0,047	0,109	1
55-64	0,709	0,151	0,035	0,105	1
65-74	0,684	0,135	0,032	0,148	1
75+	0,645	0,113	0,016	0,226	1
Moyenne	0,608	0,184	0,067	0,141	1



Le graphique symétrique montre que la distance entre les groupes d'âges (les profils-lignes) 25-34 ans , 35-54 ans et 45-54 ans est très faible, ils sont presque superposés. Cela signifie que ces groupes sont presque similaires (profils lignes similaires) .



Sur le graphique asymétrique des lignes, les colonnes sont représentées dans l'espace des lignes (coordonnées standards pour les colonnes et coordonnées principales pour les lignes). Le graphique montre que la première dimension permet d'opposer la modalité « Bon » à « Mauvais ». La tranche d'Age 16-24 ans se trouve à coté la modalité « Bon », ce qui signifie les clients plus jeunes ont tendance à répondre par « Bon », contrairement aux classes d'âge 55-64 ans, 65- 74 ans et 75 ans et plus qui répondent souvent par « Mauvais »

4.3. Analyse factorielle des correspondances multiples (AFCM)

- La méthode AFCM une est analyse factorielles d'analyse de données qualitatives multidimensionnelles, elle est une extension de la méthode ACF au cas où le nombre de variables est égal à trois et plus.
- L'AFCM est la méthode d'analyse de données la plus utilisée pour analyser les données collectées dans le cadre des enquêtes socioéconomiques où le questionnaire comporte des questions à choix multiples
- Chaque individu est décrit par p variables qualitatives
- L'AFCM s'applique sur des tableaux des tableaux disjonctifs complets.

4.3.1. Principes de l'AFCM

Dans une enquête, les n individus enquêtés sont décrits par des modalités de p variables qualitatives (i.e. chaque variable X^s comporte un nombre $m_s$de modalités, et chaque modalité est désignée par un code choisi de façon arbitraire). En effet, chaque variable est une question à choix multiples. Par conséquent, les données brutes de l'enquête seront présentées dans un tableau rectangulaire de n lignes (pour les individus) et p colonnes (pour les modalités de p variables). A titre d'exemple, la variable catégorie professionnelle a quatre modalités, alors que chaque individu est décrit par une seule modalité. Elle sera codée comme suit :

$$\text{Variable } X^1\text{: cétégrie scioprofessionnlle} \begin{cases} \text{exécutant: } 1 \\ \text{maitrise: } 2 \\ \text{cadre: } \quad 3 \\ \text{cadre supérieur : } 4 \end{cases}$$

Donc, le tableau des données contient que des codes numériques: Exemple, le tableau X ci-dessous montre 6 individus décrits par trois variables qualitatives (la première possède 2 modalités, la deuxième deux modalités, et la troisième trois modalités). X est le tableau de données brutes (données de l'enquête codifiées)

$$X_{n,p} = \begin{bmatrix} 2 & 2 & 2 \\ 2 & 1 & 1 \\ 1 & 1 & 3 \\ 1 & 2 & 3 \\ 2 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix}$$

A partir du tableau brut X, on peut définir un tableau disjonctif contenant que des codes binaires (0 et 1) ((i.e. chaque modalité d'une variable X^s est désignée par une seule colonne). Pour ce faire, on va créer 2 colonnes par la variable X^1 ($m_1 = 2$), 2 colonnes pour la variable X^2 ($m_2 = 2$), et 3 colonnes pour la variable X^3 ($m_3 = 3$) . Donc, le tableau disjonctif complet Z se présente avec 5 lignes (n=5) et 7 colonnes ($q = m_1 + m_2 + m_3 = 7$):

Tableau disjonctif complet

$$Z_{n,q} = (X^1|X^2|X^3) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Tels que :

- n : nombre d'individus
- $q = m_1 + m_2 + \cdots m_p = \sum_{s=1}^{p} m_s$: le nombre total des modalités de p variables (les p questions dans un questionnaire)
- Le tableau contient soit $Z_{ij} = 1$ ou $Z_{ij} = 0$

Propriétés

- $Z_{i.} = \sum_{j=1}^{q} Z_{ij} = p$: La somme de chaque ligne = p =3 (nombre de variables, ou nombre de questions)
- $Z_{.j} = \sum_{i=1}^{n} Z_{ij}$: La somme de chaque colonne donne l'effectif marginal des individus décrits par une modalité de la variable X^s

  Exemple, l'effectif marginal de la première modalité de la première variable 3 (i.e. 3 individu sont décrits par la modalité en question)
- $\sum_{j \in s} Z_{.j} = n = 6$: la somme de chaque sous tableau correspondant aux modalités de la variables X^s est égale au nombre total des individus
- $\sum_{i=1}^{n} \sum_{j=1}^{q} Z_{ij} = np$ = (6)(3)=18
- i[ième] ligne contient $(m_j - 1)$ fois la valeur zéro (0) et une seule fois la valeur 1 correspondant à la modalité de la variable X^s

D'où, la forme générale du tableau disjonctif se présente comme suit :

Tableau 4. 7. Tableau disjonctif complet

IND/Var.	X_1^1	X_2^1	…	$X_{m_1}^1$	...	X_1^s	…	$X_{m_s}^s$	….	X_1^p	…	$X_{m_p}^p$	Somme $Z_{i.}$
1	1	0	….	0		0	…	1		1	..	0	p
2	0	1		0		1		0		1		0	P
……													P
i													P
…													P
n													P
Effectif marginal $Z_{.j}$	$n_{X_1^1}$	$n_{X_2^1}$		$n_{X_{m_1}^1}$				$n_{X_{m_s}^s}$				$n_{X_{m_p}^p}$	

Tableau de BURT

A partir du tableau disjonctif complet, le tableau de contingence de BURT se construit comme suit :

$$B_{(q,q)} = Z'Z$$

Ci-dessous un exemple concret du tableau du Burt construit à partir des données d'une enquête extraites du cite https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-des-correspondances-multiples-acm-ou-afcm. Il s'agit des réponses de 28 individus sur 4 questions à choix multiples. La question 1 comporte deux réponses, la deuxième question 3 réponses et la troisième question 5 réponses, tandis que la dernière possède 2 réponses. Donc, le tableau de Burt croise 4 variables qualitatives deux à deux, ce qui donne au total 16 sous tableaux de contingence.

Tableau 4.8. Exemple du tableau du Burt.

	Satisfait-Non	Satisfait-Oui	Réparé-NSP	Réparé-Non	Réparé-Oui	Accueil-1	Accueil-2	Accueil-3	Accueil-4	Accueil-5	Q/Prix-Non	Q/Prix-Oui
Satisfait-Non	13	0	1	5	7	4	3	4	1	1	12	1
Satisfait-Oui	0	15	6	0	9	2	1	3	6	3	6	9
Réparé-NSP	1	6	7	0	0	0	1	2	3	1	2	5
Réparé-Non	5	0	0	5	0	2	0	2	0	1	4	1
Réparé-Oui	7	9	0	0	16	4	3	3	4	2	12	4
Accueil-1	4	2	0	2	4	6	0	0	0	0	6	0
Accueil-2	3	1	1	0	3	0	4	0	0	0	3	1
Accueil-3	4	3	2	2	3	0	0	7	0	0	6	1
Accueil-4	1	6	3	0	4	0	0	0	7	0	2	5
Accueil-5	1	3	1	1	2	0	0	0	0	4	1	3
Q/Prix-Non	12	6	2	4	12	6	3	6	2	1	18	0
Q/Prix-Oui	1	9	5	1	4	0	1	1	5	3	0	10

Source :https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-des-correspondances-multiples-acm-ou-afcm

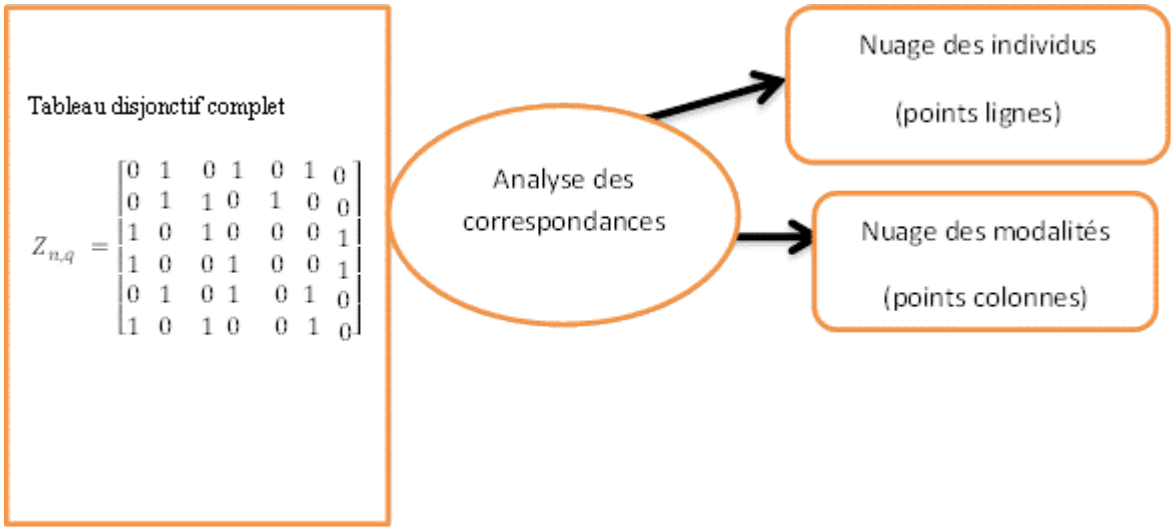
Le quatre sous tableaux de continence qui se trouvent dans la première diagonale (en jaune) fournissent les effectifs des modalités des quatre variables avec elles-mêmes (ce sont des effectifs absolus unisériés). Cependant, les six sous tableaux de contingence qui se trouvent au-dessus de la première diagonale sont identiques à ceux qui se trouvent au-dessous de la diagonale principale.

On note par *D* la matrice diagonale contenant les effectifs de la diagonale de tableau de Burt.

Table 4.9. Présentation de la matrice diagonale contenant les effectifs de la diagonale de tableau de Burt (D)

	Satisfait-Non	Satisfait-Oui	Réparé-NSP	Réparé-Non	Réparé-Oui	Accueil-1	Accueil-2	Accueil-3	Accueil-4	Accueil-5	Q/Prix-Non	Q/Prix-Oui
Satisfait-Non	13	0	0	0	0	0	0	0	0	0	0	0
Satisfait-Oui	0	15	0	0	0	0	0	0	0	0	0	0
Réparé-NSP	0	0	7	0	0	0	0	0	0	0	0	0
Réparé-Non	0	0	0	5	0	0	0	0	0	0	0	0
Réparé-Oui	0	0	0	0	16	0	0	0	0	0	0	0
Accueil-1	0	0	0	0	0	6	0	0	0	0	0	0
Accueil-2	0	0	0	0	0	0	4	0	0	0	0	0
Accueil-3	0	0	0	0	0	0	0	7	0	0	0	0
Accueil-4	0	0	0	0	0	0	0	0	7	0	0	0
Accueil-5	0	0	0	0	0	0	0	0	0	4	0	0
Q/Prix-Non	0	0	0	0	0	0	0	0	0	0	18	0
Q/Prix-Oui	0	0	0	0	0	0	0	0	0	0	0	10

L'AFCM est l'analyse des correspondances du tableau disjonctif complet. En effet, nous procédons de la même manière que l'AFC. Nous calculons les profils-lignes, les profils colonnes, ainsi que la métrique de Khi-deux (distance de khi-deux).



4.3.2. Coordonnées des points-lignes (individus) et points-colonnes (modalités)

Pour le cas du nombre de variables est supérieur à deux. Le calcul des cordonnées des individus et des modalités sont données par les formules suivantes :

Analyse dans $\mathbb{R}^p$:

Les coordonnées points-lignes sur l'axe α: $F_\alpha = D^{-1}U_\alpha$ tel que U_α le vecteur propre de la matrice $(\frac{1}{p}Z'ZD^{-1})$ associé à la valeur propre λ_α

Analyse dans $\mathbb{R}^n$:

En utilisant la formule de transition, les coordonnées points-colonnes sur l'axe α:

$$G_\alpha = \frac{1}{p\sqrt{\lambda_\alpha}} ZF_\alpha$$

Exemple d'application de l'AFCM : (cet exemple est tiré du help xl-stat. Il est disponible sur le site https://help.xlstat.com/fr/6374-analyse-des-correspondances-multiples-acm-dans-excel)

Le tableau de données est construit sur la base d'une enquête réalisée par un concessionnaire automobile auprès de 28 clients en leurs posant cinq questions :

- *Etes-vous globalement satisfait de votre visite au garage ? (Oui / Non)*
- *Considérez-vous que la réparation a bien été effectuée ? (Oui / Non / Ne sait pas)*
- *Comment jugez-vous la qualité de l'accueil ? (1 à 5)*
- *Le rapport qualité prix vous semble-t-il correct ? (Oui / Non)*
- *Reviendrez-vous dans ce garage pour une réparation ? (Oui / Non / Ne sait pas)*

Après avoir soumis le tableau de données à ne analyse de correspondances multiples, les principaux résultats sont présentés ci-dessous.

Valeurs propres :

	F1	F2	F3	F4	F5	F6	F7	F8
Valeur pro	0,611	0,321	0,278	0,260	0,215	0,165	0,090	0,058
Variabilité	30,561	16,060	13,918	13,009	10,761	8,261	4,515	2,914
% cumulé	30,561	46,622	60,540	73,549	84,311	92,572	97,086	100,000

Résultats pour les variables

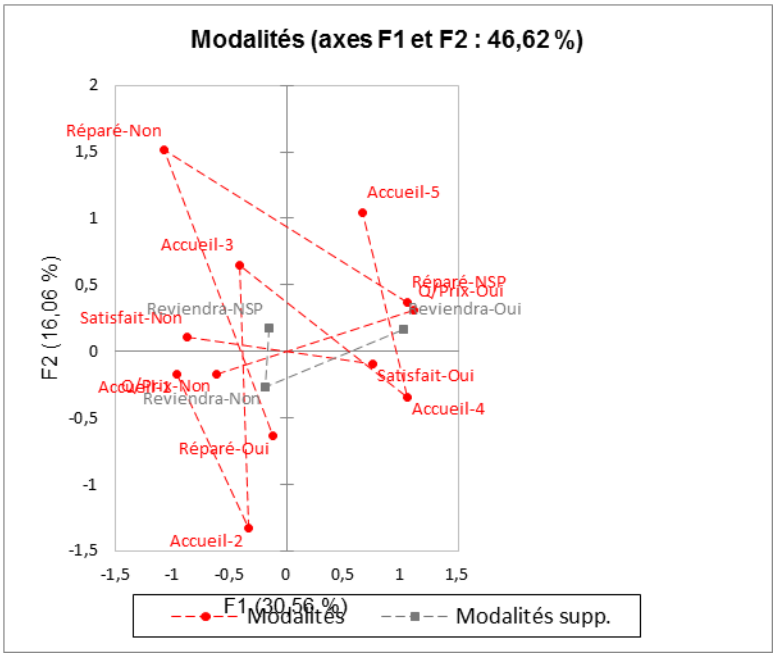
Le tableau des contributions montre que les modalités Satisfait-Non, Satisfait-Oui, Réparé-NS, accueil-4, Q/prix-Non, et Q/Prix- Oui contribuent beaucoup plus à la constitution de la première dimension, tandis que Réparé-Non, Réparé-Oui Acuei-2 et accueil-5 contribuent à la formation de la dimension 2.

Contributions (Variables) :

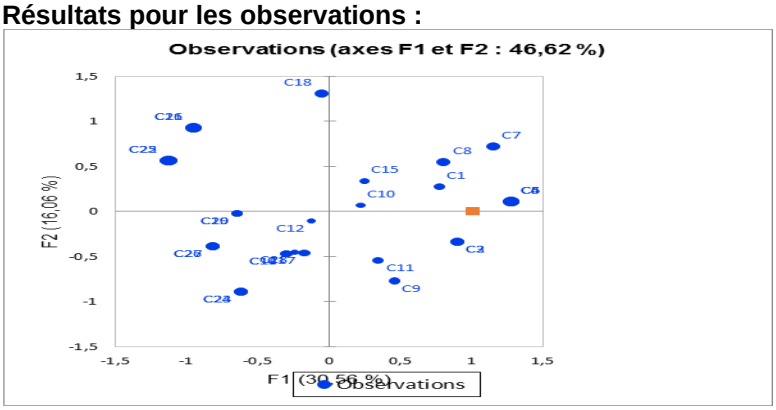
	F1	F2	F3	F4
Satisfait-Non	14,391	0,399	0,910	2,627
Satisfait-Oui	12,472	0,346	0,789	2,277
Réparé-NSP	11,352	2,688	14,273	0,001
Réparé-Non	8,554	31,798	0,585	0,129
Réparé-Oui	0,352	17,950	4,290	0,050
Accueil-1	8,132	0,498	20,013	11,206
Accueil-2	0,642	19,622	7,497	39,864
Accueil-3	1,805	8,240	29,377	3,629
Accueil-4	11,376	2,266	0,015	12,480
Accueil-5	2,590	12,164	21,087	24,794
Q/Prix-Non	10,119	1,439	0,416	1,051
Q/Prix-Oui	18,213	2,590	0,749	1,891

Valeurs test (Variables) :

	F1	F2	F3	F4
Satisfait-N	**-4,211**	0,508	0,715	-1,174
Satisfait-O	**4,211**	-0,508	-0,715	1,174
Réparé-NS	**3,161**	1,115	**2,392**	-0,022
Réparé-No	**-2,622**	**3,665**	-0,463	-0,210
Réparé-Ou	-0,737	**-3,812**	-1,735	0,182
Accueil-1	**-2,614**	-0,469	**-2,767**	**2,002**
Accueil-2	-0,703	**-2,818**	1,622	**-3,615**
Accueil-3	-1,260	1,952	**3,432**	1,166
Accueil-4	**3,164**	-1,024	0,078	**2,162**
Accueil-5	1,412	**2,219**	**-2,720**	**-2,851**
Q/Prix-Nor	**-4,325**	-1,182	0,592	0,909
Q/Prix-Oui	**4,325**	1,182	-0,592	-0,909
Reviendra-	-0,751	0,842	0,100	0,264
Reviendra-	-0,782	-1,117	-0,118	-0,561
Reviendra-	**2,161**	0,359	0,021	0,407



Résultats pour les observations :



Les deux graphiques ci-dessus représentent respectivement la carte factorielle des modalités et la carte factorielle des observations. Il est à noter que la graphique des modalités fait ressortir que les clients ont tendance à revenir auprès du concessionnaire une fois qu'ils sont satisfaits du service redu, du prix pratiqué, de l'accueil, et de la qualité du service lié à la réparation du véhicule prix.

Références ,

1. Caumont, D., Ivanaj, S. (2017). *Analyse de données*. DUNOD
2. Castell F. Cours d'analyse de données. Aix Marseille Université.
3. Husson, F., Le, S., Pagès J. (2016). *Analyse de données avec R*. PRESSES UNIVERSITAIRES DE RENNES (PUR)
4. Gilula Z., (1986). Grouping and association in contingency tables: an exploratory canonical correlation approach. J. of Amer. Statist. Assoc. 81, 773-779.
5. Grelet Y. (1993). Préparation des tableaux pour l'analyse de données : le codage des variables. In. Traitement statistique d'enquête, Grangé D., Lebart L. (eds.). DUNOD. Paris.
6. Lebart, L., Morineau A., Piron M. (1995). *Statistique exploratoire multidimensionnelle*. DUNOD
7. Joseph F Hair, Barry J. Babin, Rolph E. Anderson, William C. Black , (2018). *Multivariate Data Analysis, 8th edition Paperback*. CENGAGE
8. Saborta, G. (2011). *Probabilités, analyse de données et statistiques.* TECHNIP.

Tables des matières

Introduction	**1**
Chapitre 1. Quelques rappels de mathématiques et d'analyse bidimensionnelle	**1**
1.1. Produit scalaire dans un plan	**1**
1.2. Distance entre deux points	**1**
1.3. Calcul matriciel	**4**
1.3.1. Les valeurs propres d'une matrice	**4**
1.3.2. Vecteurs propres d'une matrice	**5**
1.4. Analyse bidimensionnelle	**7**
1.4.1. Relation entre deux variables quantitatives	**7**
1.4.2. Coefficient de corrélation linaire de Bravais-Pearson, noté $r_{X,Y}$	**8**
Exercice	**10**
Chapitre 2. Analyse factorielle générale	**12**
2.1. Analyse dans $\mathbb{R}^P$	**12**
2.2. Analyse dans $\mathbb{R}^n$	**17**
Chapitre 3. Analyse en composantes principales (ACP)	**20**
3.1. Caractéristiques relatives aux tableaux de données quantitatives multidimensionnelles	**20**
3.2. Principe de la méthode ACP	**23**
3.2.1. ACP non-normée	**25**
3.2. 2 . ACP normée	**29**
3.4. Application de la méthode « Analyse en Composantes Principales –ACP » sur SPSS	**33**
Exercices	**46**
Chapitre 4. Analyse factorielle des correspondances, et Analyse factorielle des correspondances multiples.	**61**
4.1. Analyse factorielle des correspondances (AFC)	**61**
4.2. Fondements théoriques de la méthode AFC	**63**
Exemple d'application de l'AFC	**66**
4.3. Analyse factorielle des correspondances multiples (AFCM)	**69**
4.3.1. Principes de l'AFCM	**70**
4.2. Coordonnées des points-lignes (individus) et points-colonnes (modalités)	**73**
Exemple d'application de l'AFCM	**74**
Références	**76**