République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

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

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

MANSOURI Hanane

Titre :

Analyse en Composantes Principales (ACP)

Membres du Comité d'Examen :

Dr.	**BENELMIR Imen**	UMKB	Encadreur
Dr.	**BENAMEUR Sana**	UMKB	Président
Dr.	**DHIABI Samra**	UMKB	Examinateur

Juin 2019

Dédicace

Au nom du Dieu clément et misécordieux

A mon cher père

Pour l'amour et l'éducation qu'il m'a donnée

A ma chère mère

Pour son grand amour, ses sacrifices et toute l'affection qu'elle m'a toujours offerte

A mes sœurs

Fouzia Chahinaz Hadjer Ahlam Narimane

A mon cher frère

Imade Eddine

En leurs souhaitant tout le succès...tout le bonheur

A mes chères amies

Amani Ilham Yousra

A tous les étudiants de mathématique, surtout $2^{éme}$ master groupe de **statistique**

et tous les étudiants de l'université **Mohammed Khieder**.

Mansouri Hanane.

Remerciements

Je tiens tout d'abord à remercier bien ***ALLAH*** le tout puissant et miséricordieux qui m'a donné la force et la patience d'accomplir ce modeste travail.

Je tiens à remercier avec ma plus grande gratitude mon encadreur : **Benelmir Imen** pour la suivi et l'aide qu'elle m'a apporté pour l'élaboration et pour ses précieux conseils et ses aides durant toute la période du travail.de ce mémoire.

Je tiens aussi remerciement à l'ensemble des enseignants de département de mathématique

Je remercie les membres du jury :

Benameur Sana et **Dhiabi Samra**

Enfin, je tiens également à remercier toutes les personnes qui ont participé de prés ou de loin à la réalisation de ce travail.

Merci à Tous.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'analyse des données est un sous domaine des statistiques qui se préoccupe de la description des données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différents données ainsi qu'à en tirer une information statistique qui sert à décrire les principales informations contenues dans ces dernières.

L'analyse en composantes principales est un grand classique de l'analyse des données qu'on note par la suite ACP ou principal component analysis (PCA) en anglai. C'est une méthode statistique multivariée proposée sous forme d'un tableau rectangulaire des données comportant les valeurs des variables quantitatives pour un ensemble individus qui sont utilisés pour résumer et visualiser l'information contenue dans ces données procurant ainsi des représentations géométriques de ces individus et de ces variables.

L'objectif de l'Analyse en Composantes Principales est de réduire la dimension d'un espace en essayant de déformer le moins possible les critères à la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

Le but de ce travail est de présenter et de faire une description de l'ACP, toutes expliquant comment résoudre le problème de la représentation des données étudier les relations existantes entre les individus par l'évaluation de leurs ressemblances, ainsi que les relations entre les variables par l'évaluation de leurs liaisons. Ce travail ce devise en deux chapitres :

chapitre1 : On va présenter quelques définitions, proposition, propriétés...ect. En d'autres termes, on va faire une description des données et leurs caractéristiques, les données traitées sont des individus et des variables quantitatives.

chapitre2 : On va traiter l'ACP en expliquant le principe de cette méthode avec ces éléments et ces caractéristiques. On a aussi essayé d'interprète les résultats de l'ACP.

On achève ce travail par une application faite est sur des données réelles "La température mensuelle de 15 villes de France sur 30 ans" sous logiciel **R.**

Chapitre 1

Préliminaires

L'analyse des données (aussi appelée analyse exploratoire des données ou AED) est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives permettant de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Dans ce chapitre, on s'intéresse d'abord à la description de ces données ainsi qu'à leurs caractéristiques comme le tableau des données, puis on définit les individus, les variables, la matrice des poids, le centre de gravité...ect.

1.1 Données et leurs caractéristiques

Avant tout travail, on doit procéder au préliminaires, dont le tableau des données les individus les variables et autres.

1.1.1 Tableau des données

Chaque tableau contient des lignes qui représentent les individus et des colonnes qui représentent les variables. Ce tableau rectangulaire (matrice) qu'on note par X possède des

observations à n individus et p variables. Il a la forme suivante :

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ . & & . \\ . & x_{ij} & . \\ . & & . \\ x_{n1} & \dots & x_{np} \end{bmatrix} \in \mathcal{M}_{\mathbb{R}}(n,p),$$

où x_{ij} est la valeur prise par la variable j sur l'individu i.

1.1.2 Individus et variables

Les individus et les variables sont définit ci-dessous.

Définition 1.1.1 (Individu) *Le $i^{éme}$ individu est un vecteur à p composantes réelles qu'on le note par e_i tel que*

$$e_i = (x_{i1}, x_{i2}, ..., x_{ip})^t \in \mathbb{R}^p, \text{ pour } i = \overline{1,n}.$$

Définition 1.1.2 (Variable) *La $j^{éme}$ variable est la liste des n valeurs qui elle prend sur n individus, on la note par x_j tel que*

$$x_j = (x_{1j}, x_{2j}, ..., x_{nj})^t \in \mathbb{R}^n, \text{ pour } j = \overline{1,p}.$$

1.1.3 Types de variables

Il existe deux types des variables : les variables quantitatives (ce qui est dans notre cas) et les variables qualitatives.

Définition 1.1.3 (Variable quantitative) *En statistique, une variable quantitative est une variable qui reflète une notion de grandeur, c'est-à-dire (i.e) si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesure qui sert de référence.*

Définition 1.1.4 (Variable qualitative) *En statistique, une variable qualitative est une variable catégorielle (facteur) qui prend pour valeur des modalités (catégories, niveaux), par opposition aux variables quantitatives qui mesurent sur chaque individu une quantité.*

Exemple 1.1.1 *Les observations suivantes représente les mesurés quotidiennes de trois variables indicatrices : la taille, le poids et l'âge de cinq personnes dans une certaine ville. Les résultats sont représentés dans le tableau X avec $n = 5$ et $p = 3$*

$$X = \begin{bmatrix} 1.69 & 77.85 & 22 \\ 1.53 & 55.10 & 21 \\ 1.62 & 76.55 & 19 \\ 1.53 & 62.69 & 25 \\ 1.68 & 58.00 & 21 \end{bmatrix} \in \mathcal{M}_{\mathbb{R}}(5,3),$$

où x_1 représente la taille, x_2 le poids et x_3 l'âge.

Par exemple pour le quatrième individu et la troisième variable on a :

$e_4 = (1.53, 62.69, 25)^t \in \mathbb{R}^3$.

$x_3 = (22, 21, 19, 25, 21)^t \in \mathbb{R}^5$.

1.1.4 Matrice des poids

Si les données ont été recueillies d'un tirage aléatoire, alors les probabilités de ces n individus ont toutes la même importance i.e égale à $\frac{1}{n}$, or ceci n'est pas toujours le cas. Dans le cas contraire, il est utile de travailler avec des poids qu'on note par p_i pour les différents individus où ces derniers sont regroupés dans une matrice diagonale de taille n notée D appelée matrice des poids. Elle est définit comme suit :

$$D = \begin{bmatrix} p_1 & \dots & 0 \\ . & & . \\ . & & . \\ 0 & \dots & p_n \end{bmatrix}, \text{ avec } p_i \geq 0 \text{ et } \sum_{i=1}^{n} p_i = 1.$$

Dans le cas usuel des poids égaux, on a

$$D = \frac{1}{n} I_n.$$

Preuve.

Comme on a $p_1 = p_2 = ...p_i = ... = p_n$ et $\sum_{i=1}^{n} p_i = 1$ alors

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

Par conséquent

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

Et

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

1.1.5 Centre de gravité

C'est le vecteur des moyennes arithmétiques de chaque variable, on le note par g qu'on appelle aussi individu moyen ou point moyen. Il est définit par :

$$g = (\overline{x_1}, \overline{x_2}, ..., \overline{x_p})^t \in \mathbb{R}^p,$$

où $\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$.

La forme matricielle :

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

Preuve.

$$X^t D 1_n = \begin{bmatrix} x_{11} & \dots & x_{n1} \\ . & & . \\ . & . & . \\ x_{1p} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} p_1 & \dots & 0 \\ . & & . \\ . & & . \\ 0 & \dots & p_n \end{bmatrix} \begin{bmatrix} 1 \\ . \\ . \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} p_i x_{i1} \\ . \\ . \\ \sum_{i=1}^{n} p_i x_{ip} \end{bmatrix} = \begin{bmatrix} \overline{x_1} \\ . \\ . \\ \overline{x_p} \end{bmatrix} = g.$$

■

1.1.6 Standardisation du tableau

Dans l'analyse en composantes principales les variables sont souvent normalisées. Ceci est particulièrement recommandé lorsque les variables sont mesurées dans différentes unités par exemple : (kilogrammes, kilomètres, centimètres, ...ect) ; sinon, le résultat de l'analyse obtenue sera fortement affecté.

L'objectif est de rendre les variables comparables. Généralement, les variables sont normalisées de manière à ce qu'elles aient au final

1. Un écart type égale à un.
2. Une moyenne égale à zéro.

Tableau centré associé à X

Le centrage des données nous permet de ramener toutes les colonnes de X à la même origine zéro dans une matrice notée par Y de terme général :

$$y_{ij} = x_{ij} - \overline{x_j}.$$

La forme matricielle :

$$Y = X - 1_n g^t.$$

Preuve.

$$X - 1_n g^t = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ . & & . \\ . & & . \\ x_{n1} & \dots & x_{np} \end{bmatrix} - \begin{bmatrix} 1 \\ . \\ . \\ 1 \end{bmatrix} (\overline{x_1}, \overline{x_2}, ..., \overline{x_p})$$

$$= \begin{bmatrix} x_{11} - \overline{x_1} & \dots & x_{1p} - \overline{x_p} \\ . & & . \\ . & & . \\ x_{n1} - \overline{x_1} & \dots & x_{np} - \overline{x_p} \end{bmatrix} = \begin{bmatrix} y_{11} & \dots & y_{1p} \\ . & & . \\ . & & . \\ y_{n1} & \dots & y_{np} \end{bmatrix} = Y.$$

■

Tableau réduit associé à X

La réduction des données nous permet de ramener toutes les variables à un même écart-type 1, d'après le tableau Y on construit un tableau standard noté par Z de terme général :

$$Z = \frac{y_{ij}}{s_j}.$$

Avec : $s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$.

La forme matricielle :

$$Z = Y D_{1/s}.$$

Avec :

$$D_{1/s} = \begin{bmatrix} 1/s_1 & \dots & 0 \\ . & & . \\ . & & . \\ 0 & \dots & 1/s_p \end{bmatrix}.$$

Preuve.

$$YD_{1/s} = \begin{bmatrix} y_{11} & \dots & y_{1p} \\ . & & . \\ . & & . \\ y_{n1} & \dots & y_{np} \end{bmatrix} \begin{bmatrix} 1/s_1 & \dots & 0 \\ . & & . \\ . & & . \\ 0 & \dots & 1/s_p \end{bmatrix}$$

$$= \begin{bmatrix} y_{11}/s_1 & \dots & y_{1p}/s_p \\ . & & . \\ . & & . \\ y_{n1}/s_1 & \dots & y_{np}/s_p \end{bmatrix} = \begin{bmatrix} z_{11} & \dots & z_{1p} \\ . & & . \\ . & & . \\ z_{n1} & \dots & z_{np} \end{bmatrix} = Z.$$

■

1.1.7 Matrice de variance-covariance

C'est l'ensemble des variances et des covariances, regroupées dans un tableau noté S de terme général :

$$S_{jj'} = cov(x_j, x_{j'}) = \sum_{i=1}^{n} p_i(x_{ij} - \overline{x_j})(x_{ij'} - \overline{x_{j'}}), \text{ pour } j, j' = \overline{1, p}.$$

La matrice de variance-covariance est donnée par

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

La forme matricielle :

$$S = Y^t DY = X^t DX - gg^t.$$

Dans le cas ou les poids sont égaux, la forme matricielle devient :

$$S = \frac{1}{n} Y^t Y = \frac{1}{n} X^t X - gg^t.$$

Preuve.

On a

$$\begin{aligned} Y &= X - 1_n g^t, \text{ alors} \\ S &= (X - 1_n g^t)^t D (X - 1_n g^t) \\ &= X^t DX - X^t D 1_n g^t - g 1_n^t DX + g 1_n^t D 1_n g^t \\ &= X^t DX - gg^t - gg^t + gg^t, \text{ car } 1_n^t D 1_n = \sum_{i=1}^{n} p_i = 1 \\ &= X^t DX - gg^t. \end{aligned}$$

■

1.1.8 Matrice de corrélation

C'est l'ensemble des coefficients de corrélation, regroupés dans un tableau noté par R dont les termes diagonaux valent 1. Chaque élément $r_{jj'}$ est défini par :

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}.$$

La matrice de corrélation est donnée par

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

La forme matricielle :

$$R = D_{1/s} S D_{1/s} = Z^t DZ.$$

Preuve.

On montre d'abord que $R = D_{1/s} S D_{1/s}$ On a

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

Ensuite, on montre que $Z^t DZ = R$. On a

$$
\begin{aligned}
Z^t DZ &= \left(YD_{1/s}\right)^t D \left(YD_{1/s}\right) \\
&= D_{1/s}Y^t DYD_{1/s} \\
&= D_{1/s}SD_{1/s} \\
&= R.
\end{aligned}
$$

■

Remarque 1.1.1

- *R et S sont des matrices carrées symétriques d'ordre p. Comme il ya p variables cela nous conduit donc à calculer $\dfrac{p\,(p-1)}{2}$ corrélations.*

1.2 Nuage de points (individus)

Chaque individu étant un point défini par p coordonnées est considéré comme un vecteur d'un espace vectoriel défini dans $\mathbb{R}^p$ appelé l'espace des individus. L'ensemble des n individus est un nuage de points appelé nuage des individus.

1.2.1 Ressemblance entre deux individus

Deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables. On exprime la ressemblance par la distance qui est égale à :

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

1.2.2 Métrique

En physique, la distance entre deux points dans l'espace se calcule facilement par la formule de Pythagore : le carré de la distance est la somme des carrées des différences des coordonnées, car les dimensions sont de même nature (unité). Mais en statistique il n'en est pas de même, car chaque dimension correspond à un caractère qui s'exprime avec sa propre unité.
On particulier, pour résoudre ce problème on définie la distance entre deux individus e_i et $e_{i'}$ sous la forme quadratique suivante :

$$\langle e_i, e_{i'} \rangle_M = (e_i - e_{i'})^t M (e_i - e_{i'}),$$

où M est une matrice carrée symétrique d'ordre p définie positive.
La formule de Pythagore revient à définir le produit scalaire de deux individus e_i et $e_{i'}$ de la façon suivante :

$$\langle e_i, e_{i'} \rangle_M = e_i^t M e_{i'}.$$

Les métriques les plus utilisées sont les métriques diagonale qui sont I_p et D_{1/s^2}.
I_p : représente la matrice identité d'ordre p, et

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

Ce qui revient à diviser chaque caractère par son écart-type. Ceci a pour avantage que la distance entre deux individus ne dépend plus des unités de mesure ce qui est très utile lorsque les variables ne s'expriment pas avec les mêmes unités.

Remarque 1.2.1

- *On utilise la métrique D_{1/s^2} pour le tableau Y et la métrique I_p pour le tableau Z.*

Preuve.

On a

- Le $i^{ème}$ individu du tableau Y est $e_i^y = (y_{i1}, ..., y_{ip})^t \in \mathbb{R}^p$.
- Le $i^{ème}$ individu du tableau Z est $e_i^z = (z_{i1}, ..., z_{ip})^t \in \mathbb{R}^p$.

$$\begin{aligned}
\langle e_i^y, e_i^y \rangle_{D_{1/s^2}} &= (e_i^y)^t D_{1/s^2} e_i^y \\
&= \left(\frac{y_{i1}}{s_1}\right)^2 + ... + \left(\frac{y_{ip}}{s_p}\right)^2 \\
&= \sum_{j=1}^{p} \left(\frac{y_{ij}}{s_j}\right)^2 \\
&= \sum_{j=1}^{p} (z_{ij})^2 \\
&= \sum_{j=1}^{p} \left(\frac{z_{ij}}{1}\right)^2 \\
&= (e_i^z)^t I_p e_i^z \\
&= \langle e_i^z, e_i^z \rangle_{I_p}.
\end{aligned}$$

■

1.2.3 Inertie

On appelle inertie totale du nuage de points, la moyenne des carrées des distances des n points au centre de gravité g. Elle est exprimée comme ceci :

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

On peut aussi l'écrire comme :

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

Remarque 1.2.2

1. *L'inertie en un point quelconque est définie par :*

$$I_a = \sum_{i=1}^{n} p_i d_M^2(e_i, a).$$

2. Si $g = 0$, on a

$$I_g = \sum_{i=1}^{n} p_i \|e_i\|_M^2 = \sum_{i=1}^{n} p_i e_i^t M e_i.$$

3. Formule de huyghens :

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

4.

$$I_g = tr\left(MS\right) = tr\left(SM\right).$$

Démonstration de la $3^{ième}$ remarque. Puisque $I_a = \sum_{i=1}^{n} p_i \langle e_i - a, e_i - a\rangle_M$ alors

$$\begin{aligned}
\langle e_i - a, e_i - a\rangle_M &= \langle e_i - g + g - a, e_i - g + g - a\rangle_M \\
&= \langle e_i - g, e_i - g\rangle_M + \langle e_i - g, g - a\rangle_M + \langle g - a, e_i - g\rangle_M + \langle g - a, g - a\rangle_M \\
&= \|e_i - g\|_M^2 + 2\langle e_i - g, g - a\rangle_M + \|g - a\|_M^2.
\end{aligned}$$

D'où

$$\begin{aligned}
I_a &= \sum_{i=1}^{n} p_i \left(\|e_i - g\|_M^2 + 2\langle g - a, e_i - g\rangle_M + \|g - a\|_M^2\right) \\
&= I_g + 2\sum_{i=1}^{n} p_i \langle g - a, e_i - g\rangle_M + \|g - a\|_M^2.
\end{aligned}$$

Il reste à montrer que $\sum_{i=1}^{n} p_i \langle g - a, e_i - g\rangle_M = 0$. En effet

$$\begin{aligned}
\sum_{i=1}^{n} p_i \langle g - a, e_i - g\rangle_M &= \sum_{i=1}^{n} p_i (g - a)^t M (e_i - g\rangle \\
&= (g - a)^t M \sum_{i=1}^{n} p_i (e_i - g) \\
&= (g - a)^t M (\sum_{i=1}^{n} p_i e_i - \sum_{i=1}^{n} p_i g) \\
&= (g - a)^t M (g - g), \text{ car } g = \sum_{i=1}^{n} p_i e_i \\
&= 0.
\end{aligned}$$

■

Démonstration de la $4^{ième}$ remarque. On a

$$\begin{aligned}
I_g &= \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) \\
&= tr\left(\sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g)\right) \\
&= \sum_{i=1}^{n} tr\left(p_i (e_i - g)^t M (e_i - g)\right)
\end{aligned}$$

$$= \sum_{i=1}^{n} tr(p_i M(e_i - g)(e_i - g)^t), \text{ car } tr\,(AB) = tr\,(BA)$$
$$= tr\left(M\left(\sum_{i=1}^{n} p_i (e_i - g)(e_i - g)^t\right)\right)$$
$$= tr\,(MS)\,.$$

■

Proposition 1.2.1

1. *Si $M = I_p$, l'inertie est égale à la somme des variances des p variables :*

$$I_g = \sum_{j=1}^{p} S_j^2.$$

2. *Si $M = D_{1/S^2}$, l'inertie est égale au nombre de variables :*

$$I_g = p.$$

1.3 Nuage de points (variables)

Chaque variable est associée à une suite de n nombres, elle peut être représentée comme un vecteur d'un espace défini dans $\mathbb{R}^n$ appelé espace des variables. L'ensemble des p variables constitue un nuage de points appelé nuage des variables.

1.3.1 Liaison entre deux variables

Le coefficient $r_{jj'}$ de corrélation mesure la liaison entre deux variables x_j et $x_{j'}$, qui prend ses valeurs dans $[-1, 1]$

$$r\,(x_j, x_{j'}) = \frac{cov\,(x_j, x_{j'})}{\sqrt{var\,(x_j)\,var\,(xj')}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \overline{x_j}}{s_j}\right)\left(\frac{x_{ij'} - \overline{x_{j'}}}{s_{j'}}\right), \text{ pour } j, j' = \overline{1, p}.$$

Avec :

$cov\,(x_j, x_{j'}) = (x_{ij} - \overline{x_j})\left(x_{ij'} - \overline{x_{j'}}\right)$, $\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ et $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2$.

1.3.2 Métrique des variables

Pour étudier la proximité des caractères entre eux, il faut munir cet espace d'une métrique, i.e trouver une matrice symétrique d'ordre n définie positive. Ici il n'y a pas d'hésitation comme pour l'espace des individus et le choix se porte sur la matrice diagonale des poids D pour les raisons suivantes :

1. Le produit scalaire des variables x_j et $x_{j'}$ qui est définie comme suit :

$$\langle x_j, x_{j'} \rangle_D = x_j^t D x_{j'} = \sum_{i=1}^{n} p_i x_{ij} x_{ij'}, \text{ pour } j, j^{'} = \overline{1, p}.$$

   n'est autre que la matrice de covariance $S_{jj'}$, car les caractères sont centrés.

2. La norme d'un caractère x_j est alors :

$$\| x_j \|_D^2 = S_j^2.$$

3. Dans un espace euclidien on définit l'angle $\theta_{jj'}$ entre deux vecteurs par son *cosinus* qui est égal au quotient du produit scalaire par le produit des normes des deux vecteurs :

$$\cos\theta_{jj'} = \frac{\langle x_j, x_{j'} \rangle_D}{\| x_j \|_D \| x_{j'} \|_D} = \frac{S_{jj'}}{S_j S_{j'}} = r\left(x_j, x_{j'}\right).$$

Remarque 1.3.1

1. *Dans l'espace des individus on s'intéresse aux distances entre points par contre, dans l'espace des variables on s'intéresse aux l'angle entre vecteurs.*
2. *Pour les données du tableaux standard Z le nuage des variables se trouve sur une hyper sphère de rayon égale à 1 car $\| z_j \|_D = s_{jj'} = 1$, pour $j = \overline{1, p}$.*

Chapitre 2

Analyse en composantes principales

L'analyse en composantes principales notée ACP est une méthode d'analyse statistique multivariée, qui a pour but d'étudier simultanément un nombre important de variables quantitatives. L'ACP permet d'obtenir des représentations graphique des distances entre les individus et des corrélations entre les variables.

2.1 Principe de l'ACP

On cherche une représentation des n individus $(e_1, e_2, ..., e_n)$ dans un espace F_k de $\mathbb{R}^p$ tel que k soit le plus petit possible $(k \prec p)$ i.e, on cherche à définir k nouvelles variables dites combinaison linéaire des p variables initiales contenant le plus d'informations possible.

2.1.1 Projection des individus

Dans cette partie on va parler sur la construction de sous-espace F_k à savoir le nuage de projection et les droites appelées aussi axes.

Nuage projeté

Le critère du choix de l'espace de projection s'effectue tel que la moyenne des carrées des distances entre les projections et leur centre de gravité soit la plus grande possible. Ce qui implique qu'il faut que l'inertie du nuage projeté sur ce sous espace soit maximale.

On note F_k le sous espace de projection. Pour cela on définit P une matrice (operateur) de projection M-orthogonal sur l'espace F_k, elle vérifie les deux conditions suivantes :

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

Définition 2.1.1 *Soit f_i la projection d'un individu e_i tel que $f_i = Pe_i$ d'où $f_i^t = e_i^tP^t$ c'est la $i^{éme}$ ligne du tableau XP^t.*
On écrit

$$X_{proj} = XP^t. \tag{1}$$

Proposition 2.1.1

1. *La matrice de covariance associée au nuage projeté :*

$$S_{proj} = PSP^t. \tag{2}$$

2. *L'inertie du nuage projeté :*

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

3. *Le centre de gravité projeté :*

$$g_{proj} = Pg.$$

Preuve.

1. Matrice de covariance :

$$\begin{aligned} S_{proj} &= X_{proj}^t DX_{proj} - g_{proj}g_{proj}^t \\ &= PX^tDXP^t - Pgg^tP^t, \text{ de (1)} \\ &= P\left(X^tDX - gg^t\right)P^t \\ &= PSP^t. \end{aligned}$$

2. Inertie :

$$I_{proj} = tr\left(S_{proj}M\right)$$

$$= tr\left(PSP^tM\right), \text{ de (2)}$$

$$= tr\left(PSMP\right), \text{ car } P \text{ est } M-\text{symétrique}$$

$$= tr\left(SMP^2\right)$$

$$= tr\left(SMP\right), \text{ car } P \text{ est idempotente.}$$

3. Centre de gravité :

$$g_{proj} = X_{proj}^t D1_n$$

$$= \left(XP^t\right)^t D1_n, \text{ de (1)}$$

$$= P\left(X^t D1_n\right)$$

$$= Pg.$$

■

Construction de sous-espace F_k

La détermination du sous espace de projection F_k revient a trouver la matrice de projection P $M-$orthogonale de rang k qui maximise $tr(SMP)$.

Le sous espace F_k peut être construit de proche en proche en cherchant d'abords le sous espace Δ_1 de dimension 1 d'inertie maximal puis le sous espace Δ_2 de dimension 1 $M-$orthogonale à Δ_1 et d'inertie maximal, ...ect. La somme directe de ces sous espaces de dimension 1 est F_k tel que

$$F_k = \Delta_1 \oplus \Delta_2 \oplus ... \oplus \Delta_k.$$

On peut alors dire que

$$I_{F_k} = I_{\Delta_1} \oplus I_{\Delta_2} \oplus ... \oplus I_{\Delta_k}.$$

Construction de la première droite Δ_1

On cherche dans $\mathbb{R}^p$ la droite Δ_1 de dimension 1 qui passe par le centre de gravite g et qui maximise l'inertie de nuage projeté sur cette droite.

Soit $a_1 \in \mathbb{R}^p$ un vecteur directeur de Δ_1. L'opérateur de projection M-orthogonale sur Δ_1

est

$$P_1 = a_1 \left(a_1^t M a_1\right)^{-1} a_1^t M = \frac{a_1 a_1^t M}{a_1^t M a_1}, \text{ car } \left(a_1^t M a_1\right) \in \mathbb{R}.$$

En remplaçant le projecteur P_1 par sa formule dans la définition de l'inertie totale du nuage projeté, on obtient :

$$\begin{aligned} I_{\Delta_1} &= tr\left(SMP_1\right) \\ &= tr\left(SMa_1\left(a_1 a_1^t M / a_1^t M a_1\right)\right), \text{ car } P = a_1 a_1^t M / a_1^t M a_1 \\ &= tr\left(SMa_1 a_1^t M\right) / a_1^t M a_1 \\ &= tr\left(a_1^t MSMa_1\right) / a_1^t M a_1 \\ &= a_1^t MSMa_1 / a_1^t M a_1. \end{aligned}$$

L'inertie du nuage projeté sur Δ_1 est

$$I_{\Delta_1} = tr\left(SMP_1\right) = \frac{a_1^t MSMa_1}{a_1^t M a_1}.$$

On pose $\frac{a_1^t MSMa_1}{a_1^t M a_1} = f\left(a_1\right)$, où f est une fonction (forme quadratique) définie sur $\mathbb{R}^p$. Elle atteint son maximum en la dérivant par rapport à a_1, puis en résolvant cette dernière en l'annulant.

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

$$SMa_1 = \frac{a_1^t MSMa_1}{a_1^t M a_1} a_1.$$

On pose $\frac{a_1^t MSMa_1}{a_1^t M a_1} = \lambda \in \mathbb{R}$, alors

$$SMa_1 = \lambda a_1.$$

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

Proposition 2.1.2

- *La meilleure droite Δ_1 est engendré par les k vecteurs propres de la matrice SM associée aux k plus grandes valeurs propres.*

Remarque 2.1.1

1. *Comme la matrice SM est $M-$symétrique alors ces vecteurs propres sont deux à deux $M-$orthogonaux, ce qui implique que les droites $\Delta_1, \Delta_2, ..., \Delta_K$ sont deux à deux $M-$orthogonaux.*
2. *Le premier axe est celui qui aura la plus grande valeur propre λ_1. Le deuxième axe sera celui de la deuxième valeur propre λ_2 et ainsi de suite.*

2.2 Eléments de l'ACP et ces propriétés

L'ACP repose essentiellement sur trois éléments qui sont :

2.2.1 Axes principaux

Ce sont les p vecteurs propres $a_1, ..., a_p$ de la matrice SM associée à la valeur propre λ_j, $M-$normé à 1 i.e :

$$\begin{cases} SMa_j = \lambda_j a_j. \\ \|a_j\|_M^2 = 1. \end{cases} \tag{3}$$

Propriétés des axes principaux

1. Les axes principaux a_j sont S^{-1} orthogonaux.
2. Les axes principaux a_j sont M orthonormé.

Preuve.

1. Soit $a_j, a_{j'}$ deux axes principaux tel que

$$\begin{aligned}
\langle a_j, a_{j'} \rangle_{S^{-1}} &= a_j^t S^{-1} a_{j'} \\
&= 1/\lambda_j \left(SMa_j\right)^t S^{-1} a_{j'} \\
&= 1/\lambda_j a_j^t MSS^{-1} a_{j'} \\
&= 1/\lambda_j a_j^t M a_{j'} \\
&= 1/\lambda_j \langle a_j, a_{j'} \rangle_M \\
&= \begin{cases} 1/\lambda_j & \text{si} \qquad j = j'. \\ 0 & \text{si non} \end{cases}
\end{aligned}$$

■

2.2.2 Facteurs principaux

Soit a_j un axe principal, le facteur principal noté u_j est un vecteur propre de la matrice MS associé à la valeurs propre λ_j, M^{-1}-normé à 1 i.e :

$$\begin{cases} MSu_j = \lambda_j u_j. \\ \|u_j\|^2_{M^{-1}} = 1. \end{cases} \tag{4}$$

où $u_j = Ma_j \in \mathbb{R}^p$.

Propriétés des facteurs principaux

1. u_j sont $S-$orthogonaux.
2. u_j sont $M^{-1}-$orthonormé.
3. u_j sont aussi les vecteurs propres de la matrice MS.

Preuve.

1. $$\begin{aligned}
\langle u_j, u_{j'} \rangle_S &= u_j^t S u_{j'} \\
&= a_j^t MSMa_{j'} \\
&= a_j^t M\lambda_{j'} a_{j'} \\
&= \lambda_{j'} a_j^t M a_{j'} \\
&= \lambda_{j'} \langle a_j, a_{j'} \rangle_M
\end{aligned}$$

$$= \begin{cases} \lambda_{j'} & \text{si} \qquad j = j'. \\ 0 & \text{si non} \end{cases}$$

2. $\langle u_j, u_{j'} \rangle_{M^{-1}} = u_j^t M^{-1} u_{j'}$

$$= a_j^t M M^{-1} M a_{j'}$$

$$= a_j^t M a_{j'}$$

$$= \langle a_{j'}, a_{j'} \rangle_M$$

$$= \begin{cases} 1 & \text{si} \qquad j = j'. \\ 0 & \text{si non} \end{cases}$$

3. Comme a_j est un vecteur propre de la matrice SM, de (3) on a

$$SMa_j = \lambda_j a_j$$

$$MSMa_j = \lambda_j M a_j$$

$$MSu_j = \lambda_j u_j.$$

■

2.2.3 Composantes principales

Chaque axe a_j est associé à une variable appelée composante principale. Ce sont de nouvelles variables $c_j = (c_{1j}, c_{2j}, ..., c_{nj}) \in \mathbb{R}^n$, définies en fonction des facteurs principaux i.e :

$$c_j = XMa_j = Xu_j. \tag{5}$$

Si on travailler avec le tableau centré réduit deviennent :

$$c_j = Zu_j.$$

Chaque c_j contient des coordonnés $(c_1, c_2, ..., c_n)$ qui sont les mesures algébriques des projections des individus e_i sur ces axes.

Propriétés des composantes principales

1. Les composantes principales sont non corrélées deux à deux, car les axes associés sont orthogonaux i.e :

$$cov\,(c_j, c_{j'}) = 0.$$

2. La variance d'une composante principale c_j est égale à l'inertie apportée par l'axe principal dont il est associé i.e :

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

3. Les composantes principales sont les vecteurs propres de la matrice XMX^tD i.e :

$$XMX^tDc_j = \lambda_j c_j.$$

Preuve.

1. $cov\,(c_j, c_{j'}) = c_j^t D c_{j'} - g_{c_j} g_{c_{j'}}^t$

$= u_j^t X^t D X u_{j'} - c_j^t D 1_n 1_n^t D c_{j'}$, de (5)

$= u_j^t X^t D X u_{j'} - u_j^t X^t D 1_n 1_n^t D X u_{j'}$

$= u_j^t X^t D X u_{j'} - u_j^t X^t D 1_n 1_n^t D X u_{j'}$

$= u_j^t \left(X^t D X - g g^t\right) u_{j'}$

$= \langle u_j, u_{j'} \rangle_S$

$= 0.$

2. Même démonstration que la précédente

$var\,(c_j) = c_j^t D c_j - g_{c_j} g_{c_j}^t$

$= \langle u_j, u_j \rangle_S$

$= \| u_j \|_S^2$

$= \lambda_j.$

3. Dans le cas où g est centré.

$XMX^tDc_j = XMX^tDXu_j$

$= XMSu_j$

$$= X\lambda_j u_j, \text{ de (4)}$$
$$= \lambda_j X u_j$$
$$= \lambda_j c_j.$$

■

Remarque 2.2.1

1. *Les composantes principales c_j sont des combinaisons linéaires des variables centrées et réduites. On a :*
$$c_j = \sum_{k=1}^{p} u_{kj} x_k.$$
2. *La variance d'une composante principale est égale à l'inertie portée par l'axe principal qui lui est associé.*
3. *La première composante principale doit être de variance maximale.*

2.3 ACP sur les données centrées réduites

L'ACP, toujours centrée est souvent également réduite ; on parle alors d'ACP centrée réduite ou ACP normée. Cela revient à travailler sur la matrice Z pour accorder la même importance à chaque variable. C'est préférable si les variances associées à chaque variable sont trop différentes et c'est impératif si les unités de mesure sont différentes, c'est pourquoi on utilise la métrique triviale I_p. Dans ce cas la matrice de covariance est égale a la matrice de corrélation, il n'ya pas de distinction entre les facteurs principaux ou les axes principaux i.e que :

$$u_j = Ma_j = I_p a_j = a_j,$$

qui sont les vecteurs propres de la matrice de corrélation R associées aux valeurs propres de la même matrice R où ces valeurs propres sont d'ordre décroissant i.e :

$$Ru_j = \lambda_j u_j,$$

avec $\lambda_1 \succ \lambda_2 \succ ... \succ \lambda_p$.

2.4 Interprétation des résultats de l'ACP

Le rôle de l'ACP est de construire de nouvelles variables dites artificielles et de les représenter graphiquement afin de permettre à visualiser les relations qui existent entre elles et de décrire l'existence d'éventuels groupes d'individus et de groupes de variables.

2.4.1 Interprétation des individus

On va essayer d'interpréter les résultats pour les individus :

Qualité de représentation du nuage des individus sur F_k

C'est le pourcentage d'inertie d'information sur chaque axe, s'il existe l'indépendance entre les variables. Ce pourcentage nous permet de déterminer le nombre d'axes retenus on calcul. Elle est définie comme suit

$$QLT\left(F_k\right) = \frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{I_g},$$

avec $0 \leq QLT\left(F_k\right) \leq 1$.

Plus $QLT\left(F_k\right)$ est proche de 1 plus la représentation sur F_k est bonne.

Qualité de représentation d'un individu i par rapport à l'axe l (Δ_l)

On mesure la qualité de la projection d'un individu i sur Δ_l par le carré du *cosinus* de l'angle θ_{il} formé entre le vecteur z_i et l'axe l :

$$\begin{aligned} QLT_l\left(e_i\right) &= \frac{\text{inertie de la projection de l'individu } i \text{ sur l'axe } l}{\text{inertie initiale de l'individu } i} \\ &= \cos^2(\theta_{il}), \end{aligned}$$

avec

$$\cos^2(\theta_{il}) = \frac{c_{il}^2}{\parallel z_i \parallel^2}.$$

En général, on mesure la qualité de la projection d'un individu i sur deux axes l et l' par le carré du *cosinus* de l'angle $\theta_{i(l;l')}$ entre le vecteur z_i et sa projection orthogonale sur (l, l') :

$$QLT_{l,l'}\,(e_i\,) = \cos^2(\theta_{i(l,l')}),$$

avec :

$$\cos^2(\theta_{i(l,l')}) = \frac{c_{il}^2 + c_{il'}^2}{\parallel z_i \parallel^2}.$$

On peut donc dire que : $QLT_{l,l'}\,(i) = QLT_l\,(i) + QLT_{l'}\,(i)\,.$

Plus la valeur du $\cos^2$ est proche de 1, plus la représentation graphique de l'individu est de meilleure qualité.

Contribution d'un individu i par rapport à l'axe l

La contribution de l'individu i à la composante c_l est définie par :

$$CTR_l\,(e_i\,) = \frac{p_i c_{il}^2}{\lambda_l},$$

avec :

- $\lambda_l = \sum_{i=1}^{n} p_i c_{il}^2$.
- c_{il} : valeur de la composante c_l pour le $i^{éme}$ individu.

Remarque 2.4.1

1. *La contribution d'un individu e_i est importante si : $CTR_l\,(e_i) \succ p_i$.*
2. *Si on a un groupe d'individus, la contribution est égale à la somme des contributions des individus i et $i^{'}$. Alors*

$$CTR_l\,(e_i, e_{i'}) = \frac{p_i c_{il}^2 + p_{i'} c_{i'l}^2}{\lambda_l}.$$

2.4.2 Interprétation des variables

On va essayer l'interprétation des résultats pour les variables.

Qualité de représentation du nuage des variables

Pour donner une signification à la composante principale c_l, il faut la relier aux variables initiales x_j, en calculant le coefficient de corrélation $r(x_j, c_l)$ et on s'intéresse au plus fort coefficient en valeur absolue.

Chaque variable représentée par les coordonnées : $(r(c_1, x_j), r(c_2, x_j))$ est dans un cercle de corrélation de rayon 1.

On exprime la qualité de représentation d'une variable quantitative x_j sur le $l^{iéme}$ axe factoriel, par le coefficient de corrélation linéaire $r(c_l, x_j)$ entre la variable initiale x_j et la composante principale c_l tel que :

$$r(c_l, x_j) = \sqrt{\lambda_l} u_{jl}.$$

Preuve.

Comme $r(c_l, x_j) = r(c_l, z_j) = \dfrac{cov(z_j, c_l)}{s_{c_l} s_{z_j}}$[1]

Alors

$$\begin{aligned} cov(z_j, c_l) &= z_j^t D c_l \\ &= z_j^t D z u_l, \text{ car } c_l = Z u_l \\ &= R u_l, \text{ car } Z_j^t D Z = R \\ &= \lambda_l u_l, \text{ car } R u_l = \lambda_l u_l. \end{aligned}$$

Donc

$$\begin{aligned} r(c_l, z_j) &= \frac{\lambda_l u_l}{s_{c_l} s_{z_j}} \\ &= \frac{\lambda_l u_l}{\sqrt{\lambda_l}} \\ &= \sqrt{\lambda_l} u_{jl}. \end{aligned}$$

■

[1] s_{c_l} et s_{z_j} : écarts types de S_{c_l} et S_{z_j} respectivement.

Contribution d'une variable j par rapport à l'axe l

La contribution de la variable j à la composante c_l est définie par :

$$CTR_l\left(x_j\right) = \frac{r^2\left(c_l, x_j\right)}{\sum_{j=1}^{p} r^2\left(c_l, x_j\right)}.$$

Puisque $\lambda_l = \sum_{j=1}^{p} r^2\left(c_l, x_j\right)$, on peut aussi définir la contribution comme suit :

$$CTR_l\left(x_j\right) = u_{jl}^2.$$

2.5 Représentation d'élément supplémentaire

Les éléments supplémentaires ou illustratifs peuvent être des variables ou des individus.
Les individus et les variables supplémentaires permettent d'enrichir l'interprétation des axes sans avoir a participer ni a leurs constructions ni a leurs déterminations des axes.

2.5.1 Représentation des individus supplémentaire

Pour faire la représentation des individus supplémentaires dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des individus dans le système des axes principaux.
On note par $y = (y_{1,}y_{2,}..., y_p)^t \in \mathbb{R}^p$ un nouvel individu appelé individu supplémentaire. On a le calcul suivant :

$$y^t u_1, y^t u_2, ..., y^t u_k.$$

2.5.2 Représentation des variables supplémentaire

Pour faire la représentation des variables supplémentaires dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des variables dans le système des axes principaux.

On note par $t = (t_1, t_2, ..., t_n)^t \in \mathbb{R}^n$ une nouvelle variable appelée variable supplémentaire. On a le calcul suivant :

$$\frac{t^t D c_l}{\sqrt{\lambda_l}} = r\left(t, c_l\right).$$

Remarque 2.5.1

Un exemple d'application est faite dans la partie "Annexe B" sous logiciel R voir "Annexe A", où on va étudier les températures mensuelles de 15 villes de France sur 30 ans. Les données sont prisent du Quid 1986, page 507 (éditions Robert Laffont).

Conclusion

Dans ce travail, on a présenté l'Analyse en composantes principales ACP comme une méthode de base en statistique exploratoire multidimensionnelle. L'objectif de cette méthode est d'obtenir une représentation simple du nuage des données plus proche de la réalité dans un espace de dimension faible, permettant ainsi l'étude de la ressemblance entre les individus et la corrélation entre les variables, ou ces informations pertinentes sont résumées et visualisées tableau des données.

L'ACP et ses variantes sont utilisées dans divers domaines à savoir en finance, marketing, économie, ingénierie, biologie, ...ect. Ces techniques sont originales pour mesurer par exemple la respiration, la position...ect.

Annexe A : Logiciel R

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

• R a été créé par Ross Ihaka et Robert Gentleman en 1996 du département de statistique de l'Université d'Auckland, en Nouvelle Zélande, et est maintenant développé par la R développement Core Team. Il est conçu pour pouvoir être utilisé avec les système d'exploitation Unix, Linux, Windows et MacOS.

Le R est un application n'offrant qu'une invite de commande il basé sur la notion de vecteur, ce qui simplifie les calculs mathématique et réduit considérablement le recours aux structures itératives (boucles for, ...ect). Programmes courts, en général quelques lignes de code seulement. Temps de développement très court.

Annexe B : Exemple d'application

Pour 15 villes de France, on dispose des moyennes des températures mensuelles calculées sur 30 ans (entre 1931 et 1960). Elles sont rassemblées dans le tableau (2.1), qui croise ces 15 villes en lignes (individus) et les 12 mois de l'année en colonnes (variables). Différents packages et fonctions utilisés sont disponibles dans les bibliothèques standard de *R*.

Tableau des données :

	Janv	Févr	Mars	Avri	Mai	Juin	Juil	Aoû	Sept	Octo	Nove	Déce
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3
Lille	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1
Marseille	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5
Nantes	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4

TAB. 2.1 – Températures mensuelles de 15 villes de France.

Janv : Janvier,	Avri : Avril,	Juil : Juillet,	Octo : Octobre.
Févr : Février,	Mai : Mai,	Aoû : Août,	Nove : Novembre.
Mars : Mars,	Juin : Juin,	Sept : Septembre,	Déce : Décembre.

Packages :

ade4, FactoMineR.

Fonctions :

read.table, colMeans, cov, cor, scale, dudi.pca, sum, barplot, abline, symbols, s.corcircle.

Programmation :

library(ade4) # Il Contient des fonctions d'analyse des données.

library(FactoMineR) # Analyse exploratoire multidimensionnelle des données.

setwd("D :/TP/ACP")

X=read.table("Classeur1.txt",h=T) # Importer le tableau a partir du logiciel "excel".

g =colMeans(X) # Centre de gravité g.

round($g, 3$)

3.973	4.833	8.233	10.980	14.433	17.833	19.833	19.567	16.987	12.320	7.927	4.847

$S = cov(X)$ # Matrice de covariance S.

round($S, 3$)

$$S = \begin{bmatrix} 4.029 & & & & & \\ 3.651 & 3.491 & & & & \\ 2.564 & 2.660 & 2.338 & & & \\ 1.737 & 2.012 & 1.989 & 2.002 & & \\ 1.095 & 1.537 & 1.764 & 2.028 & 2.264 & \\ \vdots & . & ... & . & . & . \end{bmatrix}.$$

R =cor(X) # Matrice de corrélation R.

round($R, 3$)

$$R = \begin{bmatrix} 1.000 & & & & & \\ 0.973 & 1.000 & & & & \\ 0.835 & 0.931 & 1.000 & & & \\ 0.611 & 0.761 & 0.920 & 1.000 & & \\ 0.363 & 0.547 & 0.767 & 0.953 & 1.000 & \\ \vdots & . & ... & . & . & . \end{bmatrix}.$$

Z =scale(X) # Tableau standard Z.

round(Z, 3)

$$Z = \begin{bmatrix} 0.810 & 0.946 & 1.352 & 1.286 & 0.908 & ... & . \\ 1.059 & 0.517 & -0.283 & -1.258 & -1.883 & & . \\ -0.684 & -0.607 & -0.480 & -0.481 & -0.421 & & . \\ -1.232 & -0.874 & -0.349 & -0.269 & 0.044 & & . \\ -0.784 & -1.035 & -1.461 & -1.470 & -1.351 & & . \\ . & . & ... & . & . & ... & . \end{bmatrix}.$$

acp=dudi.pca(X,center=T,scale=T,nf=2,scannf=F) # Utilisation de l'ACP.

vp=acp$eig # Valeurs propres λ.

round(vp,3)

9.582	2.276	0.070	0.040	0.014	0.008	0.006	0.002	0.001	0.000	0.000	0.000

pvp=(vp/sum(vp))*100 # Pourcentage des vps.

round(pvp,3)

79.848	18.970	0.583	0.331	0.117	0.067	0.050	0.015	0.012	0.004	0.002	0.000

barplot(pvp,ylab="%d'inertie",names.arg=(round(pvp,3)),col=1) # Histogram des vps.

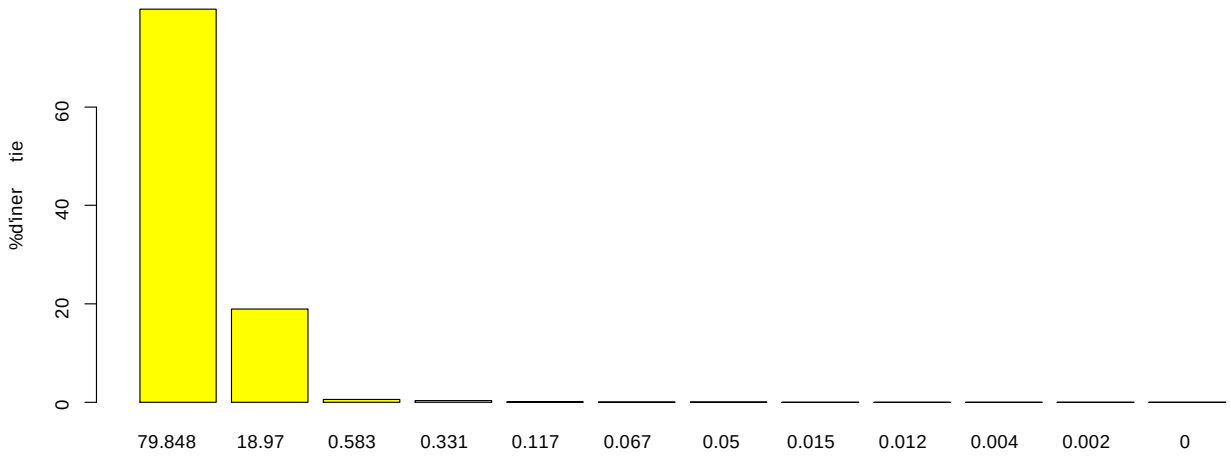

FIG. 2.1 – Eboulis des valeurs propres en %.

Commentaire :

L'inertie expliquée par la $1^{ére}$ dimension est de 79.848%, la $2^{iéme}$ dimension est de 18.970%...ect. En assemblant ces deux premiers pourcentage on obtient environ 98.818% d'inertie totale égale à $12 = I_g$ i.e une bonne qualité sur ce plan.

c_1=acp$co[,1] # $1^{ére}$ composante principale c_1.

round(c_1, 3)

c_2=acp$co[,2] # $2^{ième}$ composante principale c_2.

round(c_2, 3)

contribc=contrib$col.abs # Contribution $CTR_l\left(x_j\right)$.

round(contribc,3)

plot(c_1,c_2,type="n",ylab="comp1 :79.848%",xlab="comp2 :18.970%",main="les mois",

xlim=c(-1,1),ylim=c(-1,1),col=1)

abline(h=0,v=0)

text(c_1,c_2,row.names(acp$co),col=1) # Tracer le graphe des deux composantes c_j et $c_{j'}$.

symbols(0,0,circles=1,ylab="comp1 :79.848%",xlab="comp2 :18.970%",inches=F,add=T)

for(i in 1 :12){

arrows(0,0,c_1[i],c_2[i],angle=20,length=0.15)}

s.corcircle(acp$co) # Cercle des correlations.

	Coordonnées		Contribution	
Mois	c_1	c_2	c_1	c_2
Janv	−0.761	0.644	6.048	18.238
Févr	−0.880	0.469	8.090	9.666
Mars	−0.969	0.156	9.795	1.069
Avri	−0.969	−0.204	9.806	1.822
Mai	−0.873	−0.475	7.950	9.899
Juin	−0.864	−0.499	7.783	10.953
Juil	−0.842	−0.531	7.391	12.406
Aoû	−0.899	−0.430	8.427	8.120
Sept	−0.974	−0.208	9.901	1.902
Octo	−0.980	0.170	10.026	1.276
Nove	−0.904	0.414	8.524	7.527
Déce	−0.774	0.624	6.258	17.121

TAB. 2.2 – Composantes et Contribution des variables.

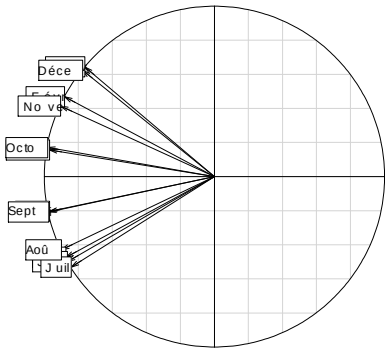

FIG. 2.2 – Représentation des variables.

Commentaire :

On observe que tout les coordonnées sur le $1^{ére}$ axe proche de 1 en valeur absolue i.e la valeur de corrélation entre ces variables et cet axe est fortement et positivement donc les variables ce bien représent sur cet axe. Sur le même tableau (2.2), on observe que la valeur de corrélation entre ces variables et ce $2^{iéme}$ axe est faible.

On conclue on peut dire que les variables sont bien représentes sur le $1^{ére}$ plan principal [Voir la représentation des variables] .

$co1$=acp$li[,1] # $1^{ére}$ composante principale $co1$.

round($co1, 3$)

$co2$=acp$li[,2] # $2^{ième}$ composante principale $co2$.

round($co2$,3)

contrib=inertia.dudi(acp,row.inertia=T,col.inertia=T)

contribl=contrib$row.abs # Contribution $CTR_l\,(e_i)$.

round(contribl,3)

plot($co1$,$co2$,ylab="axe1 :79.848%",xlab="axe2 :18.970%",xlim=c(-7,7),ylim=c(-4.5,4.5),col=1)

abline(h=0,v=0)

text($co1$,$co2$,row.names(acp$li),col=1,cex=1) # Tracer le graphe des villes celons les 2 axes.

range($co1$) # Borne du 1^{er} axe.

-6.007	4.217

range($co2$) # Borne du $2^{iéme}$ axe.

-2.172	4.093

	Coordonnées		Contribution	
villes	$co1$	$co2$	$co1$	$co2$
Bordeaux	-3.121	0.109	6.776	0.035
Brest	2.268	4.093	3.579	49.069
Clermont	1.726	-0.593	2.073	1.028
Grenoble	1.529	-1.688	1.627	8.344
Lille	4.217	0.595	12.372	1.037
Lyon	0.835	-1.788	0.485	9.365
Marseille	-4.833	-0.829	16.250	2.012
Montpellier	-4.147	-0.435	11.967	0.555
Nantes	0.281	1.115	0.055	3.638
Nice	-6.007	0.789	25.106	1.825
Paris	1.242	-0.156	1.073	0.072
Rennes	1.439	1.671	1.440	8.178
Strasbourg	4.106	-2.172	11.728	13.819
Toulouse	-1.736	-0.136	2.097	0.054
Vichy	2.201	-0.575	3.372	0.969

TAB. 2.3 – Composantes et Contribution des individus.

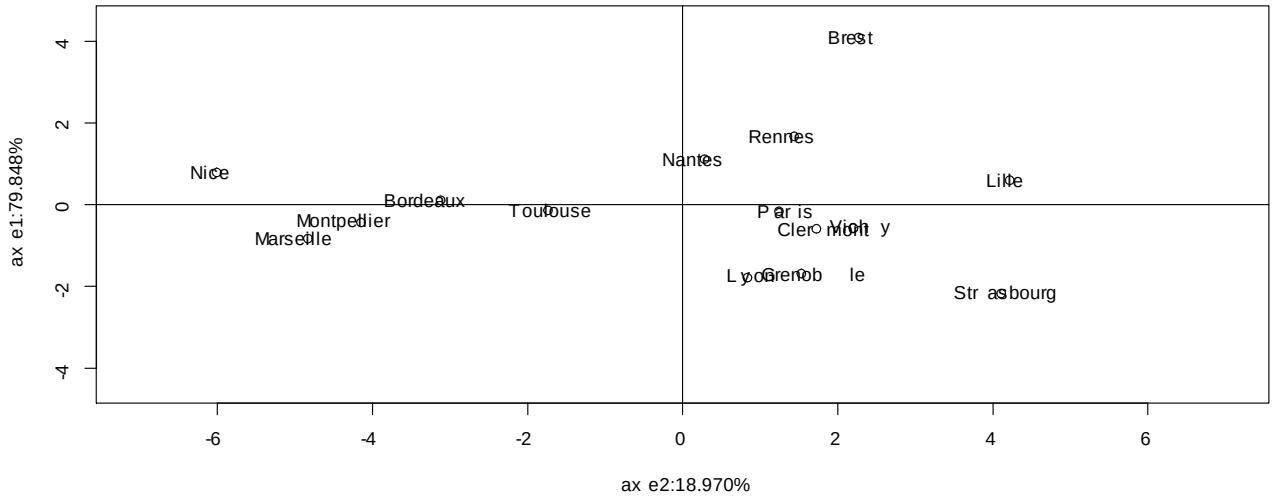

FIG. 2.3 – Représentation de nuage des individus.

Commentaire :

On compare les coordonnées de la $1^{ére}$ composante principale à la racine carrée de la $1^{ére}$ vp, i.e : $\sqrt{\lambda_1} = \sqrt{9.582} = 3.095$, où on prend seulement les individus qui ont des coordonnées supérieurs ou égales à $\sqrt{\lambda_1}$ en valeur absolue, puis on regroupe d'après ces signes.

Le tableau suivant contient six villes devisées sur 2 groupes qui sont bien représentées sur première axe

−	+
Bordeaux	Lille
Marseille	Strasbourg
Montpellier	
Nice	

On a Lille, Strasbourg, Bordeaux, Marseille, Montpellier, Nice sont bien représentées sur l'axe 1.

De la même manière on compare les coordonnées des individus par la $2^{ième}$ composante principale à la racine carrée de la $2^{ième}$ vp i.e : $\sqrt{\lambda_2} = \sqrt{2.276} = 1.509$ ou on prend seulement les individus qui ont des coordonnées supérieur ou égale à $\sqrt{\lambda_2}$ en valeur absolue puis on

regrouper d'après ces signes.

−	+
Grenoble	Brest
Lyon	Rennes
Strasbourg	

On a Brest, Rennes, Grenoble, Lyon, Strasbourg sont bien représentées sur l'axe 2, [Voir la représentation

Remarque 2.5.2

- *Clermont, Nantes, Paris, Toulouse et Vichy sont bien représentées sur le plan principal.*

Annexe C : Abréviations et Notations

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

X	:	tableau des données.
n	:	nombre des individus.
p	:	nombre des variables.
x_j	:	j-éme variable.
$\overline{x_j}$	:	moyenne de la variable x_j.
$cov(.,.)$	:	covariance.
$var\,(.,.)$	:	variance.
$d\,(e_i, e_{i'})$	:	distance entre e_i et $e_{i'}$.
S	:	matrice de variance de tableau X.
D	:	matrice de poids.
R	:	matrice de corrélation.
Z	:	tableau des données centrés réduites.
Y	:	tableau des données centrés.
M	:	métrique.
I_g	:	inertie tatale.
P	:	matrice de projection.
R^n	:	espace des nombres réels de dimention n.
S_{proj}	:	matrice de variance de nuage projeté.
f_i	:	projection de l'individu $e_{i.}$

R^p	:	espace des nombres réels de dimention p.
g	:	centre de gravité.
p_i	:	poids.
$r(x_j, x_{j'})$	:	cofficient de corrélation entre x_j et $x_{j'}$.
I_n	:	matrice d'identité de taille n.
1_n	:	vecteur unitaire d'indentité de taille n.
tr	:	trace d'une matrice.
λ	:	valeur propre.
c_{il}	:	$i-éme$ coordonnées de la composante principale c_l.
u_{jl}	:	$j-éme$ coordonnées de la facteur principal u_l.
F_k	:	sous-espace de dimention k.
$\theta_{jj'}$	:	angle entre deux vecteurs.
$proj$	:	projection.
l	:	axe pricipale.
Δ_l	:	$l-éme$ droite.
$\langle , \rangle$	:	produit scalaire.
$QLT(F_k)$	:	qualité sur F_k.
$QLT_l(e_i)$	:	qualité de e_i sur l'axe l.
$QLT_{(l,l')}(e_i)$	:	qualité de e_i sur plan (l, l').
$CTR_l(x_j)$	:	contribution de x_j sur l'axe l.
$CTR_l(e_i)$	:	contribution sur l'axe l de e_i.
$CTR_l(e_i, e_{i'})$	:	contribution sur l'axe l de couple $(e_i, e_{i'})$.
$\mathcal{M}_{\mathbb{R}}(n, p)$	:	L'ensemble des matrices de type (n, p) à coeffcients dans $\mathbb{R}$.
vps	:	valeurs propres.
i.e	:	c'est-à-dire.

Bibliographie

[1] Baccini, A., Besse, P. (septembre 2005) Data mining I Exploration Statistique, Unversité Paul Sabatier — 31062. Toulouse.

[2] Boumaza, R. (2007) Analyse des données-ACP, AFC et ACM-Mise en œuvre avec R. CPU.

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

[4] Duby, C., Robin, S. (10 Juillet 2006) Analyse en composantes principales. INA. Paris-Grignon.

[5] Escofier, B., Pagés, J. (2008) Analyse factorielle simples et multiples. Objectif, méthodes et interprétation. Dunod.

[6] Ihaka, R., Gentleman, R. (1996) R : A language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5 : 299-314.

[7] Merad, M. (22 Octobre 2015) Méthodes ACP et AFC en statistiques et leurs applications. UAB. Tlemcen.

[8] Meraghni, D. (2018) Cours de master 1. UMK. Biskra.

[9] Saporta, G. (2010). Probabilités, analyse des données et statistiques (2 éme édition). Technip, Paris.

[10] Tomalala, R.-R.(2007) Analyse en composantes principales.Univ.Lyon 2.