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

Ziza Soulaf

Titre :

Exemple sur l'analyse en composantes principales

Membres du Comité d'Examen :

Pr.	MERAGHNI DJAMEL	UMKB	Président
Pr.	BRAHIMI BRAHIM	UMKB	Encadreur
Dr.	ROUBI AFAF	UMKB	Examinatrice

Juin 2021

Dédicace

Avant tout propose, je tiens à rendre grâce à **ALLAH** qui ma guidé sur la bonne voie.

Je dédie ce modeste travail qui est le fruit de toutes mes années des études.

Tout d'abord :

A la lumiére et symbole de la vie, a la source de tendresse ≪**Ma mére**≫, pour son grand amour, ses sagrifices et toute l'affection qu'elle m'a toujours offerte. Tu es l'exemple de dévoument qui n'a pas cessé de m'encourager et de prier pour moi.

A mon secret de ma réussite, à mon adorable ≪**Mon pére**≫, qui me soutient et qui est toujours présent pour moi, tes encoragements et ton mtivation qui me réalise cette réussite.

A mes fréres :

Yahia et **Abdesslam.**

A mes soeurs.

A tous les étudiantes de **mathématique**, surtout $2^{éme}$ master goupe de **statistique** et tous les étudiants de l'université **Mohammed Khieder**.

Ziza Soulaf.

REMERCIEMENTS

En terminant mon mémoire de fin d'études. Je remercie bien **ALLAH** qui ma donné la force suffisante et la volonté pour faire ce modeste travail.

Tout d'abord :

Un mot de remerciement et de gratitude à tous ceux qui m'ont aidé à terminer mes études, et m'ont donné de l'éspoir qui m'a motivé à m'efforcer et à persévérer pour réussir. Mercie beaucoup : **AISSI ABDELKRIM**, **NEDJAHI NEDJLA** et **YEZZA AHMED**.

Un remerciement à mon encadreur : Pr. **BRAHIMI BRAHIM** pour le suivi et l'aide qu'elle m'a apporté pour l'élaboration de ce mémoire.

Je tiens aussi remerciement à tous les enseignements de département de **mathématique**. Je remercie les membres du jury :

Pr. **MERAGHNI DJAMEL** et Dr. **ROUBI AFAF.**

Je remercier toutes les personnes qui ont participé de prés ou de loin à la réalisation de ce travail.

Enfin, je remercie ma famille qui mon encouragé au long de ma vie, spécialement mes chéres parents et mes fréres.

Merci à tous.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'analyse des données est une des branches les plus vivantes de la statistique, ses principales méthodes se séparent en deux groupes :

- **Les méthodes de classification.**
- **Les méthodes factorielles.**

L'analyse en composantes principales **(ACP)** ou principal component analysis **(PCA)** en englai. Fait partie du groupe des méthodes descriptives multidimensionelle sappelées méthodes factorielles. Ces méthodes remente à **K. pearson (1901)**, elles ont été surtout développées en France dans les années **60**, en particulier par **Jean-Paul Benzékri** qui a baucoup exploité les aspects géométriques et les représentations graphiques. **l'ACP** propose à partir d'un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n individus, on proposait de réduire la dimension de l'espace en projetant le nuage des points individus sur le sous-espace de dimension inférieure pour obtenir une visualisation de l'ensemble des liaisons entre variables tout en minimisant la perte de l'information. Le but de mon travail est de mettre en évidence le rôle de **l'ACP** dans la pratique.

Ce mémoire se compose de trois chapitres principaux :

Chapitre1 : On va présenter quelques définitions, proposition, propriétés... ect. En dautres 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 caractéristique. On a aussi essayé d'interpréte les résultats de **l'ACP**.

Chapitre3 : à l'aide du l'ogiciel **R**, On va effectuer un exemple d'étude de différentes caractéristique de l'approche de **l'ACP**.

Chapitre 1

Préliminaires

1.1 Données et leurs caractéristiques

1.1.1 Tableau des données

Un tableau des observations ets une matrice notée X de type (n, p), contenant en lignes n individus $(e_1, \ldots, e_n)$, et en colonnes p variables quantitatives $(X_1, \ldots, X_p)$:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ \vdots & x_{ij} & \vdots \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \in M_{\mathbb{R}}(n, p).$$

x_{ij} la valeur de la $j^{éme}$ variable X_j mesurée sur le $i^{éme}$ individu e_i.

les lignes de la matrice X représentent des individus, et les colonnes sont des variables.

Définition 1.1.1 ***(individu)*** *La $i^{éme}$ ligne de la matrice X est représente e_i, et*

c'est un vecteur de n composantes, telles que :

$$e_i = \begin{pmatrix} x_{i1} & , \dots & , x_{ip} \end{pmatrix}^t \in \mathbb{R}^p \text{ , pour } i = \overline{1,n}.$$

Définition 1.1.2 ***(variable)*** *La $j^{éme}$ colonne de la matrice X est représente la variable X_j, qui est une liste des n valeurs prises pour n individus, telles que :*

$$X_j = \begin{pmatrix} x_{1j} & , \dots & , x_{nj} \end{pmatrix}^t \in \mathbb{R}^n, \text{ pour } j = \overline{1,p}.$$

1.1.2 Matrice des poids

Si les données ont été recueillies d'un tirage aléatoire à probabilités égales, alors les n individus ont tous même importance $\frac{1}{n}$, et ce n'est pas toujours le cas. aux contraire cas, il utile quand les individus n'ont pas la même importance, on associe à chaque individu un poid p_i, avec : $\sum_{i=1}^{n} p_i = 1.$ représentant dans la matrice diagonale notée D_p de type (n, n) tele que :

$$D_p = \begin{bmatrix} p_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n \end{bmatrix} \in M_{\mathbb{R}}(p,p) \text{ , avec } p_i \geqslant 0.$$

Dans le cas le plus usuel de poids égaux : $p_i = \frac{1}{n}$ et $D_p = \frac{1}{n} I_n$.
avec I_n est le vecteur unitaire de $\mathbb{R}^n$ ($I_n = \begin{pmatrix} 1 & , \dots & , 1 \end{pmatrix}^t \in \mathbb{R}^n$).

Preuve. Dans le cas uniforme tous les individus ont le même poids i.e :

$p_1 = p_2 = \cdots = p_n$, et on a $\sum_{i=1}^{n} p_i = 1$, alors :

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

■

par conséquent :

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

Et

$$D_p = \begin{bmatrix} \frac{1}{n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \frac{1}{n} I_n.$$

1.1.3 Centre de gravité

Nous choisissons le centre de gravité du nuage comme origine pour faciliter la représentation graphique du nuage de points. Le centre de gravité est le vecteur g, qui contient les valeurs des moyennes empiriques des variables $\left(\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \text{ pour } j = \overline{1,p}\right)$, et qui est aussi appelé point moyen ou individu moyen. Il est définit par :

$$g = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \begin{bmatrix} \overline{X_1} \\ \overline{X_2} \\ \vdots \\ \overline{X_p} \end{bmatrix} \in \mathbb{R}^p.$$

peut s'écrire sous la forme matricielle :

$$g = X^t D_p 1_n \text{ , où } 1_n = \begin{pmatrix} 1, & \ldots, & 1 \end{pmatrix}^t \in \mathbb{R}^n.$$

Preuve. On a :

$$X^t D_p 1_n = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_n \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n p_i x_{i1} \\ \sum_{i=1}^n p_i x_{i2} \\ \vdots \\ \sum_{i=1}^n p_i x_{ip} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i2} \\ \vdots \\ \sum_{i=1}^n x_{ip} \end{bmatrix} = \begin{bmatrix} \overline{X_1} \\ \overline{X_2} \\ \vdots \\ \overline{X_p} \end{bmatrix}.$$

Donc

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

■

1.1.4 Standardisation du tableau

Dans l'analyse en composantes principales (**l'ACP**). La transformation des données principales se fait de deux façons :

1. **Centrer les données** : les variables (colonnes) de moyenne nulle.
2. **Réduire les données** : les variables (colonnes) de variance égale à 1.

Tableau centré

A partire du tableau X, on obtient un tableau centré Y en centrant les variables autour de leur moyenne :

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

La forme matricielle :

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

Preuve.

$$X - 1_n g^t = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ \vdots & x_{ij} & \vdots \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} - \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{pmatrix} \overline{X_1}, & \cdots, & \overline{X_j}, & \cdots, & \overline{X_p} \end{pmatrix}$$

$$= \begin{bmatrix} x_{11} - \overline{X_1} & \cdots & x_{1p} - \overline{X_p} \\ & \vdots & \\ \vdots & x_{ij} - \overline{X_j} & \vdots \\ & \vdots & \\ x_{n1} - \overline{X_1} & \cdots & x_{np} - \overline{X_p} \end{bmatrix} = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ & \vdots & \\ \vdots & y_{ij} & \vdots \\ & \vdots & \\ y_{n1} & \cdots & y_{np} \end{bmatrix} = Y \in M_{\mathbb{R}}(n,p).$$

■

Tableau centré-réduite

A partire du tableau centré Y, on obtient un tableau centré et réduite Z, on divise les coordonnées de chaque colonne par l'écart-type correspondant. Plus précisément,

notons :

$$\sigma_j^2 := \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - X_j\right)^2, j = \overline{1,p}, \text{ (la variance empirique de la variable } X_j).$$

les données centrées et réduites :

$$z_{ij} = \frac{y_{ij}}{\sigma_j} = \frac{x_{ij} - \overline{X_j}}{\sigma_j}, \text{ pour } i = \overline{1,n} \text{ et } j = \overline{1,p}.$$

La forme matricielle. On définit la matrice poids $D_{1/\sigma}$, des inverses des écart-types :

$$D_{1/\sigma} = \begin{bmatrix} \frac{1}{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_p} \end{bmatrix} \in M_{\mathbb{R}}(p,p).$$

On peut écrire

$$Z = YD_{1/\sigma}.$$

Preuve.

$$YD_{1/\sigma} = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ & \vdots & \\ \vdots & y_{ij} & \vdots \\ & \vdots & \\ y_{n1} & \cdots & y_{np} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_p} \end{bmatrix}$$

$$= \begin{bmatrix} y_{11}/\sigma_1 & \cdots & y_{1p}/\sigma_p \\ & \vdots & \\ \vdots & y_{ij}/\sigma_j & \vdots \\ & \vdots & \\ y_{n1}/\sigma_1 & \cdots & y_{np}/\sigma_p \end{bmatrix} = \begin{bmatrix} \frac{x_{11}-\overline{X_1}}{\sigma_1} & \cdots & \frac{x_{1p}-\overline{X_p}}{\sigma_p} \\ & \vdots & \\ \vdots & \frac{x_{ij}-\overline{X_j}}{\sigma_j} & \vdots \\ & \vdots & \\ \frac{x_{n1}-\overline{X_1}}{\sigma_1} & \cdots & \frac{x_{np}-\overline{X_p}}{\sigma_p} \end{bmatrix}$$

$$= \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ & \vdots & \\ \vdots & z_{ij} & \vdots \\ & \vdots & \\ z_{n1} & \cdots & z_{np} \end{bmatrix} = Z \in M_{\mathbb{R}}(n,p).$$

■

1.1.5 Matrice de variance-covariance

La matrice de variance-covariance des variables X_j et $X_{j'}$; $j = \overline{1,p}$, c'est une matrice carrée de dimension p notée par V :

$$V = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \\ \sigma_{p1} & \cdots & & \sigma_p^2 \end{bmatrix} \in M_{\mathbb{R}}(p,p).$$

Où, $\sigma_{jj'}$: est la covariance des variables X_j et $X_{j'}$:

$$\sigma_{jj'} = cov(X_j, X_{j'}) = \sum_{i=1}^{n} p_i \left(x_{ij} - \overline{X_j}\right)\left(x_{ij'} - \overline{X_{j'}}\right) \ ; \ j, j' = \overline{1,p}.$$

σ_j^2 : est la variance de la variable X_j :

$$\begin{aligned} \sigma_j^2 = \sigma_{jj} &= cov(X_j, X_j) = var(X_j) \\ &= \sum_{i=1}^{n} p_i (x_{ij} - X_j)^2 \ ; \ j = \overline{1,p}. \end{aligned}$$

peut s'écrire sous la forme matricielle :

$$\begin{aligned} V &= Y^t D_p Y \\ &= Y X^t D_p X - g g^t. \end{aligned}$$

Dans le cas uniforme $\left(p_1 = p_2 = \cdots = p_n = \frac{1}{n}\right)$:

$$\begin{aligned} V &= \frac{1}{n} Y^t Y \\ &= \frac{1}{n} X^t X - g g^t. \end{aligned}$$

Preuve. On a :

$$\begin{aligned} V &= Y^t D_p Y \\ &= \left(X - 1_n g^t\right)^t D_p \left(X - 1_n g^t\right) \qquad ; \left(\text{car } Y = X - 1_n g^t\right). \\ &= X^t D_p X - X^t D_p 1_n g^t - g 1_n^t D_p X + g 1_n^t D_p 1_n g^t \\ &= X^t D_p X - g g^t - g g^t + g g^t \qquad ; \left(\text{car } 1_n^t D_p 1_n = \sum_{i=1}^{n} p_i = 1 \ \text{ et } \ X^t D_p 1_n = g\right). \\ &= X^t D_p X - g g^t. \end{aligned}$$

■

Remarque 1.1.1 *:*

1. Comme $V^t = \left(Y^t D_p Y\right)^t = Y^t D_p Y = V$, la matrice V est **symétrique**.
2. La matrice V admet p valeurs propres.

1.1.6 Matrice de corrélation

La matrice de corrélation des variables X_j et $X_{j'}$, c'est une matrice carrée de dimension p notée par R :

$$R=\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \in M_{\mathbb{R}}(p,p).$$

avec $r_{jj'} = \frac{cov\left(X_j,X_{j'}\right)}{\sigma_j\sigma_{j'}} = \frac{\sigma_{jj'}}{\sigma_j\sigma_{j'}}$, et $r_{jj} = \frac{cov(X_j,X_j)}{\sigma_j\sigma_j} = \frac{\sigma_j^2}{\sigma_j^2} = 1$.

La forme matricielle :

$$\begin{aligned} R &= D_{1/\sigma} V D_{1/\sigma} \\ &= Z^t D_p Z. \end{aligned}$$

Preuve. On a

$$D_{1/\sigma}VD_{1/\sigma} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_p} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \\ \sigma_{p1} & \cdots & & \sigma_p^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_p} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \sigma_{12}/\sigma_1\sigma_2 & \cdots & \sigma_{1p}/\sigma_1\sigma_p \\ \sigma_{21}/\sigma_2\sigma_1 & 1 & & \vdots \\ \vdots & & \ddots & \\ \sigma_{p1}/\sigma_p\sigma_1 & \cdots & & 1 \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \vdots \\ \vdots & & \ddots & \\ r_{p1} & \cdots & & 1 \end{bmatrix} = R\,.$$

pour $Z^t D_p Z = R$, On a :

$$\begin{aligned} Z^t D_p Z &= \left(Y D_{1/\sigma}\right)^t D_p \left(Y D_{1/\sigma}\right) \\ &= D_{1/\sigma}\ Y^t D_p Y\ D_{1/\sigma} \\ &= D_{1/\sigma} V D_{1/\sigma} \\ &= R. \end{aligned}$$

Remarque 1.1.2 *:*

1. *Comme $R^t = (Z^t D_p Z)^t = Z^t D_p Z = R$, la matrice R est* ***symétrique.***
2. *Comme il ya p variables, cela nous conduit donc à calculer $\frac{p(p-1)}{2}$ corrélations.*
3. *R est la matrice de variance-covariance des données centrées et réduites et résumé la structure des dépendance linéaires entre les p variables prise deux à deux.*

■

1.2 Nuage de points (individus)

Pour chaque individu i on associer un vecteur contenant ses observations sur les p variables (i.e. la $i^{émé}$ ligne de X). Chaque individu est considéré comme un point d'un espace vectoriel de dimension p : c'est l'espace des individus. Alors l'ensemble des n individus est un nuage de points, et g est son centre de gravité.

1.2.1 Ressemblance entre deux individus

On considéré n individus observés sur p variables quantitatives. Deux individus se ressemblant (sont proches), s'ils possédent des valeurs proches pour lensemble des

variables. En Analyse en composantes principales (**ACP)** on utilise le plus souvent la distance euclidienne, la distance entre deux individus e_i et $e_{i'}$ est égale à :

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

1.2.2 Métrique

En générale : La distance utilisée entre deux individus e_i et $e_{i'}$, est définie par la forme quadratique :

$$d^2(e_i, e_{i'}) = (e_i - e_{i'})^t M (e_i - e_{i'}) = \langle e_i, e_{i'} \rangle_M .$$

M : matrice symétrique définie positive de type $(p \times p)$.

Les métriques les plus utilisées en analyse en composantes principales (**ACP)**, sont les métriques. I_p(la matrice identité d'rdre p) et D_{1/σ^2}(la matrice diagonale des inverses des variances) :

$$D_{1/\sigma^2} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_p} \end{bmatrix} \in M_{\mathbb{R}}(p, p) .$$

Remarque 1.2.1 *:*

1. *Pour le tableau centré Y, on utilise la métrique $M = D_{1/\sigma^2}$.*
2. *Pour le tableau centré et réduite Z, on utilise la métrique $M = I_p$.*

Preuve. *On pose :*

$\bullet e_i^y = \begin{pmatrix} y_{i1} & , \dots & , y_{ip} \end{pmatrix}^t \in \mathbb{R}^p$, *(la $i^{émé}$ ligne de tableau Y).*

•$e_i^z = \begin{pmatrix} z_{i1} & , \dots & , z_{ip} \end{pmatrix}^t \in \mathbb{R}^p$, *(la $i^{émé}$ ligne de tableau Z).*

On a :

$$\begin{aligned}
\left\langle e_i^y, e_{i'}^y \right\rangle_{D_{1/\sigma^2}} &= \left(e_i^y - e_{i'}^y\right)^t D_{1/\sigma^2} \left(e_i^y - e_{i'}^y\right) \\
&= \sum_{j=1}^{p} \left(\frac{y_{ij} - y_{i'j}}{\sigma_j} \right)^2 \\
&= \sum_{j=1}^{p} \left(z_{ij} - z_{i'j}\right)^2 \\
&= \sum_{j=1}^{p} \left(\frac{z_{ij} - z_{i'j}}{1} \right)^2 \\
&= \left(e_i^z - e_{i'}^z\right)^t I_p \left(e_i^z - e_{i'}^z\right) \\
&= \left\langle e_i^z, e_{i'}^z \right\rangle_{I_p}.
\end{aligned}$$

■

1.2.3 Inertie

L'inertie totale du nuage de points est la moyenne pondérée des carrés des distances des points au centre de gravité g :

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

• Si $g = 0$ (tableau centré) alors $I_g = \sum_{i=1}^{n} p_i \left(e_i\right)^t M \left(e_i\right)$.

Remarque 1.2.2 *:*

Si ce l'inertie est grande, alors le nuage de points est trés dispersé, et à l'inverse s'il est petit, alors le nuage est trés concentré sur son centre de gravité g. L'inertie en

un point quelconque $a \in \mathbb{R}$:

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

La notion de **Huyghens** :

$$\begin{aligned} I_a &= I_g + (g-a)^t M (g-a) \\ &= I_g + \|g-a\|^2. \end{aligned}$$

Proposition 1.2.1 *:*

1. $I_g = Tr(MV) = Tr(VM)$.
2. *Si $M = I_p$, alors l'inertie est égale à la somme des variances des p variables :*

$$I_g = \sum_{j=1}^{p} \sigma_j^2 \ , \ j = \overline{1,p}.$$

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

$$I_g = p.$$

1.3 Nuage de points (variables)

Pour chaque variable j on associer un vecteur X_j contenant ses observations pour tous les individus (i.e. la $j^{émé}$ colonne de X). Chaque variable est considéré comme un point d'un espace vectoriel de dimension n : c'est l'espace des variables.

1.3.1 Liaison entre deux variables

En l'analyse en composantes principales **(ACP)**, mesurer la liaison entre deux variables X_j et $X_{j'}$par le coefficient de corrélation linéaire noté $r_{jj'}$:

$$r\left(X_j, X_{j'}\right) = \frac{cov\left(X_j, X_{j'}\right)}{\sqrt{var\left(X_j\right) var\left(X_{j'}\right)}}$$

Où la covariance observé entre deux variables X_j et $X_{j'}$, et donné par :

$$cov\left(X_j, X_{j'}\right) = \sigma_{jj'} = \sum_{i=1}^{n} p_i \left(x_{ij} - \overline{X_j}\right)\left(x_{ij'} - \overline{X_{j'}}\right) \text{ , pour } j, j' = \overline{1, p}.$$

avec $cov\left(X_j, X_j\right) = var\left(X_j\right)$ et $r\left(X_j, X_j\right) = 1$.

Propriété 1.3.1 *:*

1. $-1 \leqslant r\left(X_j, X_{j'}\right) \leqslant 1$.
2. X_j *et* $X_{j'}$ *sont linéairement liées* $\Longleftrightarrow \left|cov\left(X_j, X_{j'}\right)\right| = 1$.

1.3.2 Métrique des variables

1. Le produit scalaire de deux variables est :

$$\left\langle X_j, X_{j'}\right\rangle_{D_p} = X_j^t D_p X_{j'} = \sum_{i=1}^{n} p_i x_{ij} x_{ij'}, \text{ pour } j, j' = \overline{1, p}.$$

   si les deux variables sont centrées, alors $\left\langle X_j, X_{j'}\right\rangle_{D_p} = \sigma_{jj'}$.

2. La norme de X_j :

$$\left\|X_j\right\|_{D_p}^2 = \sigma_j^2.$$

3. L'angle $\theta_{jj'}$ entre deux variables centré Y_j et $Y_{j'}$:

$$\cos\left(\theta_{jj'}\right) = \frac{\left\langle Y_j, Y_{j'}\right\rangle_{D_p}}{\left\|Y_j\right\|_{D_p}\left\|Y_{j'}\right\|_{D_p}} = \frac{\sigma_{jj'}}{\sigma_j \sigma_{j'}}.$$

Remarque 1.3.1 *:*

On s'intéresse aux distances entre les points dans l'éspace des individus. Dans l'éspace des variables, on s'intéressera aux l'angles entre les vecteurs.

Chapitre 2

Analyse en composantes principales

2.1 Principe de l'ACP

On obtient une représentation des n individus dans un sous-espace F_k de $\mathbb{R}^p$ (de dimension faible $k < p$), c'est à dire, on cherche à définire k nouvelles variables dites combinaison linéaire des p variables initieles. En perdant le moins posible d'information.

2.1.1 Projection des individus sur un sous-espace

Le choix de l'espace de projection est tel que la moyenne des carrés des distances entre les projections et leur centre de gravité soit la plus grande possible. En d'autre terme, l'inertie du nuage projeté sur le sous-espace F_k soit maximale.

Soit le sous-espace de projection F_k. On définit P l'opérateur de projection M-orthogonale sur l'espace F_k tel que :

1. $P^2 = P$ (P est idempotente).

2. $P^tM = MP$ (P est M-symétrique).

$\star$ à chaque individu e_i (la $i^{éme}$ ligne de X) se projette sur F_k selon un vecteur colonne f_i, tel que : $f_i = Pe_i$; d'où $f_i^t = e_i^t P^t$ (i.e. la $i^{éme}$ ligne de tableau XP^t).

On écrit :

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

Proposition 2.1.1 *:*

Pour le tableau de nuage projeté X_{proj} :

$\star$ Matrice de variance-covariance :

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

$\star$ L'inertie :

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

$\star$ Centre de gravité :

$$g_{proj} = Pg.$$

Construction de sous-espace F_k

On décompose le sous-espace F_k comme la somme directe de ces sous-espace Δ_i de dimension 1 et orthogonaux entre eux :

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

On peut alors dire que :

$$I_{\mathbf{F}_k} = I_{\Delta_1} + I_{\Delta_2} + \cdots + I_{\Delta_k}.$$

Construction de la premiére droite Δ_1

On cherche dans $\mathbb{R}^p$ la droite Δ_1 de dimension 1, qui passe par le centre de gravité g, et qui maximise l'inertie de nuage projeté sur cette droite. Soit $a_1 \in \mathbb{R}^p$ (vecteur directeur de Δ_1), le projecteur $M-$orthogonale sure Δ_1 donnée par :

$$\begin{aligned} p_1 &= a_1 \left(a_1^t M a_1\right)^{-1} a_1^t M \\ &= \frac{a_1 a_1^t M}{a_1^t M a_1} \text{ , car } \left(a_1^t M a_1\right) \in \mathbb{R}. \end{aligned}$$

$\star$ En remplaçant le projecteur p_1 par sa formule dans la définition de l'inertie totale du nuage projeté I_{Δ_1}, on obtient :

$$\begin{aligned} I_{\Delta_1} &= Tr\left(VMp_1\right) \\ &= Tr\left(VM\frac{a_1 a_1^t M}{a_1^t M a_1}\right) \\ &= \frac{1}{a_1^t M a_1} Tr\left(VMa_1 a_1^t M\right) \\ &= \frac{1}{a_1^t M a_1} Tr\left(a_1^t MVMa_1\right) \text{ , car } Tr\left(AB\right) = Tr\left(BA\right) \\ &= \frac{a_1^t MVMa_1}{a_1^t M a_1}. \end{aligned}$$

Donc

$$I_{\Delta_1} = \frac{a_1^t MVMa_1}{a_1^t M a_1}; \left(a_1^t MVMa_1\right) \in \mathbb{R}.$$

Pour obtenir le maximum de $\frac{a_1^t MVMa_1}{a_1^t M a_1}$, il suffit d'annuler la dérivée de cette exprission par rapport à a_1, puis en résolvant cette derniére en l'annulant. On obtient :

$$VMa_1 = \frac{a_1^t MVMa_1}{a_1^t M a_1} a_1.$$

On pose $\frac{a_1^t MVMa_1}{a_1^t Ma_1} = \lambda \in \mathbb{R}$, alors :

$$VMa_1 = \lambda a_1.$$

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

Remarque 2.1.1 *:*

Le premier axe est celui qui aura la plus grande valeur propre λ_1, et le deuxieme axe sera celui de la deuxiéme valeur propre λ_2, est ainsi de suit.

2.2 Elémentes principaux de l'ACP

2.2.1 Axes principaux

On appelle axes principaux les p vecteurs propres $a_1, a_2, \ldots, a_p$ de la matrice VM associée à les valeurs propres λ_j, $M-$normés à 1. De plus ils sont $V^{-1}-$orthogonaux et $M-$orthonormé :

$$\begin{cases} VMa_j = \lambda_j a_j \quad \text{, pour } j = \overline{1,p}. \\ \|a_j\|_M^2 = 1 \end{cases}$$

Preuve. Soit a_j et $a_{j'}$ 2 axes principux.

On a :

$$\begin{aligned}
\langle a_j, a_{j'} \rangle_{V^{-1}} &= a_j^t V^{-1} a_{j'} \\
&= \frac{1}{\lambda_j} \left(VMa_j\right)^t V^{-1} a_{j'} \\
&= \frac{1}{\lambda_j} a_j^t MVV^{-1} a_{j'} \\
&= \frac{1}{\lambda_j} a_j^t M a_{j'} \\
&= \frac{1}{\lambda_j} \langle a_j, a_{j'} \rangle_M \\
&= \begin{cases} \frac{1}{\lambda_j} & \text{, si } j = j'. \\ 0 & \text{, si non.} \end{cases}
\end{aligned}$$

■

2.2.2 Facteurs principaux

Les facteurs principaux u_j sont les vecteurs propres M^{-1}-normés à 1, de la matrice MV associée à les valeurs propres λ_j, $j = \overline{1,p}$. De plus ils sont$V-$orthogonaux et $M^{-1}-$orthonormé :

$$\begin{cases} MVu_j = \lambda_j u_j & \text{, pour } j = \overline{1,p}\,. \\ \|a_j\|^2_{M^{-1}} = 1 \end{cases} \text{, où } u_j = Ma_j \in \mathbb{R}^p.$$

2.2.3 Composantes principales

Les composantes principales sont les vecteurs $c_j = (c_{1j}, \ldots, c_{nj})$de taille n, définies en fonction des facteurs principaux u_j par :

$$c_j = XMa_j = Xu_j.$$

Si on travailler avec le tableau centré et réduite alors :

$$c_j = Zu_j.$$

$\star$ Chaque composante c_j contient les coordonnées des projections $M-$orthogonales des n individus centrés sur les axes a_j.

Propriétés des composantes principales

1. Les c_j ne sont pas corrélées deux à deux i.e :

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

2. La variance d'une composante principale c_j est :

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

3. Les composantes principale c_j sont des combinaisons linéaires des variables initiales. De plus il sont les vecteurs propres de la matrice XMX^tD_p :

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

2.2.4 ACP sur les données centrées-réduites

En pratique pour accorder la même importance à chaque variable, on travaille sur le tableau centrée et réduite Z avec la métrique $M = I_p$, qui est utilisé lorsque les unités de mesure et les variances associées à chaque variable sont différentes. Dans ce cas les facteurs et les axes principaux sont les mêmes, car la matrice de covariance

V est égale à la matrice de corrélation R :

$$u_j = Ma_j = I_p a_j = a_j.$$

Qui sont les p vecteurs propres orthonormés de la matrice R associées aux ces valeurs propres sont d'ordre décroissant ($\lambda_1 > \lambda_2 > \cdots > \lambda_p$) :

$$Ru_j = \lambda_j u_j.$$

2.3 Interprétation et qualité de représentation

Le but de **l'ACP** est de construire de nouvelles variables dites artificielle, et d'obyenire une représentation graphique du nuage des individus sur un sous-espace de dimonsion k plus faible que p. Et fournit des représentations graphiques permettant de visualiser les relations entre les variables, ainsi que l'existence d'éventueles de groupes d'individus et ceux de variables.

2.3.1 Interprétation des individus

On va essayer de représente quelque définition sur l'interprétation des résultats pour les individus.

Qualité de représentation du nuage des individus sur F_k

La qualité de représentation obtenue par k valeurs propres est la proportion de l'inertie expliquée :

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

Où $0 \leqslant QLT\left(F_k\right) \leqslant 1$.

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

La qualité de représentation de l'individu i par rapport à l'axe l qui est donnée par :

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

avec θ_{il} : est l'angle formée entre le vecteur Z_i et l'axe l.

Remarque 2.3.1 *:*

1. En général, la qualité de la projection d'un individu i sur le plan (l, l') qui est donnée par :

   $$QLT_{l,l'}\left(e_i\right) = \cos^2\left(\theta_{i(l,l')}\right) = \frac{c_{il}^2 + c_{il'}^2}{\left\|Z_i\right\|^2}.$$

   Donc, on peut dire que :

   $$QLT_{l,l'}\left(e_i\right) = QLT_l\left(e_i\right) + QLT_{l'}\left(e_i\right).$$

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

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

La contribution d'un individu i à la composante c_l est définie par :

$$CTR_l\left(e_i\right) = \frac{p_i c_{il}^2}{\sum_{i=1}^{n} p_i c_{il}^2} = \frac{p_i c_{il}^2}{\lambda_l} \text{ , avec } l = \overline{1,k}.$$

Où, c_{il} : valeur de la composante principale l pour l'individu i.

Remarque 2.3.2 *:*

1. Si $CTR_l(e_i) > p_i$, alors la contribution de l'individu e_i est importante.
2. Si on a un grope d'individus, alors la contribution est égale à la somme des contributions des individus i et i' :

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

2.3.2 Interprétation des variables

On va essayer de représente quelque définition sur l'interprétation des résultats pour les variables.

Qualité de représentation du nuage des variables

Pour donner une signification à la composante principale c_l est de la relier aux variables initiales X_j en calculant les coefficients de corrélation linéaire $r(c_l, X_j)$, et en s'intéressant aux plus forts coefficients en valeur absolue.

On exprime la qualité de représentation d'une variable quantitative X_j sur le $l^{éme}$ axe factoriel, par le coefficient de corrélation linéaire $r(c_l, X_j)$ tel que :

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

Où :

λ_l : valeur propre associée à c_l.

u_{jl} : la $j^{éme}$ coordonnée de la facteur principale u_l.

Et on a aussi :

$$r(c_l, X_j) = r(c_l, Z_j) = \frac{c_l^t D_p Z_j}{\sqrt{\lambda_l}}.$$

Remarque 2.3.3 *:*

Chaque variable représentée par les corrélations d'une variable X_j avec un couple de composantes principales c_1 et c_2, est dans un cercle de corrélation de rayan 1.

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

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

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

Comme $\lambda_l = \sum_{i=1}^{n} p_i c_{il}^2$, on peut aussi définir la contribution par :

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

2.4 Représentation d'élement supplémentaire

Les éléments supplémentaires peuvent être des variables ou des individus.

2.4.1 Représentation des individus supplémentaire

Pour faire la représentation des individus supplémentaires sur le sous-espace de projection F_k, il suffit de calculer les coordonnées des individus dans le systéme des axes principaux.

On note par $\xi = \left(\xi_1, \xi_2, \ldots, \xi_p\right)^t \in \mathbb{R}^p$ un nouvel individu appelé individu supplémentaire, tel que :

$$\xi^t u_1, \xi^t u_2, \ldots, \xi^t u_k.$$

2.4.2 Représentation des variables supplémentaire

Pour faire la représentation des variables supplémentaires sur le sous-espace de projection F_k, il suffit de calculer les coordonnées des variables dans le systéme des axes principaux.

On note par $w = (w_1, w_2, \ldots, w_p)^t \in \mathbb{R}^p$ un nouvel variable appelé variable supplémentaire, tel que :

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

Chapitre 3

L'ACP sur R

Dans la pratique, on retient un nombre $k < p$ d'axes principaux, sur lesquels on va projeter notre nuage de points. On doit alors proposer une interprétation des nouveaux axes obtenus, ou de façon équivalente des composantes principales. On peut s'aider pour cela des outils définis dans les sections précédentes (contribution des individus et des variables dans la définition des axes). Plus précisément, on réalisera les étapes successives suivantes :

1. Centrage de la matrice de données.
2. Réduction de la matrice de données si nécessaire.
3. Calcul des valeurs propres de V, et choix du nombre d'axes à retenir en fonction du pourcentage d'inertie que l'on souhaite concerver. On peut également utiliser des régles "empirique" comme la régle de kaiser dans le cas d'une **ACP** sur variables réduites, qui consiste à ne garder que les axes pour lesqueles la valeur propre λ_j est plus grande que $1/p$, ou la régle du "coude" qui consiste à repérer un "coude" dans le graphe des valeurs propres.
4. Interprétation des nouvelles variables, à l'aide des cercles de corrélations attention, les variables doivent être proces du bord du cercle pour être bien représentées dans le plan factoriel considéré.

5. Complément pour l'interprétation des nouveaux axes à l'aide des individus et de leurs contributions à la fabrication des axes. Par exemple, on peut retenir les individus dont la contribution est supérieure à la contribution moyenne 1. Atention, si on souhaite comparer des individus entre eux, il faut d'abord s'assurer qu'ils sont bien représentés dans le plan factoriel considéré.

Effet taille. Lorsque les variables X_j sont corrélées positivement entre elles, la premiére composante principale définit ce que l'on appelle un efet taille. En efet, on sait qu'une matrice symétrique dont tous les termes sont positifs admet un premier vecteur propre dont touts les composantes sont de même signe (c'est le théorème de frobenius). La première composante principale est alors corrélée positivement avec toutes les variables. Si un tel efet se produit, on veillera à ne pas considérer cet axe dans l'interprétation des résultats, que l'on débutera avec le deuxième axe factoriel.

3.1 Etude de cas (packages R)

Plusieurs fonctions, de différents packages, sont disponibles dans le logiciel **R** pour le calcul de **l'ACP** :

. **prcomp()** et **princomp()** [fonction de **base**, package **stats**] .

. **PCA()** [package **FactoMineR**] .

. **dudi.pca()** [package **ade4**] .

. **epPCA()** [package **ExPosition**] .

Peu importe la fonction que vous décidez d'utiliser, vous pouvez facilement extraire et visualiser les résultats de **l'ACP** en utilisant les fonctions **R** fournies dans le package **Factoextra**. Ici, nous utiliserons les deux packages : **FactoMineR** (pour l'analyse) et **Factoextra** (pour la visualisation, des données, basée sur **ggplot2**).

FactoMineR est une package **R** dédié à l'analyse exploratoire multidimensionnelle

de données (à la Française). Il à été développé et il est maintenu par **François Husson, Julie Josse, Sébastien Lê** (d'Agrocampus Rennes), et **J. Mazet**.

3.1.1 Procesdures

install.packages(c("FactoMineR","Factoextra")).

library("FactoMineR").

library("Factoextra").

Nous utiliserons les jeusx de données de démonstration **decathlon2** du package **Factoextra** :

data(decathlon2).

Comme l'ulistre la figure 3.1, les données utilisées ici décrivent la performance des athlétes lors de deux événements sportifs (Decastara et OlympicG). Elles continnent 27 individus (athlétes) décrits par 13 variables.

name	100m	Long.jump	//	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58		63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4		60.15	301.5	2	8122	Decastar
Macey	10.89	7.47		58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74		55.39	278.05	5	8343	OlympicG
\\								
Zsivoczky	10.91	7.14		63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19		57.76	264.35	7	8237	OlympicG
Pogorelov	10.95	7.31		53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3		60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99		64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3		50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6		51.77	278.1	6	8030	Decastar
Nool	10.8	7.53		61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38		51.53	274.21	19	7926	OlympicG

Active individuals
Active variables
Supplementary **quantitative** variables
Supplementary **qualitative** variable
Supplementary individuals

FIG. 3.1 – Les données

Notez que selement certains de ces individus et variables seront utilisés pour effectuer l'analyse en composantes principales. Les coordonnées des individus et des variables

	X100m	Long.jump	Shot.put	High.jump	X400m	X110m.hurdle
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69
CLAY	10.76	7.40	14.26	1.86	49.37	14.05
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31

TAB. 3.1 – Les individus actifs et les variables actives pour l'ACP

restants serant prédites aprés **l'ACP**.

• **Individus actifs :** (en bleu clair, lignes 1 : 23) individus qui sont utilisés lors de l'analyse en composantes principales.

• **Individus suplémentaires :** (en bleu foncé, lignes 24 : 27) les coordonnées de ces individus seront prédites en utilisant l'information et les paramétres de **l'ACP** obtenue avec les individus/variables actifs.

• **Variables actives :** (en rose, colonnes 1 : 10) variables utilisées pour **l'ACP**.

• **Variables suplémentaires :** comme les individus suplémentaires, les coordonnées de ces variables seront également prédites. On distingue des :

1. Variables quantitatives suplémentaires (rouge) : les colonnes 11 et 12 correspondent respectivement au rang et aux points des athlétes.
2. Variables qualitatives suplémentaires (vert) : colonne 13 correspondant aux deux rencontres sportives (Jeux olympique de 2004 ou Décastar 2004). Il s'agit d'une variable catégorielle. Elle peut être utilisée pour colorer les individus par groupes.

Nous commençons par extraire les individus actifs et les variables actives pour **l'ACP** :

decathlon2.active<-decathlon2$[\mathbf{1:23}, \mathbf{1:10}]$.

head(decathlon2.active$[, \mathbf{1:6}], \mathbf{4})$**.**

3.1.2 Standardisation des données

Dans l'analyse en composantes principales, les variables sont souvent normalisées. Ceci est particuliérement recommandé lorsque les variables sont mesurées dans différentes unités (par exemple : kilogrammes, kilométres, centimétres, ...) ; sinon, le résultat de **l'ACP** obtenue sera fortement affecté. L'objectif est de rendre les variables comparables. Généralement, les variables sont normalisées de maniére à ce qu'elle aient au final, i) un écart type égal à un et, ii) une moyenne égale à zéro. Techniquement, l'approche consiste à transformer les données en soustrayant à chaque valeur une valeur de référence (la moyenne de la variable), et en la divisant par l'écart-type. A l'issue de cette transformation les données obtenues sont dites données centrées-réduites. **L'ACP** appliquée à ces données transformées est appelée **ACP normée**. La standardisation des données est une approche beaucoup utilisée dans le contexte de l'analyse des données d'expression de génes avant les analyses de type **PCA** et de **clustering**.

Lors de la normalisation des variables, les données peuvent être transformées comme suit :

$$\frac{X_i - mean(X)}{sd(X)}.$$

Où $mean(X)$ est la moyenne des valeurs de X, et $sd(X)$ est l'écart-type.

La fonction **scale()** peut être utilisée pour normaliser les données.

Notez que, par défaut, la fonction **PCA()** [dans **FactoMineR**], normalise les données automatiquement pendant **l'ACP** ; donc, vous n'avez pas besoin de faire cette transformation avant **l'ACP**.

Format simplifié :

PCA(x,scale.unit = TRUE,ncp = 5,graph = TRUE).

• **x** : jeu de données de type data frame. Les lignes sont des individus numériques.

##name	description
##1"$eig"	"eigenvalues"
##2"$var"	"results for the variables"
##3"varcoord"	"coord.for the variables"
##4"varcor"	"correlations variables-dimensions"
##5"varcos2"	"cos2 for the variables"
##6"varcontrib"	"contributions of the variables"
##7"$ind"	"results for the individuals"
##8"indcoord"	"coord.for the individuals"
##9"indcos2"	"cos2 for the individuals"
##10"indcontrib"	"contributions of the individuals"
##11"$call"	"summary statistics"
##12"$call$centre"	"mean of the variables"
##13"$call$ecart.type"	"standard errorofthe variables"
##14"$call$row.w"	"weights for the individuals"
##15"$call$col.w"	"weights for the variables"

TAB. 3.2 – Résultats de l'Analyse en Composantes Principales (ACP)

• **scale.unit** : une valeur logique. Si TRUE, les données sont standardisées/normalisées avant l'analyse.

• **ncp** : nombre de dimonsions conservées dans les résultats finaux.

• **graph** : une valeur logique. Si **TRUE** un graphique est affiché.

Calculer **l'ACP** sur les individus /variables actifs :

library("factoMineR").

res.pca<-PCA(decathlon2.avtive,graph = FALSE).

print(res.pca).

Le résultat de la fonction **PCA()** est une liste, contenant les éléments suivants :

##**Results for the Principal Component Analysis (PCA)**.

##The analysis was performed on 10 individuals, described by 1 variables.

##*The results are available in the following objects :

L'objet crée avec la fonction **PCA()** contient de nombreuses informations trouvées

dans de nombreuses listes et matrices différentes. Ces valeurs sont décrites dans la section suivante.

3.1.3 Visualisation et interprétation

Les fonctions suivantes, de **Factoextra**, seront utilisées :

• **get_eigenvalue(res.pca)** : Extraction des valeurs propres/variables des composantes principales.

• **fviz_eig(res.pca)** : Visualisation des valeurs propres.

• **get_pca_ind(res.pca), get_pca_var(res.pca)** : Extraction des résultats pour les individus et les variables, respectivement.

• **fviz_pca_ind(res.pca), fviz_pca_var(res.pca)** : Visualisez les résultats des individus et des variables, respectivement.

• **fviz_pca_biplot(res.pca) :** Création d'un biplot des individus et des variables.

Dans les sections suivantes, nous allons illustrer chacune de ces fonctions.

3.1.4 Valeurs propres/Variances

Comme décrit dans les sections précédentes, les valeurs propres (eigenvalues en anglais) mesurent la quantité de variance expliquée par chaque axe principal. Les valeurs propres sont grandes pour les premiers axes et petits pour les axes suivants. Autrement dit, les premers axes correspondent aux directions portant la quantité maximale de variation contenue dans le jeu de données. Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération. Les valeurs propres et la proportion de variances (i.e. information) retenues par les composantes principales peuvent être extraites à l'aide de la fonction **get_eigenvalue()** [**package Factoextra**].

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	4.124		
Dim.2	1.839		
Dim.3	1.239		
Dim.4	0.819		
Dim.5	0.702		
Dim.6	0.423		
Dim.7	0.303		
Dim.8	0.274		
Dim.9	0.155		
Dim.10	0.122		

TAB. 3.3 – Les valeurs propres

library("Factoextra").

eig.val <- get_eigenvalue(res.pca).

eig.val.

La somme de toutes les valeurs propres donne une variance totale de 10.

La proportion de variance expliquée par chaque valeur propre est donnée dans la deuxiéme colonne. Par exemple, 4.124 divisé par 10 est égal à 0.4124, ou, environ 41.24% de la variation est expliquée par cette premiére valeur propre. Le pourcentage cumulé expliqué est obtenu en ajoutant les proportions successives de variances expliquées. Par exemple, 41.242% plus 18.385% sont égaux à 59.627%, et ainsi de suite. Par conséquent, environ 59.627% de la variance totale est expliquée par les deux premiéres valeurs propres.

Les valeurs propres peuvent être utilisées pour déterminer le nombre d'axes principaux à concerver aprés **l'ACP** (kaiser 1961) :

• Une valeur propre > 1 indique que la composante principale concernée représente plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. Ceci est généralement utilisé comme seuil à partir du quel les composantes principales sont concervés. A noter que cela ne s'aplique que lorsque les

données sont normalisées.

• Vous pouvez également limiter le nombre d'axes à un nombre qui représente une certaine fraction de la variance totale. Par exemple, si vous êtes satisfaits avec 70% de la variance totale expliquée, utilisez le nombre d'axes pour y parvenir.

Malheuresment, il n'existe pas de méthode objective bien acceptée pour décider du nombre d'axes principaux qui suffisent. Cela dépendra du domaine d'application spécifique et du jeu de donées spécifiques. Dans la pratique, on a tendance à regarder les premiers axes principaux afin de trouver des profils intéressants dans les données.

Dans notre analyse, les trois premiéres composantes principales expliquent 72% de la variation. C'est un pourcentage acceptable.

Une autre méthode pourdéterminer le nombre de composantes principales est de regarder le graphique des valeurs propres (appelé scree plot). Le nombre d'axes est déterminer par le point, au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables (Jollife 2002, Peres-Neto, Jackson, and Somers (2005)).

Le graphique des valeurs peopres peut être généré à l'aide de la fonction **fviz_eig()** ou **fviz_screeplot()** [**packagr Factoextra**].

fviz_eig(res.pca,addlabels = TRUE,ylim = c(0, 50)).

Du graphique ci-dessus, nous pourrions vouloir nous arrêter à la cinquiéme copo-sante principale. 87% des informations (variances) contenues dans les données sont conservées par les cinq premiéres composantes principales.

3.1.5 Graphique des variables

Une méthode simple pour extraire les résultats, pour les variables, à partir de **l'ACP** est d'utiliser la fonction **get_pca_var()** [**package Factoextra**]. Cette fonction retourne une liste d'éléments contenant tous les résultats pour les variables actives

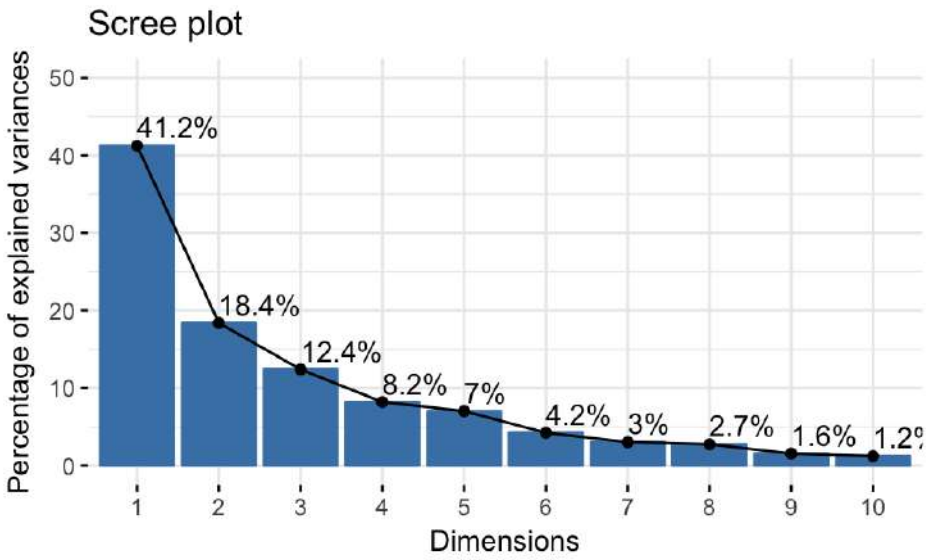

FIG. 3.2 – Le graphique des valeurs propres

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
X100m	17.54	1.751	7.340	0.138	5.39
Long.jump	15.29	4.290	2.930	1.625	7.75
Shot.put	13.06	0.397	21.620	2.014	8.82
High.jump	9.02	11.772	8.790	2.550	23.12

TAB. 3.4 – Les résultats pour les variables actives (coordonnées, corrélation entre variables et les axes, cosinus-carré et contributions)

(coordonnées, corrélation entre variables et les axes, cosinus-carré et contributions).

var<- get_pca_var(res.pca).

var.

Plus la valeur de la contribution est importante, plus la variable contribue à la composante principale en question.

La fonction **fviz_contrib()** [**package Factoextra**], peut être utilisée pour créer un bar plot de la contribution des variables. Si vous données contiennent de nombreuses variables, vous pouvez décider de ne montrer que les principales variables contributives. Le code **R** ci-dessous montre le top 10 des variables contribuant le plus aux

composantes principales :

#Contribution des variables à PC1.

fviz_contrib(res.pca, choice = "var", axes = 1, top = 10).

#Contribution des variables à PC2.

fviz_contrib(res.pca, choice = "var", axes = 2, top = 10).

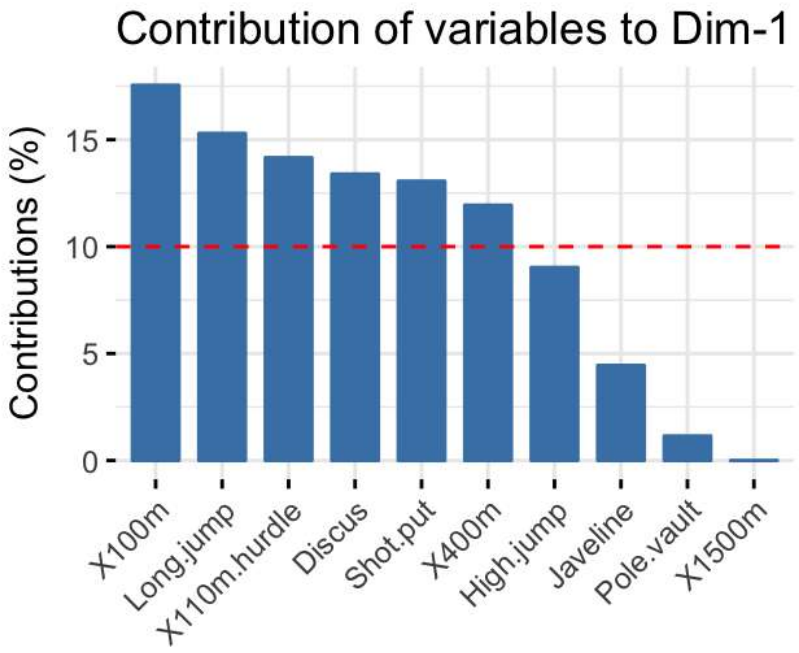

FIG. 3.3 – Contribution totale à PC1

La contribution totale à PC1 et PC2 est obtenue avec le code **R** suivant :

fviz_contrib(res.pca, choice = "var", axes = 1 : 2, top = 10).

La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moiyenne attendue. Si la contribution des variables était uniforme, la valeur attendue serait $1/length(variables) = 1/10 = 10\%$. Pour une composante donnée, une variable avec une contribution supérieure à ce seuil pourrait être considérée comme importante pour contribuer à la composante.

Notez que la contribution totale d'une variable donnée, pour expliquer la variance

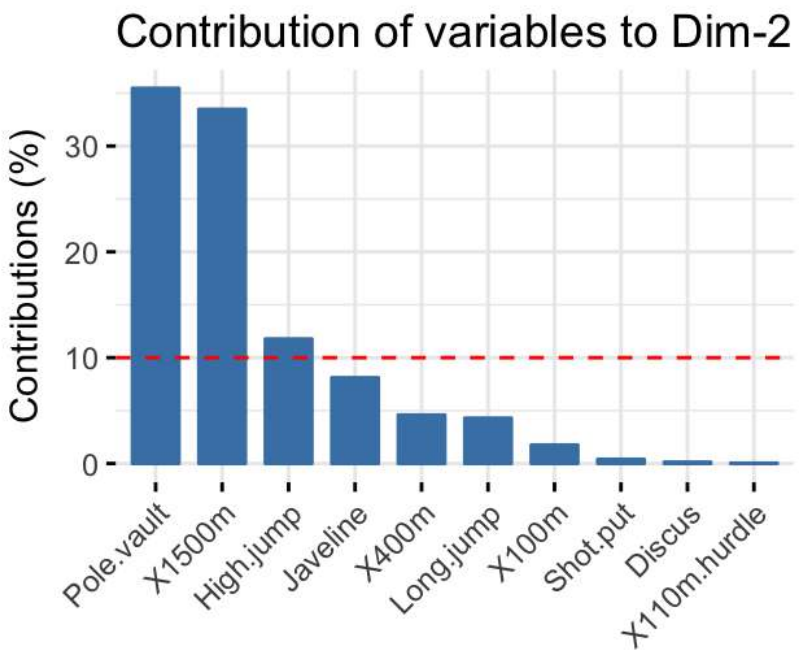

FIG. 3.4 – Contribution totale à PC2

retenue par deux composantes principales, disons PC1 et PC2, est calculée comme :

$$contrib = [(C1 * \text{Ei}\, g1) + (C2 * \text{Ei}\, g2)]\,/(\text{Ei}\, g1 + \text{Ei}\, g2).$$

Où

• C1 et C2 sont les contributions de la variable aux axes PC1 et PC2, respectivement.

• Eig1 et Eig2 sont les valeurs propres de PC1 et PC2, respectivement. Rappelons que les valeurs propres mesurent la quantité de variation retenue par chaque composante principale. Dans ce cas, la contribution moyenne attendue (seuil) est calculée comme suit :

Comme mentionné ci-dessus, si les contributions des 10 variables étaient uniformes, la contribution moyenne attendue pour une composante principale donnée serait $1/10 = 10\%$. La contribution moyenne attendue d'une variable pour PC1 et PC2

est :

$$[(10 * \text{Ei}\, g1) + (10 * \text{Ei}\, g2)]\,/(\text{Ei}\, g1 + \text{Ei}\, g2).$$

On peut voir que les variables - $X100m$, $Long.jump$ et $Pole.vault$ - contribuent le plus aux dimension 1 et 2.

Les variables les plus importantes (ou, contributives) peuvent être mise en évidence sur le graphe de corrélation comme suit :

fviz_pca_var(res.pca, col.var = "contrib", gradiant.cols = c("#00AFBB", "#E7B800","#FC4E07")).

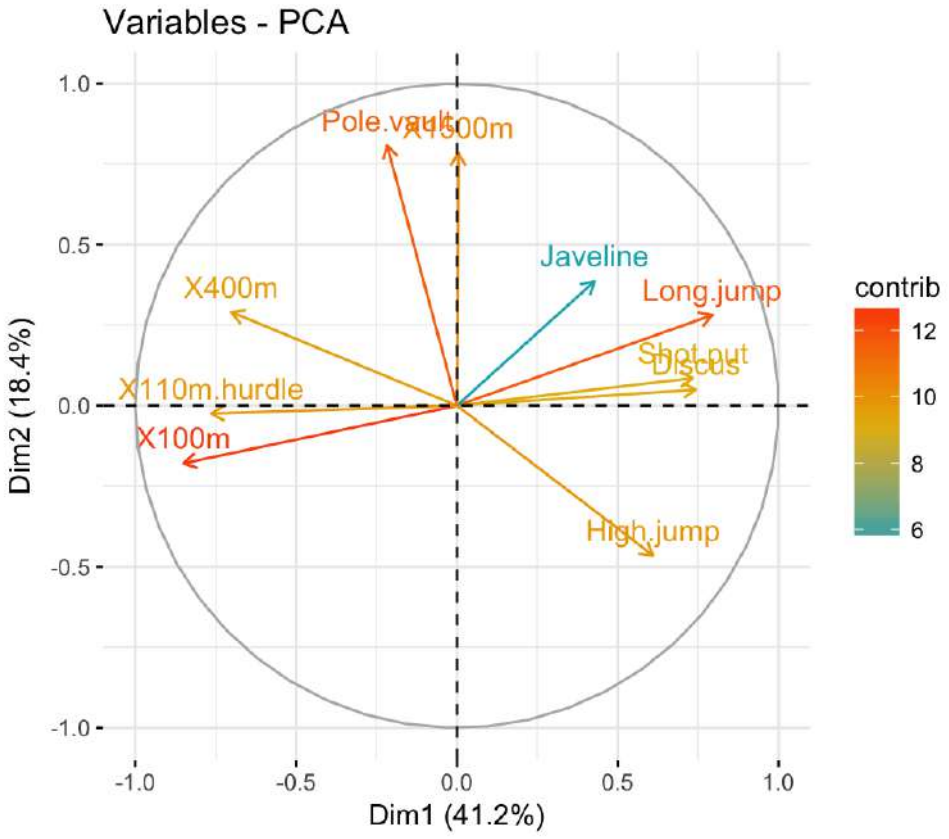

FIG. 3.5 – Graphique de corrélation des variables

Notez qu'il est aussi possible de modifier la transparence des variables en fonction de leurs contributions en utilisant l'option **alpha.var="contrib"**. Par exemple, tapez ceci :

	correlation	*p.value*
Long.jump	0.743	0.734 $6.72e-05$
Discus	0.428	$4.15e-02$
Shot.put	-0.702	$1.91e-04$
High.jump		
Javeline		
$X400m$		
$X110m.hurdle$		
$X100m$		

TAB. 3.5 – Description de la dimension 1

#Changez la transparence en fonction de contrib.

fviz_pca_var(res.pca, alpha.var= "contrib").

3.1.6 Description des dimensions

Dans les sections précédentes, nous avons décrit comment mettre en évidence les variables en fonction de leurs contributions aux composantes principales.

Notez également que la fonction **dimdesc()** [dans **FactoMineR**], pour dimension description (en anglais), peut être utilisée pour identifier les variables les plus significativement associées avec une composante principale donnée. Elle peut être utilisée comme suit :

res.desc<-dimdesc(res.pca,axes = c(1, 2),proba = 0.05).

#Description de la dimension 1.

res.desc$Dim.1.

#Description de la dimension 2.

res.desc$Dim.2.

Name	*Description*
1"$*coord*"	"Coordinates for the individuals"
2"$*cos*2"	"Cos2 for the individuals"
3"$*contrib*"	"contributions of the individuals"

TAB. 3.6 – Résultats de l'analyse en composantes principales pour les individus

3.1.7 Graphique des individus

Les résultats, pour les individus, peuvent être extraits à l'aide de la fonction **get_pca_ind()** [**package Factoextra**]. comme **get_pca_var()**, la fonction **get_pca_ind()** retourne une liste de matrices contenant tous les individus (coordonnées, corrélation entre individus et axes, cosinus-carré et contributions).

ind<-get_pca_ind(res.pca).

ind.

##Principal Component Analysis Results for individuals.

Pour accéder aux différents éléments, utilisez ceci :

#Coordonnées des individus.

head(ind$coord).

#Qualité des individus.

head(ind$cos2).

##ontributions des individus.

head(ind$contrib).

3.1.8 Graphique : qualité et contribution

La fonction **fviz_pca_ind()** est utilisée pour produire le graphique des individus.

Pour créer un graphique simple, tapez ceci :

fviz_pca_ind(res.pca).

Comme les variables, il est également possible de colorer les individus en fonction de leurs valeurs de **cos2** :

fviz_pca_ind(res.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FACA4E07"),repel =TRUE # Évite le chevauchement de texte).

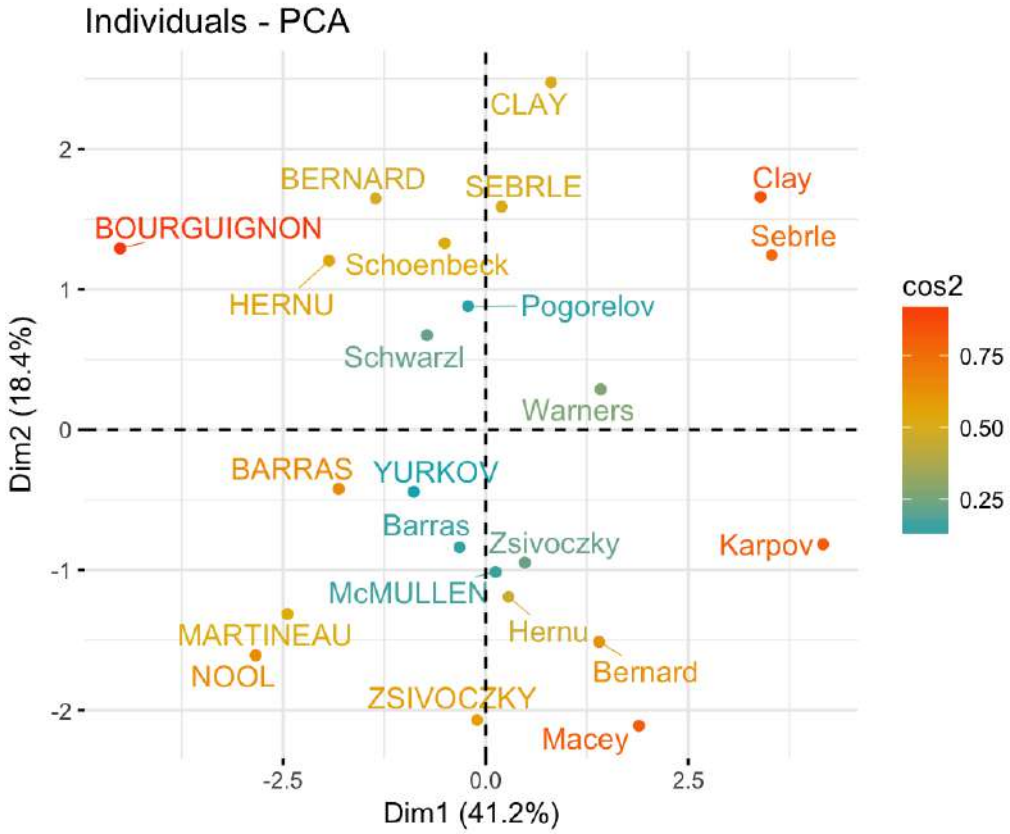

FIG. 3.6 – Graphique simple des individus

Notez que les individus qui sont similaires sont regropés sur le graphique.

Vous pouvez également modifier la taille des points en fonction du cos2 des individus correspondants :

fviz_pca_ind(res.pca, pointsize= "cos2",pointshape = 21, fill = "#E7B800",repel =TRUE # Évite le chevauchement de texte).

Pour créer un bar plot de la qualité de représentation (cos2) des individus, vous pou-

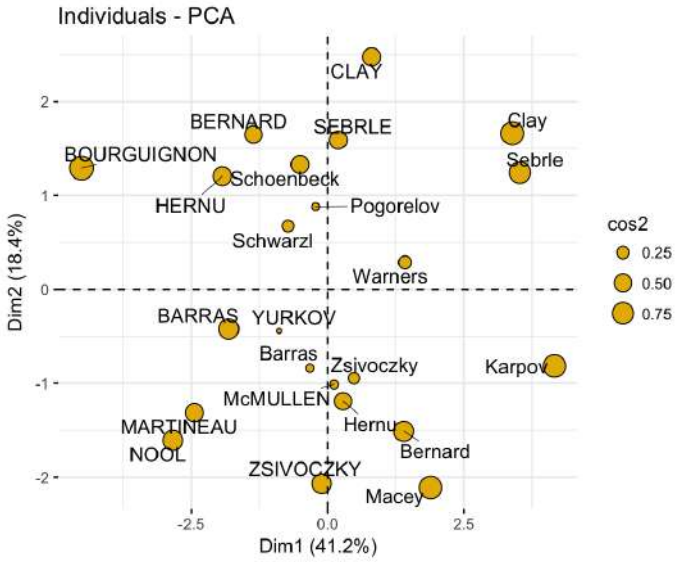

FIG. 3.7 – Graphique des individus

vez utiliser la fonction **fviz_cos2()** comme décrit précédemment pour les variables :

fviz_cos2(res.pca, choice = "ind").

Pour visualiser la contribution des individus aux deux premiéres composantes principales, tapez ceci :

#Contribution totale sur PC1 et PC2.

fviz_contrib(res.pca, choice = "ind", axes = 1 :2).

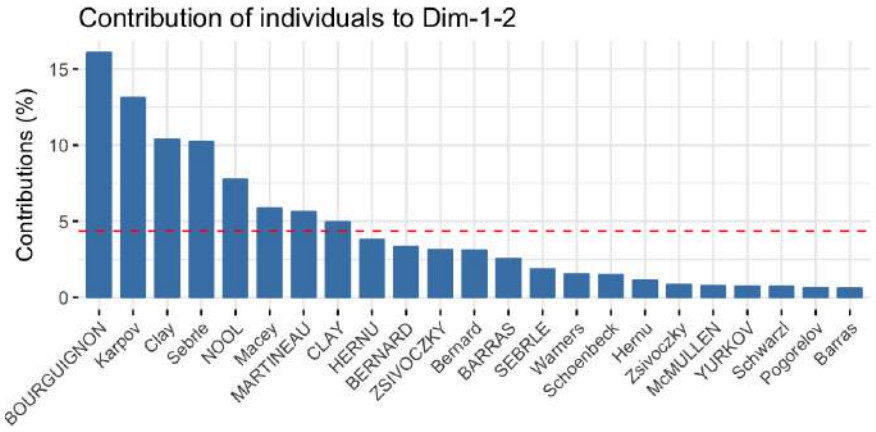

FIG. 3.8 – Contribution totale sur PC1 et PC2

Colorer en fonction d'une variable continue quelconque.

Conclusion

Dans ce mimoire, on a présenté l'Analyse en composantes principales ***(ACP)*** *qui est une des premiéres analyses factorielles, et certainement aujourd'hui l'une des plus employées. Elle est sent doute à la base de la compréhension avtuelle des analyses factorielles. Sons utilisation a cependant été plus tardive avec l'essor des capacités de calculs. 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'etude de la ressemblance entre les individus et la corrélation entre les variables, où ces informatins pertinentes sont résumées et visualisées tableau des donées.*

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

Bibliographie

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

[2] .Castell, F. (2004). Cours d'Analyse des données. Aix Marseille Université.

[3] .Duby, C., & Robin, S. (2006). Analyse en composantes principales. Institut National Agronomique, Paris-Grignon, 80.

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

[5] .Kassambara, A. (2017). http ://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/

[6] .Lasgouttes,J-M.(2014). Cours d'analyse des données.[http ://who. rocq. inria. fr/2014/cours-acp-2014-09-23. pdf].

[7] .Martin, A. (2004). L'analyse de données. Polycopie de cours ENSIETA-Réf : 1463.

[8] .Necer, A. (2020). cours de master 1 Analyse des données. Université Mohammed Khider, Biskra.

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

Annexe A : 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$^{\text{éme}}$ variable.
e_i	:	i$^{\text{éme}}$ individu.
D_p	:	Matrice des poids.
g	:	Centre de gravité.
Y	:	Tableau des données centrées.
$\overline{X_j}$	:	Moyenne de la variable X_j.
Z	:	Tableau des données centrées-réduites.
σ_j	:	Ecart-type de la variable X_j.
$\mathbb{R}^n$	:	Espace des nombres réels de dimension n.
V	:	Matrice de variance de tableau X.
R	:	Matrice de corrélations de tableau X.
$cov\,(.,.)$	:	Covariance.

$var\,(.,.)$	:	Variance
$d\,(e_i, e_{i'})$	:	Distance entre e_i et $e_{i'}$.
M	:	Métrique.
I_g	:	Inertie totale.
$\langle .,. \rangle$	:	Produit scalaire.
$cor\,(.,.)$	:	Corrélation.
Tr	:	Traced'une matrice.
F_k	:	Sous espace de dimension k.
P	:	Matrice de projection.
λ	:	Valeur propre.
f_i	:	Projection de l'individu e_i.
X_{proj}	:	Tableau de nuage projeté.
V_{proj}	:	Matrice de variance de nuage projeté.
I_{proj}	:	Inertie totale de nuage projeté.
g_{proj}	:	Centre de gravité de nuage projeté.

a_j	:	Axe principale.
u_j	:	Fateur principale.
c_j	:	Composante principale.
$QLT\left(F_k\right)$	:	Qualité sur F_k.
$QLT_l\left(e_i\right)$	:	Qualité de e_i sur l'axe l.
$QLT_{l,l'}\left(e_i\right)$	:	Qualité de e_i sur plan $\left(l, l'\right)$.
$CTR_l\left(e_i\right)$	:	Contribution sur l'axe l de e_i.
$CTR_l\left(X_j\right)$	:	Contribution de X_j sur l'axe l.
$CTR_l\left(e_i, e_{i'}\right)$	:	Contribution sur l'axe l de couple $\left(e_i, e_{i'}\right)$.
ξ	:	Individu suplimentaire.
w	:	Variable suplimentaire.
$M_{\mathbb{R}}\left(n, p\right)$	:	L'encemble des matrices de type $\left(n, p\right)$ à coefficients dans $\mathbb{R}$.
$proj$	:	projection.
$i.e$	:	C'est-à-dire.

Résumé :

L'objectif de ce travail est d'étudier l'Analyse en composantes principales (ACP), qui est outil extrêmement puissant de synthèse de l'information, très utile lorsque l'on est en présence d'une somme importante de données quantitatives à traiter. Il permet également de voire 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, et obtenir une représentation simple du nuage des données dans un espace de dimension faible plus proche de la réalité. Nous avons appliqué cette méthode en utilisant le logiciel R.

Abstract:

The objective of this work is to study principal component analysis (PCA), which is an extremely powerful tool for synthesizing information, very useful when we are in the presence of a large amount of quantitative data to treat. It also makes it possible to see the existing relations between individuals by evaluating their similarities, as well as the relations between variables by evaluating their links, and to obtain a simple representation of the data cloud in a space of low dimension more close to reality. We applied this method using R.

الملخص:

الهدف من هذا العمل هو دراسة تحليل المكونات الرئيسية، ؤ هي أداة قوية للغاية لتجيع المعلومات، ومفيدة عند وجود كمية كبيرة من البيانات الكمية التي يجب معالجتها. كما أنه يجعل من الممكن رؤية العلاقات القائمة بين الأفراد من خلال تقييم أوجه التشابه بينهم، وكذلك العلاقات بين المتغيرات من خلال تقييم روابطهم، والحصول على تمثيل بسيط لسحابة البيانات في مساحة ذات بعد منخفض أقرب إلى الواقع. طبقنا هذه الطريقة باستخدام برنامج الأر.