République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes

Départementde Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathématiques**"

Option : **Probabilités et statistique**

Par HAMIDAT Nassima

Titre :

Analyse en Composantes Principales

Devant le Jury :

Dr.	CHINE Amel	UMKB	Président
Dr.	BENELMIR Imane	UMKB	Encadreur
Dr.	BENBRAIKA Ghozlane	UMKB	Examinatrice

Soutenu Publiquement le 03/06/2025

Dédicace

À Allah, source de toute force, pour Sa guidance à chaque étape.

À ma chère **mère Meryem**,

ton amour et tes prières sont le pilier de ma réussite.

À mes **frères** et **sœurs**,

merci pour votre soutien et votre confiance.

À mes **ami(e)s**,

merci d'avoir illuminé ce chemin par votre présence.

Remerciements

Merci à Allah de m'avoir donné la santé, la patience et le courage pendant de longues

années d'étude afin de finir mes études et arriver à cet humble travail.

Mes remerciements à mon Encadreur, **Dr. BENELMIR Imane** pour m'avoir aidée, conseillée et encadrée cette année. Je suis très heureuse d'avoir travaillé avec elle.

Mes remerciements aux membres de Jury : **Dr. CHINE Amel (**Président) et **Dr. BENBRAIKA Ghozlane (**Examinatrice**),** pour avoir accepté d'étudier et d'évaluer ce travail.

Merci à tous mes professeurs et mes amis proches.

Merci à tous.

Notations et symbols

X	: Tableau des données.
n	: Nombre des individus.
p	: Nombre des variables.
x_j	: j-éme variable.
e_i	: i-éme individu.
$\overline{x}_j$	: Moyenne de variable x_j.
$\mathbb{R}^n$	: Espace des nombres réels de dimension p.
$\mathbb{R}^p$	: Espace des nombres réels de dimension n.
D	: Matrice de poids.
p_i	: Poids.
I_n	: Matrice d'identité de taille n.
1_n	: Vecteur unitaire d'identité de taille n.
g	: Centre de gravité.
Y	: Tableau des données centrées.
$\mathbb{Z}$	: Tableau des données centrées reduites.
σ_j	: L'écart-type.
$d(e_i, e_{i'})$	: Distance entre e_i et $e_{i'}$.
cov	: Covariance.
var	: Variance.

V	: Matrice de variance-covariance.
R	: Matrice de corrélation.
$r_{jj'}$	: Coeffcient de corrélation entre x_j et $x_{j'}$.
M	: Métrique.
I_g	: Inertie totale.
tr	: Trace d'une matrice.
$\langle ., . \rangle$	: Produit scalaire.
$i.e$	: C'est-à-dire.
$\mathcal{M}_{(n,p)}$	: L'encemble des matrices de type(n ;p)
F_k	: Sous-espacedeprojectiondedimension k.
P	: Operateur de projection.
f_i	: Projection de l'individu e_i.
X_{proj}	: Tableau de donnée de nuage projeté.
g_{proj}	: Centre de gravité projeté.
V_{proj}	: Matrice de variance-covariance du nuage projeté.
I_{proj}	**:** L'inertie du nuage projeté.
Δ	: Sous-espace vectorieil.
a_j	: Axes principaux.
u_j	: Facteurs principaux.
c_j	: Composantes principales.
λ_j	: Valeur propre.
QLT	: Qualité.
CTR	: Contribution.

Table des matières

Table des figures

Liste des tableaux

Introduction

On appelle statistique descriptive multidimensionnelle, l'ensemble des méthodes de la statistique descriptive qui traitent plusieurs variables à la fois.

Les méthodes factorielles sont les plus utilisées en statistique descriptive multidimensionnelle, car elles permettent de résumer et visualiser des données comportant plusieurs variables.

Les domaines d'application de ces méthodes sont nombreux et variés, couvrant des secteurs tels que la santé, la démographie, le marketing, l'économie et la psychologie.

Il existe plusieurs méthodes factorielles qui permettent de traiter diverses structures de données :

L'analyse en composantes principales pour un tableau de variables quantitatives, l'analyse factorielle des correspondances pour les tables de contingence, l'analyse factorielle multiple pour les variables qualitatives et l'analyse discriminante pour la prise en compte d'une partition des individus en groupe.

L'origine de ces méthodes remonte à l'année **1901** par [10]. Elles ont été particulièrement développées en France dans les années **1960**, notamment par [2], qui a largement exploité les aspects géométriques et les représentations graphiques.

Ainsi, à partir d'un tableau rectangulaire de données contenant les observations de p variables quantitatives sur n individus, il est possible d'obtenir des représentations géométriques de ces individus et de ces variables dans un sous-espace de dimension réduite grâce à l'Analyse en Composantes Principales notée ACP.

Le but de mon travail est de mettre en évidence le rôle de l'ACP dans la pratique.

Ce travail est organisé en deux parties théorique et pratique. La première partie se compose en deux chapitres.

- Le premier chapitre est dédié à l'introduction des notions fondamentales utilisées dans l'ACP à savoir, la matrice des poids, le centre de gravité, le nuage des individus et des variables, la matrice de covariance et de corrélation ainsi que la matrice d'inertie. Ces éléments constituent la base mathématique de la méthode.
- Le deuxième chapitre est consacré au principe de cette méthode, la représentation graphique ainsi que l'Interprétation des résultats obtenus.

La deuxième partie est consacrée à l'application de cette méthode à l'aide du l'ogiciel R qui est implimenté d'un exemple d'étude des différentes caractéristiques de l'approche de l'ACP.

Ce modeste travail s'achève avec une conclusion générale.

Chapitre 1

Généralité

Ce chapitre est une introduction à l'analyse en Composantes Principales (ACP). Les connaissances nécessaires pour aborder ce chapitre sont les données et leurs caractéristiques comme le tableau des données, puis on définit les individus, les variables, la matrice des poids, le centre de gravité...etc.

1.1 Données multivariées

Avant de débuter un travail, il est essentiel de passer par les étapes préliminaires, telles que la création du tableau des données contenant les individus, les variables et divers autres éléments.

1.1.1 Tableau des données

Les valeurs observées de p variables pour n individus sont rassemblées dans un tableau rectangulaire X à n lignes et p colonnes définis ci-dessous [11]

$$X = \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ & \ddots & & & \\ x_{i1} & & x_{ij} & & x_{ip} \\ & & & \ddots & \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \in \mathcal{M}_{\mathbb{R}}(n,p). \tag{1.1}$$

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

1.1.2 Variables et individus

Les colonnes de la matrice X représentent les variables et les lignes représentent les individus.

Définition 1.1.1 (Variables) *La $j^{ème}$ variable notée x_j est un vecteur de dimension n [1], tel que*

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

Définition 1.1.2 (Individu) *Le $i^{ème}$ individu noté e_i est un vecteur à p composantes réelles [1], tel que*

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

1.1.3 Types des variables statistiques

Les variables sont classées en deux groupes quantitatives et qualitatives.

- **Variables quantitatives :** elles sont caractérisées par des valeurs numériques tel que le poids ou la taille. [7]
- **Variables qualitatives** : elles sont caractérisées par des valeurs non numériques tel que le sexe, la couleur des yeux, ou bien par des valeurs numériques réparties en classes [7] tel que la classe d'âge entre $[0, 20[$ ou $]20, 40]$ ans.

1.1.4 Matrice des poids

On suppose que chaque individu est muni d'un poids p_i tel que $p_i \geq 0$ et $\sum_{i=1}^{n} p_i = 1$. Dans certains cas, il est utile de travailler avec des poids p_i différents (données regroupées$\cdots$).

Ces poids sont regroupés dans une matrice diagonale notée D de taille n, définie comme suit [12]

$$D = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & p_n \end{bmatrix}$$

Dans le cas le plus usuel les poids sont égales à $\frac{1}{n}$, dans ce cas la matrice diagonale devient égale à

$$\boxed{D = \tfrac{1}{n} I_n.} \tag{1.2}$$

où I_n est la matrice d'identité de taille n.

Preuve. Puisque les poids sont égaux i.e $p_1 = p_2 = \cdots = p_n = p$ et $\sum_{i=1}^{n} p_i = 1$ alors

$$\begin{aligned}\sum_{i=1}^{n} p_i &= p_1 + \cdots + p_n \\ &= p + \cdots + p \\ &= np \\ &= 1.\end{aligned}$$

par consequent $p = \dfrac{1}{n}$.

Donc

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

1.1.5 Centre de gravité

Le centre de gravité noté g est un vecteur des moyennes arithmétiques de chaque variable. Il est défini comme suit [9]

$$\boxed{g = (\overline{x}_1, \overline{x}_{2,}..., \overline{x}_p)^t \in \mathbb{R}^p \text{ où } \overline{x}_j = \sum_{i=1}^{n} p_i x_{ij}, \text{ pour } j = \overline{1, p}.}$$

On peut aussi l'écrire sous forme matricielle en fonction de X et de D comme suit (1.1,1.2)

$$\boxed{g = X^t D 1_n.} \tag{1.3}$$

où 1_n désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1.

Preuve.

On a :

$$\begin{aligned} X^t D 1_n &= \begin{bmatrix} x_{11} & \dots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} p_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_n \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{n} p_i x_{i1} \\ \vdots \\ \sum_{i=1}^{n} p_i x_{ip} \end{bmatrix} \\ &= \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix} = g. \end{aligned}$$

■

1.1.6 Standardisation du tableau

Dans l'analyse en composantes principales notée ACP, il existe deux types de transformations utilisées sur les données initiales.

Tableau centré

Le centrage des données nous permet de ramener toutes les colonnes de X à la même origine zéro (une moyenne égale à zéro) dans une matrice notée par Y de terme général. [2]

$$\boxed{y_{ij} = x_{ij} - \overline{x}_j.}$$

La forme matricielle est :

$$\boxed{Y = X - 1_n g^t.} \tag{1.4}$$

Preuve.

$$
\begin{aligned}
X - 1_n g^t &= \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} (\overline{x}_1, \overline{x}_2, ..., \overline{x}_p) \\
&= \begin{bmatrix} x_{11} - \overline{x}_1 & \dots & x_{1p} - \overline{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \overline{x}_1 & \dots & x_{np} - \overline{x}_p \end{bmatrix} \\
&= \begin{bmatrix} y_{11} & \dots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \dots & y_{np} \end{bmatrix} = Y.
\end{aligned}
$$

■

Tableau centré réduit

À partir du tableau centré Y, on obtient un tableau centré et réduit Z, en divisant les coordonnées de chaque colonne par l'écart-type correspondant.

On note par σ_j^2 la variance empirique de la variable X_j définit par

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{X_j})^2, j = \overline{1, p}.$$

Les données centrées et réduites sont

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

On définit la matrice des poids notée $D_{1/\sigma}$ comme suit

$$D_{1/\sigma} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_p \end{bmatrix}.$$

On peut écrire

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

Preuve.

$$\begin{aligned} YD_{1/\sigma} &= \begin{bmatrix} y_{11} & ... & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & ... & y_{np} \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_p \end{bmatrix} \\ &= \begin{bmatrix} y_{11/}\sigma_1 & \cdots & y_{1p}/\sigma_p \\ \vdots & \ddots & \vdots \\ y_{n1}/\sigma_1 & \cdots & y_{np}/\sigma_p \end{bmatrix} \\ &= \begin{bmatrix} (x_{11} - \overline{X}_1)/\sigma_1 & \cdots & (x_{1p} - \overline{X}_p)/\sigma_p \\ \vdots & \ddots & \vdots \\ (x_{n1} - \overline{X}_1)/\sigma_1 & \cdots & (x_{np} - \overline{X}_p)/\sigma_p \end{bmatrix} \\ &= \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix} = Z. \end{aligned}$$

■

1.1.7 Matrice de variance-covariance

C'est une matrice carrée symétrique d'ordre p notée V définie comme suit

$$V = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{bmatrix},$$

où $\sigma_{jj'}$ est la covariance entre les variables x_j et $x_{j'}$.

$\sigma_{jj'}$ peut être calculé comme suit

$\sigma_{jj'} = cov(x_j, x_{j'}) = \sum_{i=1}^{n} p_i \left(x_{ij} - \overline{x}_j\right)\left(x_{ij'} - \overline{x}_{j'}\right), j, j' = \overline{1, p}.$

Et σ_j^2 est la variance de la variabe x_j tel que

$\sigma_{jj} = \sigma_j^2 = var\left(x_j\right) = \sum_{i=1}^{n} p_i \left(x_{ij} - \overline{x}_j\right)^2, j = \overline{1, p}.$

La forme matricielle[4] :

$$\boxed{V = Y^t DY = X^t DX - gg^t.} \tag{1.5}$$

Preuve. On a : $Y = X - 1_n g^t$, alors

$$\begin{aligned} V &= Y^t DY \\ &= (X - 1_n g^t)^t D(X - 1_n g^t) \\ &= X^t DX - X^t D1_n g^t - g1_n^t DX + g1_n^t D1_n g^t \quad \blacksquare \\ &= X^t DX - gg^t - gg^t + gg^t, \text{ car } 1_n^t D1_n = 1. \\ &= X^t DX - gg^t. \end{aligned}$$

Remarque 1.1.1 *La matrice V admet p valeurs propres.*

1.1.8 Matrice de corrélation

La matrice de corrélation est une matrice symétrique qui regroupe tous les coefficients de corrélation entre les p variables prises deux à deux, elle est notée par R [4], avec

$$R = \begin{bmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{bmatrix}.$$

où $r_{jj'} = \dfrac{\sigma_{jj'}}{\sigma_j \sigma_{j'}}$.

La forme matricielle :

$$\boxed{R = D_{1/\sigma} V D_{1/\sigma} = Z^t D Z.}$$

Preuve.

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

$$= \begin{bmatrix} 1 & \cdots & \frac{\sigma_{1p}}{\sigma_1 \sigma_p} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sigma_p \sigma_1} & \cdots & 1 \end{bmatrix} = R. \quad \blacksquare$$

Remarque 1.1.2

- Comme il y a p variables, cela nous conduit donc à calculer $p(p-1)/2$ corrélations.
- R est la matrice de variance-covariance des données centrées réduites, elle résume la structure des dépendances linéaires entre les p variables prises deux à deux.

1.2 Nuage des individus

Chaque individu est représenté par un point défini par p coordonnées dans un espace vectoriel $\mathbb{R}^p$ appelé l'espace des individus. L'ensemble des n individus forme ainsi un "nuage" de points dans $\mathbb{R}^p$.[12]

1.2.1 Ressemblance entre deux individus

Deux individus se ressemblant (sont proches), s'ils possédent des valeurs proches pour l'ensemble des variables. En ACP, on utilise le plus souvent la distance euclidienne, la distance entre deux individus e_i et $e_{i'}$ est égale à

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

1.2.2 Métrique

En générale, la distance utilisée entre deux individus e_i et $e_{i'}$ est définie par la forme quadratique suivante [12]

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

où M est une matrice symétrique définie positive de type $(p \times p)$.

Les métriques les plus utilisées en ACP sont la matrice identité d'ordre p notée I_p et la matrice diagonale des inverses des variances D_{1/σ^2}.

Il est a noté que pour le tableau centré Y, on utilise la métrique $M = D_{1/\sigma^2}$. Aussi pour le tableau centré et réduit Z, on utilise la métrique $M = I_p$. [9]

1.2.3 Inertie

On appelle inertie totale du nuage des points, la moyenne pondérée des carrées des distances des points $(e_i)_{1\leq i\leq n}$ au centre de gravité g.[11]

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

L'inertie dans un point a quelconque est définie par

$$I_a = \sum_{i=1}^{n} p_i (e_i - a)^t M (e_i - a).$$

On a la relation de **Huygens** :

$$\boxed{I_a = I_g + (g - a)^t M (g - a) = I_g + \left\| g - a \right\|^2 .}$$

Si $g = 0$ i.e.la matrice X est centrée, alors

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

Proposition 1.2.1 ***1.*** *$I_g = Tr(MV) = Tr(VM)$. (Tr est la trace d'une matrice).*

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

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

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

$$I_g = p.$$

1.3 Nuage des variables

Chaque variable. x_j est en fait une liste de n valeurs numériques. On la considère comme un vecteur d'un espace $\mathbb{R}^n$ à n dimensions appelé espace des variables. [11]

1.3.1 Liaison entre deux variables

Le coeffcient $r_{jj'}$ de corrélation mesure la liaison entre deux variables x_j et $x_{j'}$, qui prend ses valeurs dans l'intervalle $[-1, 1]$. Il est définie comme suit

$$\boxed{r\left(x_j, x_{j'}\right) = \frac{cov\left(x_j, x_{j'}\right)}{\sqrt{var\left(x_j\right) var\left(x_{j'}\right)}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \overline{x}_j}{\sigma_j}\right)\left(\frac{x_{ij'} - \overline{x}_{j'}}{\sigma_{j'}}\right); j, j' = \overline{1, p},}$$

avec $cov\left(x_j, x_{j'}\right) = \left(x_{ij} - \overline{x}_j\right)\left(x_{ij'} - \overline{x}_{j'}\right)$.

1.3.2 Métrique des variables

Pour étudier la proximité des variables entre elles, il faut munir cet espace d'une métrique i.e. trouver une matrice d'ordre n définie positive symétrique. 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 pour les raisons suivantes : [11]

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

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

   Il n'est autre que la matrice de covariance $\sigma_{jj'}$ car les caractères sont centrés.

2. La norme d'un caractère x_j est alors

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

3. L'angle $\theta_{jj'}$ entre deux variables centrées est donné

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

Remarque 1.3.1 *Dans l'espace des individus on s'intéresse aux distances entre points, et dans l'espace des variables on s'intéresse à l'angle entre les vecteurs.*

Chapitre 2

Analyse en composantes principales

L'ACP est une méthode d'analyse des données qui consiste à résumer les données statistiques (qui peuvent être trop nombreuses au départ) et à les présenter sous la forme de graphes afin de faciliter l'exploitation de leur structure. L'ACP traite les tableaux croisant des individus et des variables quantitatives. [3]

2.1 Principe de l'ACP

Le principe de l'ACP est d'obtenir une représentation approchée du nuage des individus dans un espace de dimension faible. Ceci se fait par projection des individus sur un sous-espace de $\mathbb{R}^p$ choisi de telle sorte que les distances entre les points projetés ressemblent le plus possible aux distances entre les points initiaux. En d'autres termes, elle revient à déformer le moins possible les distances en projection.[9]

2.1.1 Projection des individus sur un sous-espace

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

Nuage projeté

On note par F_k le sous-espace de projection avec $k < p$. Puisque l'opération de projection a pour effet de diminuer la distance, alors F_k doit être telle que la moyenne des carrés des distances entre les points projetés soit la plus grande possible, i.e. que l'inertie du nuage projeté sur F_k doit être maximale.

Soit P la matrice (l'operateur) de projection M-orthogonal sur l'espace F_k. On a

- $P^2 = P$ où P est idempotente.
- $MP = P^tM$ où P est $M-$symétrique.

On note par f_i la projection d'un individu e_i sur F_k telque

$$f_i = Pe_i; i = \overline{1, n}.$$

Le nuage projeté associé au tableau sera donnée par

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

Proposition 2.1.1 ***1.*** *Le centre de gravité projeté est :* $g_{proj} = Pg$.

2. *La matrice de variance-covariance du nuage projeté est :* $V_{proj} = PVP^t$.

3. *L'inertie du nuage projeté est :* $I_{proj} = tr(VMP)$.

Preuve.

1. *Centre de gravité projeté :*

$$\begin{aligned} g_{proj} &= X_{proj}^t D 1_n \\ &= \left(XP^t\right)^t D 1_n \\ &= P\left(X^t D 1_n\right) \\ &= Pg. \end{aligned}$$

2. *Matrice de variance-covariance du nuage projeté :*

$$\begin{aligned} V_{proj} &= X_{proj}^t D X_{proj} - g_{proj} g_{proj}^t \\ &= P X^t D X P^t - P g g^t P^t \\ &= P\left(X^t D X - P g g^t\right) P^t \\ &= P V P^t. \end{aligned}$$

3.*L'inertie du nuage projeté :*

$$\begin{aligned} I_{proj} &= Tr(V_{proj} M) \\ &= Tr\left(P V P^t M\right) \\ &= Tr\left(P V M P\right), \ P \ est \ M - symétrique \quad \blacksquare \\ &= Tr\left(V M P^2\right) \\ &= Tr\left(V M P\right), \ P \ est \ idempotente. \end{aligned}$$

Construction du 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 $Tr(VMP)$. 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 maximale,...ect. La somme directe de ces sous

espaces de dimension 1 est F_k [11] tel que

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

On peut alors dire que l'inertie maximale du sous espace F_k est

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

Construction de la première droite 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 $a_1 \in \mathbb{R}^p$ un vecteur directeur de Δ_1. L'opérateur de projection $M-$orthogonale sur Δ_1 [9] est

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

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

$$\begin{aligned} I_{\Delta_1} &= Tr\left(VMP_1\right) \\ &= Tr\left(VM\frac{a_1 a_1^t M}{a_1^t M a_1}\right) \\ &= \frac{Tr(VMa_1 a_1^t M)}{a_1^t M a_1} \\ &= \frac{Tr(a_1^t MVMa_1)}{a_1^t M a_1}, \text{car } tr(AB) = tr(BA) \\ &= \frac{a_1^t MVMa_1}{a_1^t M a_1}. \end{aligned}$$

Donc $I_{\Delta_1} = \dfrac{a_1^t MVMa_1}{a_1^t Ma_1}$, tel que $a_1^t MVMa_1 \in \mathbb{R}$.

Pour obtenir le maximum de $\dfrac{a_1^t MVMa_1}{a_1^t Ma_1}$, il suffit d'annuler la dérivée de cette expression par rapport à a_1. En appliquant la règle de dérivation d'une forme quadratique par rapport à un vecteur, on obtient

$$VMa_1 = \frac{a_1^t MVMa_1}{a_1^t Ma_1} a_1.$$

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

$$VMa_1 = \lambda a_1.$$

Remarque 2.1.1 *Le a_1 est un vecteur propre de VM avec M matrice régulière et λ est la plus grande valeur propre associée à la matrice VM.*

Construction des autres droites

De la même manière on déterminera le deuxième axe Δ_2 passant par g et M-orthogonale à Δ_1 et maximisant l'inertie du nuage projeté sur Δ_2. La droite Δ_2 va correspondre au vecteur propre de VM associé à la deuxième plus grande valeur propre. Ainsi de suite, on obtient toutes les droites permettant de construire le sous-espace F_k. La M-orthogonalité des droites $(\Delta_1, \Delta_2, ..., \Delta_k)$ entre elles est garanties par le fait que la matrice VM est M-symétrique et donc, elle a des vecteurs propres M-orthogonaux deux à deux.[11]

Théorème 2.1.1 *Le sous-espace F_k de dimension k est engendré par les k vecteur propres de VM associés aux k plus grandes valeurs propres.*

Remarque 2.1.2 *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 Éléments principaux

2.2.1 Axes principaux

On appelle axes principaux notés a_j les vecteurs propres de VM associés à la valeur propre λ_j, $M-$normés à 1. Ils sont aussi $V^{-1}-$orthogonaux et $M-$orthonormés [10] i.e

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

2.2.2 Facteurs principaux

On appelle facteurs principaux notés u_j les vecteurs propres de MV associés à la valeurs propre λ_j, $M^{-1}-$normés à 1. Ils sont aussi $V-$orthogonaux et $M^{-1}-$orthonormés [12] i.e

$$\boxed{\begin{cases} MVu_j = \lambda_j u_j; j = \overline{1,p} \\ \|u_j\|_{M^{-1}}^2 = 1. \end{cases}} \tag{2.1}$$

Sachant que

$$\boxed{u_j = Ma_j.} \tag{2.2}$$

2.2.3 Composantes principales

Les composantes principales sont les variables c_j définies par les facteurs principaux u_j comme suit

$$\boxed{c_j = XMa_j = Xu_j.} \tag{2.3}$$

Les composantes principales c_j est un vecteur renfermant les coordonnées des

projections $M-$orthogonales des individus sur l'axe défini par a_j.

$$c_{ij} = \langle e_i, a_j \rangle_M .$$

Propriétés des composantes principales

1. Les c_j ne sont pas corrélées deux à deux donc $cov(c_j, c_{j'}) = 0$ avec $j \neq j'$.
2. La variance d'une composante principale c_j est $var(c_j) = \lambda_j$.
3. Les composantes principales c_j sont des combinaisons linéaires des variables initiales. De plus, elles sont les vecteurs propres de la matrice XMX^tD. Donc $XMX^tDc_j = \lambda_j c_j$.

Preuve.

1.
$$\begin{aligned} cov(c_j, c_{j'}) &= c_j^t D c_{j'} - g_{c_j} g_{c_{j'}}^t \\ &= u_j^t X^t D X u_{j'} - c_j^t D 1_n 1_n^t D c_{j'} \\ &= u_j^t X^t D X u_{j'} - u_j^t X^t D 1_n 1_n^t D X u_{j'} \\ &= u_j^t \left(X^t D X - g g^t \right) u_{j'} \\ &= u_j^t V u_j, \text{ d'aprés 1.5} \\ &= \langle u_j, u_{j'} \rangle_V \\ &= 0, \text{ car } u_j \text{ et } u_{j'} \text{ sont orthogonaux.} \end{aligned}$$

2. D'aprés 2.1, on a $Vu_j = \lambda_j M^{-1} u_j$ et puisque dans la démonstration précédente on a trouvé que $cov(c_j, c_{j'}) = u_j^t V u_j$, alors la $var(c_j) = \lambda_j u_j^t M^{-1} u_j = \lambda_j$.

3. D'aprés 2.1et 1.5, on a $MX^tDXu_j = \lambda_j u_j$.

En multipliant à gauche par X et en remplaçant Xu_j par c_j on trouve que $XMX^tDc_j = \lambda_j c_j$. ■

2.2.4 ACP sur les données centrées réduites

En pratique, on travaille sur le tableau centré réduit Z pour accorder la même importance à chaque variable, avec la métrique $M = I_p$ qui est utilisée lorsque les unités de mesure est les variances associées à chaque variable sont différentes. Dans ce cas, la matrice de covariance V est égale à la matrice de corrélation R, donc les facteurs et les axes principaux sont les mêmes [11] i.e.

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

où ces derniers sont les vecteurs propres de la matrice de corrélation R associées aux valeurs propres où ces valeurs propres sont d'ordre décroissant

$$Ru_j = \lambda_j a_j.$$

Donc, les composantes principales sont données par

$$\boxed{c_j = Zu_j; j = \overline{1, p.}} \tag{2.4}$$

La première composante principale $c_1 = Zu_1 \in \mathbb{R}^p$ est la combinaison linéaire des variables centrées réduites ayant la variance maximale λ_1.[10]

En conclusion, on peut dire que l'ACP revient à remplacer les variables initiales $x_1, x_2, ..., x_p$ qui sont corrélées entre elles, par de nouvelles variables $c_1, c_2, ...$appelées composantes principales et qui sont des combinaisons linéaires des x_j, non corrélées entre elles.

Ces combinaisons linéaires $c_1, c_2, ...$sont les plus liées aux x_j parmi toutes les combinaisons linéaires des x_j.

Remarque 2.2.1 *La meilleure représentation plane du nuage des individus ou des variables est obtenue sur le premier plan principal, i.e. le plan engendré par les axes principaux correspondant aux deux plus grandes valeurs propres de R.*

2.3 Interprétation des résultats de l'ACP

Le but de l'ACP est d'obtenir une représentation des individus ou des variables dans un espace de dimension k plus faible que celle de l'espace initial de dimension p. Ceci se fait en construisant de nouvelles variables appelées composantes principales.

Celles-ci fournissent une image approchée du nuage des individus. Il y'a donc une perte d'information liée à la réduction de la dimension initiale p.

2.3.1 Interprétation des individus

Qualité de représentation du nuage des individus sur F_k

La qualité de la représentation obtenue par k valeurs propres est la proportion de l'inertie expliquée, la mesure de cette qualité ce fait à l'aide du critère de pourcentrage d'inertie [5]

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

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

Remarque 2.3.1 *Si $QLT(F_k)$ est proche de 1, alors 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

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

avec θ_{il} est l'angle formé entre le vecteur z_i et l'axe l.

Remarque 2.3.2

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

$$\begin{aligned} QLT_{(l,l')}(e_i) &= \frac{\textit{Inertie de la projection de l'individu } i \textit{ sur le plan } (l,l')}{\textit{inertie initiale de l'individu } i} \\ &= \cos^2(\theta_{i(l,l')}) = \frac{c_{il}^2 + c_{il'}^2}{\|z_i\|^2}. \end{aligned}$$

Donc, on peut dire que

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

2. *Plus la valeur du* $\cos^2$ *est proche de* 1*, donc la représentation graphique 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 :

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

Où, c_{il} est la valeur de la composante principale l pour l'individu i.[9]

Remarque 2.3.3 ***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 $i\prime$:

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

2.3.2 Interprétation des variables

Qualité de représentation du nuage des variables

Pour donner une signification à la composante principale c_l, il est important 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.[5]

On exprime la qualité de représentation d'une variable quantitative x_j sur le $l^{ème}$ 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 est la valeur propre associée à c_l,et u_{jl} est la $j^{ème}$ coordonnée du facteur principale u_l.

On a aussi

$$\boxed{r(c_l, x_j) = r(c_l, z_j) = \frac{c_l^t D z_j}{\sqrt{\lambda_l}}.}$$

Remarque 2.3.4 *Les corrélations d'une variable x_j avec les premieres composantes principales c_1 et c_2, s'expriment dans une figure appelée cercle des corrélations 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 donnée par la formule suivante

$$CTR_l(x_j) = \frac{r^2(c_l, x_j)}{\sum_{i=1}^{p} r^2(c_l, x_j)} = \frac{r^2(c_l, x_j)}{\lambda_l}.$$

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

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

Remarque 2.3.5 *En 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. Cela peut être fait en utilisant la 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 le choix du nombre d'axes à retenir en fonction du pourcentage d'inertie que l'on souhaite conserver.*

4.*Interprétation des nouvelles variables à l'aide des cercles de corrélations, il est importante de noter que les variables doivent être proches 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.[10]*

Chapitre 3

Application de l'ACP sur le logiciel R

Ce chapitre est consacré à la mise en œuvre de l'ACP à l'aide du logiciel R, un environnement de programmation largement utilisé pour le traitement et l'analyse des données. L'objectif est de montrer, à travers des données réelles, comment réaliser une ACP en exploitant les fonctionnalités offertes par ce logiciel.

Il existe de nombreux packages et fonctions pour l'itulisation de cette méthode d'anlyse, les plus populaires et utilisés sous R sont :

3.1 Packages et fonctions

• Il existe diverses packages, on cite quelques un :

- FactoMineR : Analyse de données exploratoire multivariée et fouille de données.

- factoextra : Visualisation des résultats des analyses multivariées et aide à leurs interprétations.

- ggplot2 : Représentation élégante et personnalisable des données selon une structure graphique claire.

- readr : Importation des données dans le logiciel R.

• Il existe diverses fonctions, on cite quelques un :

- read.csv : Insérer un tableau à partir de l'Excel.

- round : Arrondir les valeurs.

- PCA : Analyse en Composantes Principales.

- fviz_eig : Histogramme des valeurs propres.

3.2 Exemple d'application

Le tableau ci-dessous représente les moyennes de 29 élèves de première année lycée dans 12 matières différentes (arabe, mathématiques, physique, sciences naturelles, sciences islamiques, histoire, langue française et anglaise, informatique, technologie, dessin et sport) dans un établissement à Biskra.(voir 3.1)

3.3 Programme

```
X=read.csv("D :/Resultats/Resultats.csv",sep =" ;",dec = ",",header=TRUE,
  stringsAsFactors=FALSE,row.names = "Nom",fileEncoding = "latin1")
View(X)                    # Tableau des données X.
dim(X)                     # Matrice de 29 lignes et 12 colonnes.
summary(X)                 # Statistique descriptive.
```

g=colMeans(X,na.rm=TRUE) # Centre de gravité g.

arb	math	phy	sec.na	sec.isl	his	fr	ang	inf	techn	dess	sport
10.771	11.405	11.155	10.572	13.338	13.426	12.583	10.178	10.672	9.367	12.836	15.974

$Z = scale(X)$ # Tableau standard Z.

$$Z = \begin{bmatrix} 2.139 & 1.663 & 2.082 & 1.849 & 1.387 & \dots \\ 1.917 & 1.663 & 1.923 & 1.571 & 1.443 & \cdots \\ 1.647 & 1.981 & 0.781 & 1.067 & 1.500 & \dots \\ 1.043 & 1.595 & 1.020 & 1.496 & 0.712 & \dots \\ 0.804 & 1.981 & -0.121 & 1.496 & 1.275 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$V = cov(X)$ # Matrice de variance-covariance V.

$$V = \begin{bmatrix} 9.893 & & & & & \\ 9.517 & 13.426 & & & & \\ 10.189 & 10.741 & 14.200 & & & \\ 10.666 & 11.383 & 13.094 & 15.709 & & \\ 10.936 & 11.531 & 13.059 & 14.098 & 19.735 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$R = cor(X)$ # Matrice de corrélation R.

$$R = \begin{bmatrix} 1.000 & & & & & \\ 0.826 & 1.000 & & & & \\ 0.860 & 0.778 & 1.000 & & & \\ 0.856 & 0.784 & 0.877 & 1.000 & & \\ 0.783 & 0.708 & 0.780 & 0.801 & 1.000 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Valeurs propres et inerties :

res.pca=PCA(X,scale.unit=FALSE,ncp=6,graph=FALSE)

res.pca$eig # Pourcentage des valeurs propres.

round(res.pca$eig,2)

Valeurs propres	8.98	1.10	0.40	0.34	0.28	0.25	0.20	0.15	0.10	0.08	0.06	0.05
pourcentages d'inertie(%)	74.87	9.19	3.36	2.86	2.35	2.06	1.69	1.25	0.83	0.66	0.47	0.42
pourcentages cumulés(%)	74.87	84.06	87.42	90.28	92.63	94.68	96.37	97.62	98.45	99.11	99.58	100

TAB. 3.2 – Valeurs propres et inerties.

fviz_eig(res.pca,addlabels = TRUE,barfill = "purple",barcolor = "black")

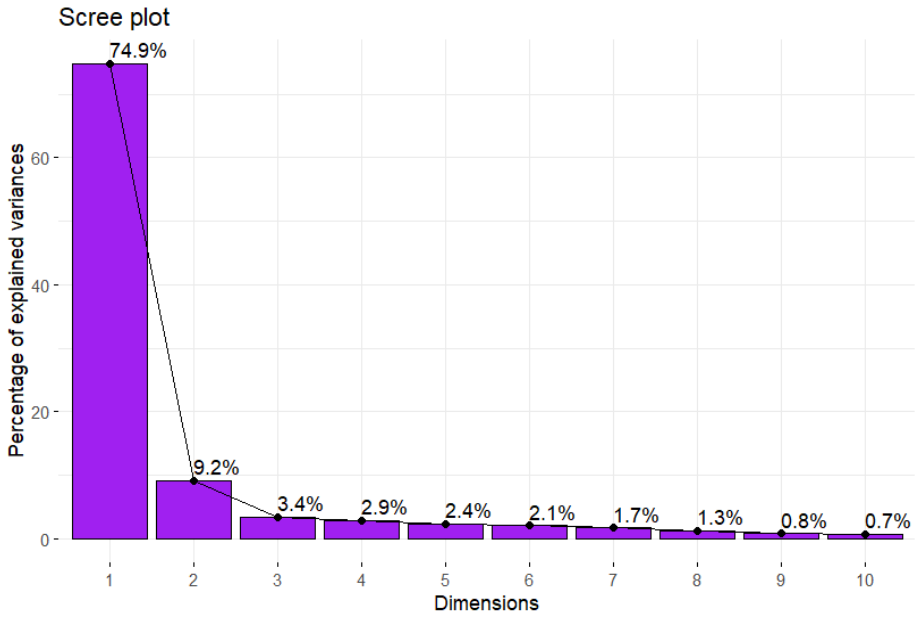

FIG. 3.1 – Pourcentage des valeurs propres.

Commentaire :

D'après la table 3.2 et l'histogramme 3.1, on remarque que le $1^{ér}$ axe représente environ $74,9\%$ de valeur propre λ_1, tandis que le $2^{ème}$ axe représente environ $9,2\%$ de valeur propre λ_2.

Pour cela, on ne prend que les deux premiers axes principaux, parce qu'ils traduisent à eux seuls $84,1\%$ de l'information disponible. On dit alors que l'analyse est de très bonne qualité sur les deux premières dimensions (plan).

Représentation de nuage des individus :

```
coord_ind=res.pca$ind$coord        # Coordonnées des individus.
cos2_ind=res.pca$ind$cos2          # Cosinus carrés des individus.
contrib_ind=res.pca$ind$contrib    # Contributions des individus.
coord=res.pca$ind$coord            # Composante principale.
c1 =coord[,1]                      # 1ér composante principale.
c2 =coord[,2]                      # 2éme composante principale.
round(range(c1),3)                 # Borne du 1er axe.
−5.859        5.706
round(range(c2),3)                 # Borne du 2éme axe.
−8.273        1.961
```

fviz_pca_ind(res.pca, label = "ind", geom.ind = "text", col.ind = "darkgreen",repel = TRUE, addEllipses = FALSE)

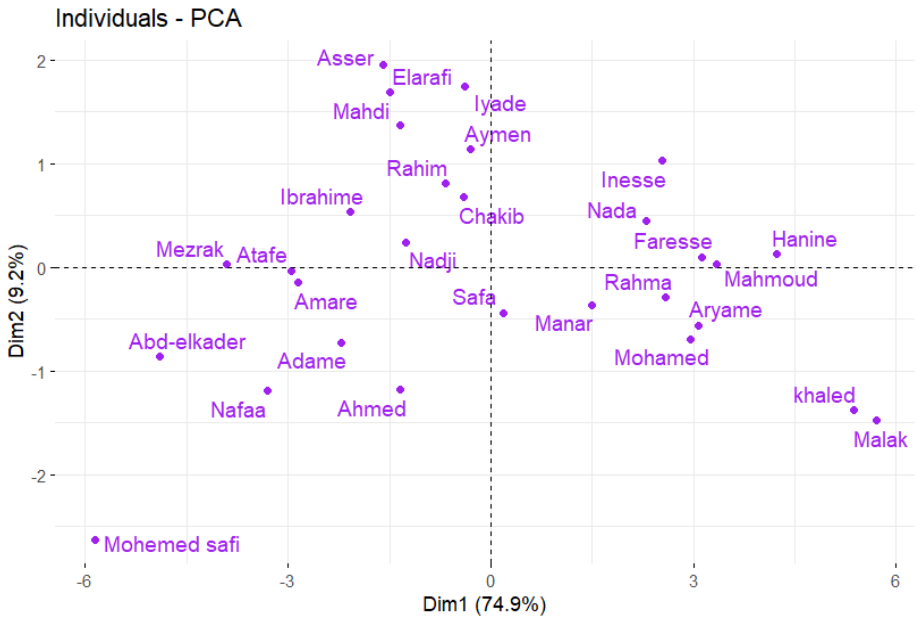

FIG. 3.2 – Représentation de nuage des individus.

Commentaire :

D'après l'histogramme 3.2 la projection des individus selon l'ACP. Les deux premières dimensions, *Dim* 1 et *Dim* 2, expliquent ensemble $84,1\%$ de la variance totale, avec une forte contribution de la première $74,9\%$. La majorité des individus sont regroupés autour du centre, ce qui indique une certaine similarité entre eux. Cependant, quelques individus comme "**mohemed safi**", "**khaled**" et "**malak**" sont éloignés, ce qui reflète des caractéristiques distinctes par rapport aux autres.

Tableau des coordonnées, contribution et qualité de représentation des individus :

```
tableau_individus=data.frame(Coord_c1=round(coord[,1],3),
Coord_c2=round(coord[,2],3),Contrib_c1=round(contrib[,1],3),
Contrib_c2=round(contrib[,2],3),Cos2_c1=round(cos2[,1],3),
Cos2_c2=round(cos2[,2],3))
```

(voir 3.3)

Commentaire :

Le tableau 3.3 présente les résultats d'analyse des individus selon les deux premiers axes principaux issus de l'ACP.

Les colonnes $c_1$et c_2 indiquent les coordonnées des individus sur les axes 1 et 2, Les colonnes CTR_1 et CTR_2 expriment la contribution relative de chaque individu à la formation des axes, exprimée en pourcentage. Enfin, les colonnes QLT_1 et QLT_2 mesurent la qualité de la représentation de chaque individu sur les axes, i.e. la proportion de l'inertie totale expliquée par l'axe considéré.

Une valeur élevée de la contribution indique que l'individu influence fortement la direction de l'axe correspondant. De plus, la QLT proche de 1 signifie que l'individu est bien représenté sur l'axe, alors qu'une valeur faible signale une mauvaise représentation.

À partir du tableau, on constate que certains individus, comme **"Malak"** et **"Mohamed Safi"**, possèdent des contributions élevées sur le premier axe (CTR_1), ce qui suggère leur importance dans l'interprétation de cet axe. D'autres individus, tels que **"Ayman"** et **"Amare"**, présentent une bonne qualité de représentation (QLT) proches de 1, indiquant qu'ils sont bien projetés sur les deux premiers axes principaux.

Représentation de nuage des variables :

```
coord_var=res.pca$var$coord        # Coordonnées des variables.
cos2_var=res.pca$var$cos2          # Cosinus carrés des variables.
contrib_var=res.pca$var$contrib    # Contributions des variables.
coord=res.pca$var$coord            # Composant principale.
c1=coord[,1]                       # 1^ér composant principale.
c2=coord[,2]                       # 2^éme composant principale.
round(range(c1),3)                 # Borne du 1^ér axe.
−8.945      15.503
round(range(c2),3)                 # Borne du 2^éme axe.
−8.273      4.943
fviz_pca_var(res.pca,label="var",geom.var="text",col.var="red",repel=TRUE,
             addEllipses=FALSE)
```

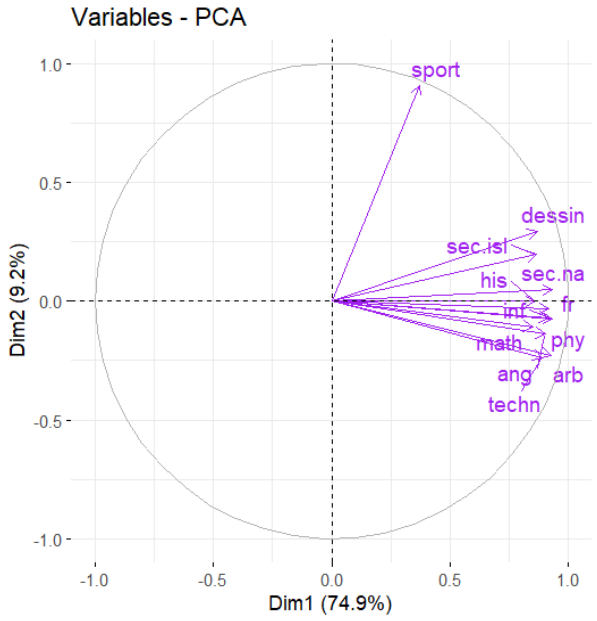

FIG. 3.3 – Représentation de nuage des variables.

Commentaire :

Ce graphique représente le cercle des corrélations 3.3 issu d'une ACP sur les moyennes des matières d'une classe. L'objectif est d'identifier les relations entre les matières et les dimensions principales ($Dim1$ et $Dim2$) expliquant la majorité de la variance.

"Sport" est isolé sur $Dim2$, ce qui indique un comportement différent par rapport aux autres matières.

"Math", "Phy", "Inf", "Ang", "Arb", "Techn" sont fortement corrélées entre elles, formant un groupe cohérent de matières scientifiques et linguistiques.

"Dessin", "Sec.ist", "Sec.na" sont moyennement représentées et moins liées aux autres.

On conclut que l'ACP révèle des groupes de matières au comportement similaire. Ces résultats peuvent aider à mieux comprendre les profils d'élèves et orienter le soutien scolaire. La matière "Sport" mérite une attention particulière.

Tableau des coordonnées, contribution et qualité de représentation des variables :

```
res.pcaPCA(X,scale.unit=TRUE,ncp=5,graph=TRUE)
tableau_variables=data.frame(Coord_c1=round(coord[,1],3),
Coord_c2=round(coord[,2],3),Contrib_c1=round(contrib[,1],3),
Contrib_c2=round(contrib[,2],3),Cos2_c1=round(cos2[,1],3),
Cos2_c2=round(cos2[,2],3))
```

(voir 3.4)

Commentaire :

Le tableau 3.4 représente les résultats de l'analyse des variables selon les deux premiers axes principaux issus de l'ACP.

Les colonnes $c_1$et c_2 indiquent les coordonnées des variables sur les axes 1 et 2. Les colonnes CTR_1 et CTR_2 expriment la contribution relative de chaque variables à la formation des axes, exprimée en pourcentage. Enfin, les colonnes QLT_1 et QLT_2 mesurent la qualité de la représentation de chaque variables sur les axes, i.e. la proportion de l'inertie totale expliquée par l'axe considéré.

Une valeur élevée de contribution indique que les variables influencent fortement la direction de l'axe correspondant. De plus, une QLT proche de 1 signifie que la variable est bien représentée sur l'axe, alors qu'une valeur faible signale une mauvaise représentation.

À partir du tableau, on remarque que les variables telles que "**arab**", "**phy**", "**inf**" et "**ang**" présentent des contributions élevées et de bonnes qualités de représentation sur le premier axe. En revanche, la variable "**sport**" est mieux représentée sur le deuxième axe avec une forte contribution.

Conclusion

Dans ce mémoire, on a étudié des méthodes multidimensionnelles de l'analyse des données à savoir l'ACP. L'objectif de cette analyse était de traiter un nombre très important de données afin de visualiser ces derniers dans le meilleur espace réduit. Cette analyse est réalisable lorsqu'il est possible de réduire l'espace multidimensionnel en un espace à deux ou trois dimensions, de telle sorte que cet espace réduit conserve une part importante de l'information qui était contenue dans l'espace multidimensionnel d'origine.

On a présenté le principe de l'ACP qui est une technique d'analyse statistique principalement descriptive, qui consiste à représenter sous forme graphique le plus d'informations possibles contenues dans un tableau. Cette méthode permet de visualiser un espace à p dimensions à l'aide des espaces de dimensions plus petites.

On a également mené une étude et une interprétation des résultats obtenus par l'ACP sur des données réelles et ceci en utilisant les différents packages du logiciel R.

Il existe galement d'autres méthodes utilisées pour differentes données tel que l'analyse factorielle en correspondance notée AFC, analyse des correspondance multiples notée ACM, ect.

Annexe A : Logiciel *R*

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. *R* a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets. Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, les différentes méthodes d'analyse des données,... Plusieurs paquets, tels *ade4, FactoMineR, MASS, multivariate, scatterplot3d et rgl* entre autres sont destinés à l'analyse des données statistiques multidimensionnelles.

Il a été initialement créé par [6]du département de statistique de l'Université d'Auckland en Nouvelle Zélande.

Annexe B : Tableau des données

	arb	math	phy	se-n	se-is	his	fr	ang	inf	tech	des	sport
Malak	17.50	17.50	19.0	17.9	19.50	19.63	19.1	18.0	17.95	18.00	16.75	15.00
khaled	16.80	17.50	18.4	16.8	19.75	18.13	18.8	17	18.55	18	17	15.00
Hanine	15.95	15	14.1	14.8	20	19.25	17.2	16.8	13.75	17	16.25	17.50
Mohamed	14.05	17.25	15	16.5	16.5	15.88	15.6	12.8	17.25	10.2	13.5	15.50
Faresse	13.3	15	10.7	16.5	19	18.13	16.5	15.2	16.85	11.9	14	16.75
Aryame	14	14.25	15.3	14.6	13.25	17.38	17	17.4	12.35	13.9	16.5	16.00
Mahmoud	14.4	12.25	15.5	12.2	18.13	18.13	16.7	14	17.25	13	17.5	16.50
Rahma	13	17.5	15.8	11.1	18	13.5	14.8	15.9	12.6	11.2	17.25	15.75
Inesse	11.05	11.5	14.9	13.4	14	15.13	17.1	17.1	13.1	12.6	18	18.25
Nada	10.8	11.75	14.3	11.7	16.13	17.13	15.7	11.5	14.7	14.15	18	16.75
Manar	13.7	13	12.2	12.7	13	12.75	15.8	15.9	12.15	10	14.5	16.00
Safa	11.85	9.00	13.0	12.9	14.25	14.63	10.5	8.4	10.00	11.10	13.25	15.00
Iyade	11.70	12.00	11.0	10.0	16.50	9.75	10.0	5.40	11.15	7.70	12.75	19.00
Chakib	9.55	13.00	9.20	8.90	18.50	12.63	13.1	10.3	6.50	5.70	14.50	16.00
Aymen	8.70	9.00	11.0	11.1	14.63	11.50	10.8	10.7	9.80	10.10	14.75	17.50
Rahim	8.50	11.50	9.70	10.7	10.13	15.88	9.60	5.80	12.05	10.00	12.00	17.50
Nadji	9.50	15.75	8.80	9.10	8.13	14.38	10.4	7.20	10.05	4.40	10.75	16.50
Ahmed	9.70	11.00	10.8	9.80	10.63	9.00	10.0	11.8	10.20	9.40	9.00	14.25
Elarafi	7.15	10.75	10.5	8.30	13.13	9.25	10.8	4.90	10.10	6.90	13.25	18.25
Mahdi	8.25	7.25	10.0	10.4	13.00	13.63	11.2	11.3	6.05	4.20	12.75	17.50
Ibrahime	8.35	8.25	10.2	9.60	11.50	7.88	11.1	6.80	8.05	6.80	10.25	16.50
Asser	8.35	8.25	7.50	9.70	13.00	12.88	10.8	5.60	6.80	5.70	14.75	18.25
Adame	10.1	9.50	7.20	6.80	8.75	12.38	10.90	9.10	6.95	7.40	9.25	14.75
Atafe	8.75	8.75	7.90	7.30	12.0	13.00	8.80	6.25	4.50	5.30	5.50	15.75
Nafaa	09.10	09.00	6.50	05.60	09.13	10.63	7.70	6.05	08.55	05.80	07.50	13.50
Amare	8.65	6.50	5.90	3.90	9.25	12.38	13.60	3.50	9.00	7.10	9.50	15.25
Mezrak	7.50	6.50	7.40	7.50	8.00	10.50	5.90	3.80	6.35	3.50	9.00	15.00
Abd-elkader	6.35	7.00	6.20	4.20	3.00	7.50	10.20	4.05	3.70	3.90	7.75	13.75
Mohemed safi	5.75	5.25	5.50	2.60	6.00	6.50	5.20	2.60	3.20	6.70	6.50	10.00

TAB. 3.1 – Tableau des données.

arb : arabe des : dessin phy : physique se-n : sciences naturelles

his : histoire ang : anglais inf : informatique se-is : science islamique

fr : français sport : sport tech : technologie math : mathématiques

Annexe C : Tableau des Coordonnées, Contribution et qualité de représentation des individus

	c_1	c_2	CTR_1	CTR_2	QLT_1	QLT_2
Malak	5.706	−1.477	12.495	6.824	0.930	0.062
khaled	5.371	−1.379	11.072	5.946	0.928	0.061
Hanine	4.228	0.128	6.861	0.051	0.916	0.001
Mohamed	2.950	−0.692	3.340	1.496	0.749	0.041
Faresse	3.123	0.103	3.744	0.033	0.791	0.001
Aryame	3.075	−0.560	3.630	0.979	0.843	0.028
Mahmoud	3.333	0.035	4.265	0.004	0.879	0.000
Rahma	2.590	−0.291	2.575	0.265	0.689	0.009
Inesse	2.530	1.030	2.457	3.315	0.648	0.107
Nada	2.302	0.446	2.035	0.622	0.705	0.026
Manar	1.502	−0.359	0.866	0.403	0.567	0.032
Safa	0.178	−0.437	0.012	0.567	0.014	0.083
Iyade	−0.391	1.747	0.059	9.544	0.025	0.491
Chakib	−0.395	0.679	0.060	1.441	0.037	0.109
Aymen	−0.301	1.142	0.035	4.081	0.035	0.498
Rahim	−0.679	0.811	0.177	2.056	0.114	0.162
Nadji	−1.265	0.200	0.575	0.117	0.248	0.006
Ahmed	−1.334	−1.311	0.636	5.005	0.341	0.336
Mahdi	−1.345	1.391	0.646	5.631	0.316	0.345
Elaraf	−1.496	1.601	0.822	7.464	0.335	0.380
Asser	−1.591	1.963	0.920	11.214	0.363	0.553
Ibrahime	−2.085	0.455	1.591	0.604	0.705	0.033
Adame	−2.223	−0.755	1.800	1.661	0.781	0.090
Amare	−2.853	−0.158	3.008	0.073	0.699	0.002
Atafe	−2.958	−0.069	3.228	0.014	0.768	0.000
Nafaa	−3.309	−1.302	4.048	4.939	0.825	0.126
Mezrak	−3.909	−0.034	5.661	0.003	0.942	0.000
Abd-elkader	−4.898	−0.859	9.207	2.310	0.913	0.028
Mohemed safi	−5.859	−2.627	13.176	21.581	0.803	0.161

TAB. 3.3 – Tableau des Coordonnées, Contribution et qualité de représentation des individus.

Annexe D : Tableau des Coordonnées, Contribution et qualité de représentation des variables

	c_1	c_2	CTR_1	CTR_2	QLT_1	QLT_2
arb	0.928	−0.228	9.588	4.729	0.861	0.052
math	0.848	−0.110	8.004	1.101	0.719	0.012
phy	0.934	−0.076	9.701	0.522	0.872	0.006
sec.na	0.933	0.049	9.682	0.217	0.870	0.002
sec.isl	0.865	0.197	8.337	3.512	0.749	0.039
his	0.859	−0.001	8.205	0.000	0.737	0.000
fr	0.918	−0.031	9.376	0.087	0.842	0.001
ang	0.901	−0.139	9.041	1.754	0.812	0.019
inf	0.915	−0.069	9.321	0.434	0.837	0.005
techn	0.891	−0.238	8.827	5.157	0.793	0.057
dessin	0.868	0.292	8.382	7.756	0.753	0.086
sport	0.371	0.908	1.535	74.732	0.138	0.824

TAB. 3.4 – Tableau des Coordonnées, Contribution et qualité de représentation des variables.

Bibliographie

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

[2] Benzécri, J.-P. (1973). *L'analyse des données. Tome 2 : L'analyse des correspondances*. Paris : Dunod.

[3] Duby, C., & Robin, S. (2006, juillet 10). *Analyse en composantes principales*. INA, Paris- Grignon.

[4] Escofier, B., & Pagès, J. (2008). Analyses factorielles simples et multiples : objectifs, méthodes et interprétation (4 éd.). Paris : Dunod.

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

[6] Fenelon, J.-P. (1981). *Qu'est-ce que l'analyse des données*(Vol. 3, No. 1).Lefonen.

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

[8] Meraghni, D. (2018). *Analyse en Composantes Principales*. Cours de master 1, Université Mohamed Khider, Biskra.

[9] Necir, A., (2024) *Analyse en Composantes Principales*. Cours de master 1, Université Mohamed Khider, Biskra.

[10] Pearson, K. (1901). *On Lines and Planes of Closest Fit to Systems of Points in Space*. Philosophical Magazine, Series 6.

[11] Saporta,G., & Niang,N.(2003).*Analyse en composantes principales.*In G. Govaert (Ed.), *Analyse des données* (pp. 19–42). Paris : Hermès.

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

Résumé

Ce travail a pour but d'étudier l'Analyse en Composantes Principales (ACP), un outil très efficace pour résumer et interpréter une grande quantité de données quantitatives. L'ACP permet de mettre en évidence les corrélations entre les variables et les ressemblances entre les individus tout en réduisant l'information dans un espace plus simple et lisible. Dans notre cas, on a appliqué cette méthode aux résultats de 29 élèves évalués dans 12 matières différentes à l'aide du logiciel statistique R.

Mots clés : Analyse des Données, ACP, Variables Quantitatives, Matrice des poids.

Abstract

This work aims to study Principal Component Analysis (PCA), a powerful statistical tool used to summarize and interpret large amounts of quantitative data. PCA highlights the correlations between variables and the similarities between individuals, while reducing the information to a simpler and more readable space. In our study, we applied this method to the results of 29 students evaluated in 12 different subjects, using the statistical software R.

Keywords : Data Analysis, PCA, Quantitative Variables, Loading Matrix.

الملخص

يهدف هذا العمل إلى دراسة التحليل بالمكونات الرئيسية(PCA) ، وهو أداة إحصائية قوية تُستخدم لتلخيص وتفسير كميات كبيرة من البيانات الكمية. يبرز هذا التحليل العلاقات بين المتغيرات والتشابهات بين الأفراد، مع تقليص المعلومات إلى فضاء أبسط وأسهل في القراءة. في دراستنا، قمنا بتطبيق هذه الطريقة على نتائج 29 تلميذًا تم تقييمهم في 12 مادة مختلفة، باستخدام البرنامج الإحصائي.R

الكلمات المفتاحية: تحليل البيانات، تحليل المكونات الرئيسية، المتغيرات الكمية، مصفوفة الأوزان.