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

MIMI Amel

Titre :

Analyse en Composantes Principales

Membres du Comité d'Examen :

Pr.	MERAGHNI Djamel	UMKB	Président
Dr.	CHINE Amel.	UMKB	Encadreur
Dr.	BENLMIR Imen	UMKB	Examinateur

Septembre 2020

Dédicace

Avant tout propos, 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

Tous d'abord :

A la lumière et symbole de la vie, à la source de tendresse «Ma mère» qui m'encourage, me réconforte et qui ne cesse de sacrifier pour assurer

A mon secret de ma réussite, à mon adorable «Mon père» qui me soutient et qui est toujours présent pour moi, tes encouragements et ton motivation qui me réalise cette réussite.

Qui Dieu vous garde en bonne santé et vous procure une longue vie.

A mes chers frères

A toute ma famille

A toute mes amies

A tous les étudiants de Mathématique, surtout 2_eme année Master et surtout Groupe de STATISTIQUE

et tous les étudiants de l'université MOHAMMED KHIEDER.

Mimi Amel.

REMERCIEMENTS

Tous d'abord je tiens a remercie Allah pour la santé, la force et la patience qui ma donner pour terminer ce travail.

Je tiens tout particulièrement à remercier mon encadreur

Dr. Chine Amel pour la suivi et l'aide qu'elle m'a apporté pour l'élaboration et pour ses précieux conseils et ses aides durant toute la période du travail de ce mémoire.

Je remercie les membres du jury :

Pr.MERAGHNI Djamel et Dr.BENLMIR Imen.

Et tous les enseignements de Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie en particulier Département de Mathématique et tous ces profs.

Enfin, je remercie ma famille qui mon encouragé au long de ma vie, spécialement mes chères parents et mes frères et à tous, pour tous.

Table des matières

Table des figures

Liste des tableaux

Introduction

On désigne par statistique descriptive multidimensionnelle l'ensemble des méthodes de la statistique descriptive (ou exploratoire) permettant de traiter simultanément un nombre quelconque de variables. Les méthodes les plus classiques de la statistique descriptive multidimensionnelle sont les méthodes factorielles. Les domaines d'utilisation de ces méthodes sont nombreux et diversifiés : biologie, économétrie, médecine, etc . . .

Il existe une multitude de méthodes factorielles permettant de traiter différences 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 au moins 'a K.Pearson (1901), mais leur pratique n'est devenue courante que depuis l'ère informatique. Elles ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzekri qui a beaucoup exploité les aspects géométriques et les représentations graphiques.

L'Analyse en Composantes Principales (ACP) ou principal component analysis (PCA) en anglai est une méthode d'analyse statistique multidimensionnelle. Permettant d'étudier simultanément un grand nombre des variables statistiques, dans le but de mettre en valeur des liaisons qui peuvent exister entre elles (les variables). Elle consiste à réduire les données statistiques pouvant être trop nombreuses au départ et les présenter sous forme de graphiques afin de voir leur structure, l'ACP est la méthode qui traite les tableaux

croisant des individus et des variables quantitatives.

Le but de ce mémoire alors est de présenter, de faire une description de l'ACP, de savoir comment résoudre le problème de la représentation des données très nombreuses où les variables quantitatives et étudier la ressemblance entre les individus et la liaison entre les variables. Ce travail ce devise en deux chapitres :

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

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

Finalement, à l'aide du logiciel R, on va effectuer un exemple d'étude de différentes caractéristiques de l'approche de l'ACP " l'exemple des notes des élèves". En donnant des remarques et des résultats d'ont l'obtention à travers la réalisation de l'ACP.

Chapitre 1

Préliminaires

L'analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple par utilisation des plusieurs méthodes , dont les plus importances sont : l'analyse en composantes principales (ACP), l'analyse factorielles des correspondance(AFC), l'analyse canonique(AC). Ces méthodes sont beaucoup utilisées dans un grand nombre de domaines : les domaines scientifiques et industriel et aussi en marketing, en météorologie. [1]

Dans ce chapitre, on s'intéresse d'abord à la description de ces données ainsi qu'à leurs caractéristiques comme le tableau des données, puis on définit les individus, les variables, la matrice des poids, le centre de gravité...etc.

1.1 Données et leurs caractéristiques

1.1.1 Tableau des données

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

$$X = \begin{bmatrix} x_{11} & .... & x_{1p} \\ . & & . \\ . & x_{ij} & . \\ . & & . \\ x_{n1} & .... & x_{np} \end{bmatrix} \in M_{\mathbb{R}}(n,p). \tag{1.1}$$

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

Dans une optique purement descriptive on identfiera une variable à la colonne de X correspondante : une variable n'est rein d'autre que la liste des n valeurs qu'elle prend sur les n individus :

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

On identifiera de même l'individu i au vecteur e_i à p composantes :

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

1.1.2 Types de variables

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

Définition 1.1.1 (variable quantitative) *:Ses valeurs sont des nombres exprimant une quantité, sur lesquels les opérations arithmétiques (somme, etc.) ont un sens. La variable peut alors être discrète ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre (valeurs isolées ou intervalle de R).*

Définition 1.1.2 (Variable qualitative) *:Ses valeurs sont des modalités, (ou catégories, ou caractères) exprimées sous forme littérale ou par un codage numérique sur lequel des opérations arithmétiques n'ont aucun sens. On distingue des variables qualitatives ordinales ou nominales, selon que les modalités peuvent être naturellement ordonnées ou pas. [2]*

Exemple 1.1.1 *Dans une entreprise employant, 7 personnes ($n = 7$), on étudie les variables salaire mensuel (en euros) et âge ($p = 2$) comme :*

$x_1 = (2157, 2053, 2924, 1862, 3106, 2754, 1027)^t \in \mathbb{R}^7$ *et*

$x_2 = (20, 30, 35, 55, 26, 45, 39)^t \in \mathbb{R}^7$*,alors :*

$$X = \begin{bmatrix} 2157 & 20 \\ 2053 & 30 \\ 2924 & 35 \\ 1862 & 55 \\ 3106 & 26 \\ 2754 & 45 \\ 1027 & 39 \end{bmatrix}. \tag{1.2}$$

1.1.3 Matrice de poids

On affècte à chaque individu un poids p_i reflétant son importance par rapport aux autres individus. On appelle matrice des poids la matrice diagonale (n, n) dont les éléments diagonaux sont les poids p_i. Elle sera notée :

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

Dans le cas usuel des poids égaux, on a :

$$D = \frac{1}{n}\mathbf{I}_n. \tag{1.3}$$

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

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

Par conséquent

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

Et

$$D = \begin{bmatrix} \frac{1}{n} & .... & 0 \\ : & & : \\ : & & : \\ 0 & .... & \frac{1}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & .... & 0 \\ : & & : \\ : & & : \\ 0 & .... & 1 \end{bmatrix} = \frac{1}{n}\mathbf{I}_n.$$

■

Exemple 1.1.2 *D'après l'exemple (1.2) on a $n = 7$. On va calculer la matrice de poids comme suit :*

$$D = \begin{bmatrix} \frac{1}{7} & 0 & ... & 0 \\ 0 & \frac{1}{7} & & 0 \\ \vdots & & \ddots & 0 \\ 0 & ... & 0 & \frac{1}{7} \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & .. & 0 \\ \vdots & & \ddots & 0 \\ 0 & ... & 0 & 1 \end{bmatrix}. \tag{1.4}$$

1.1.4 Centre de gravité

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

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

où , $\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} = \sum_{i=1}^{n} p_i x_{ij}$ pour $j = 1, 2, ...p$.

On peut l'écrire sous forme matricielle de X et $D(1.1)$, (1.3) :

$$g = X^t D \mathbf{1}_n \qquad \text{tel que } \mathbf{1}_n = (1, 1, ...1)^t \in \mathbb{R}^n. \tag{1.5}$$

Preuve. on a

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

$$= \begin{bmatrix} \sum_{i=1}^{n} p_i x_{i1} \\ : \\ : \\ \sum_{i=1}^{n} p_i x_{ip} \end{bmatrix} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_p} \end{bmatrix} = g.$$

■

Exemple 1.1.3 *D'après l'exemple (1.2) et (1.4), on va calculer le centre de gravité comme suit :*

$$g = X^t D\mathbf{1}_n = \begin{bmatrix} 2157 & 2053 & 2924 & 1862 & 3106 & 2754 & 1027 \\ 20 & 30 & 35 & 55 & 26 & 45 & 39 \end{bmatrix} \begin{bmatrix} \frac{1}{7} & .. & .. & 0 \\ & \frac{1}{7} & & \\ & & \ddots & \\ 0 & & & \frac{1}{7} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2157 & 2053 & 2924 & 1862 & 3106 & 2754 & 1027 \\ 20 & 30 & 35 & 55 & 26 & 45 & 39 \end{bmatrix} \begin{bmatrix} \frac{1}{7} \\ \vdots \\ \vdots \\ \frac{1}{7} \end{bmatrix} = \begin{bmatrix} 2269 \\ 35.71 \end{bmatrix}.$$

alors $g = (2269, 35.71)^t$.

1.1.5 Standardisation des données

Il existe deux types de transformations utilisées sur les données initiales :

Tableau centré : L'analyse centrée consiste à modifier les données du tableau X en remplaçant les valeurs des x_{ij} par :

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

La forme matricielle :

$$Y = X - \mathbf{1}_n g^t. \tag{1.6}$$

Si on remplace g par sa formule matricielle (1.5) on obtient :

$$\begin{aligned} Y &= X - \mathbf{1}_n (X^t D \mathbf{1}_n)^t \\ &= X - \mathbf{1}_n (\mathbf{1}_n^t D X) \\ &= (\boldsymbol{I}_n - \mathbf{1}_n \mathbf{1}_n^t D) X. \end{aligned}$$

Tableau centré réduite : L'analyse centrée réduite ou encore normée, que nous présentons ici, est liée à la transformation des données du tableau X en remplaçant les valeurs des x_{ij} par :

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{\delta_j} = \frac{y_{ij}}{\delta_j}. \tag{1.7}$$

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

La forme matricielle utilisant la formule(1.6) :

$$Z = Y D_{\frac{1}{\delta}}.$$

On note $D_{\frac{1}{\delta}}$ la matrice de taille p diagonale, des inverses des écarts-types :

$$D_{\frac{1}{\delta}} = \begin{bmatrix} \frac{1}{\delta_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\delta_p} \end{bmatrix}.$$

1.1.6 Matrice de variance-covariance

C'est une matrice carrée symétrique d'ordre p notée par V :

$$V = \begin{bmatrix} \delta_1^2 & \dots & \delta_{1p} \\ \vdots & \ddots & \vdots \\ \delta_{p1} & \dots & \delta_p^2 \end{bmatrix}.$$

où $\delta_{jj'}$ est la covariance des variables x_j et $x_{j'}$ on peut la calculer comme suit :

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

Et δ_j^2 est la variance de la variable x_j tel que :

$$\delta_{jj} = \delta_j^2 = \text{var}(x_j) = \sum_{i=1}^{n} p_i(x_{ij} - \overline{x_j})^2.$$

La forme matricielle :

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

1.1.7 Matrice de corrélation

C'est une matrice $(p \times p)$ qui regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux notée R tel que :

$$R = \begin{bmatrix} 1 & r_{12} & ... & r_{1p} \\ r_{21} & 1 & ... & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & ... & 1 \end{bmatrix}.$$

son terme général est :

$$r_{jj'} = \frac{\text{cov}(x_j, x_{j'})}{\delta_j \delta_{j'}} = \frac{\delta_{jj'}}{\delta_j \delta_{j'}}. \tag{1.8}$$

La forme matricielle :

$$R = D_{\frac{1}{\delta}} V D_{\frac{1}{\delta}} = Z^t D Z.$$

Preuve. on a

$$D_{\frac{1}{\delta}} V D_{\frac{1}{\delta}} = \begin{bmatrix} \frac{1}{\delta_1} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \frac{1}{\delta_p} \end{bmatrix} \begin{bmatrix} \delta_1^2 & ... & \delta_{1p} \\ \vdots & \ddots & \vdots \\ \delta_{p1} & ... & \delta_p^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\delta_1} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \frac{1}{\delta_p} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & ... & \frac{\delta_{1p}}{\delta_1 \delta p} \\ \vdots & \ddots & \vdots \\ \frac{\delta_{p1}}{\delta_p \delta_1} & ... & 1 \end{bmatrix} = R.$$

■

Remarque 1.1.1 *La matrice R est la matrice de variance-covariance des données centrées réduites et résume la structure des dépendances linéaires entre les p variables prise 2 à 2. [8]*

1.2 Nuage des individus

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

1.2.1 Ressemblance entre deux individus

Définition 1.2.1 *Deux individus se ressemblent, ou sont proches, s'ils possèdent des valeurs proches pour l'ensemble des variables. [6]*

Cette définition sous entend une notion de proximité qui se traduit par une distance :

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

1.2.2 Rôle de la métrique

Comment mesurer la distance entre deux individus ? Cette question primordiale doit être résolue avant toute étude statistique car les résultats obtenus en dépendent dans une large mesure.

En physique, la distance entre deux points de l'espace se calcule facilement par la formule de **Pythagore**, car les dimensions sont de même nature, ce sont des longueurs que l'on mesure avec la même unité. Mais en statistique il n'en est pas de même où chaque dimension correspond à un caractère qui s'exprime avec son unité particulière. Alors, la distance utilisée en générale entre deux individus e_i et $e_{i'}$ est définie par la forme quadratique suivante :

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

où M est une matrice symétrique de taille p définie positive. L'espace des individus est

donc muni de produit scalaire comme suit :

$$\prec e_i, e_{i'} \succ_M \quad = e_i^t \; M \; e_{i'}.$$

En pratique les métriques usuelles en Analyse en Composantes (acp) sont en nombre réduit par exemple la métrique $M = \boldsymbol{I}_p$ qui revient à utiliser le produit scalaire usuel mais la plus utilisée est la métrique diagonale des inverses des variances :

$$M = D_{\frac{1}{\delta^2}} \begin{bmatrix} \frac{1}{\delta_1^2} & 0 & ... & 0 \\ 0 & \frac{1}{\delta_2^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \frac{1}{\delta_p^2} \end{bmatrix}$$

Ce qui revient à diviser chaque caractère par son écart-type, entre autres avantages, la distance entre deux individus ne dépend plus des unités de mesure puisque les nombres $\frac{x_{ij}}{\delta_j}$ sont sans dimension, ce qui est très utile lorsque les variables ne s'expriment pas avec les mêmes unités.

Remarque 1.2.1 *Quand on travaille sur le tableau Y on utilise la métrique $M = D_{\frac{1}{\delta^2}}$ mais quand on travaille sur le tableau Z on utilise la métrique $M = \boldsymbol{I}_p$ tellque $\boldsymbol{I}_p$ la matrice identité d'ordre p. [8]*

1.2.3 Inertie

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

$$I_g = \sum_{i=1}^{n} p_i (e_i - g)^t M \; (e_i - g) = \sum_{i=1}^{n} p_i \parallel e_i - g \parallel_M^2 = \sum_{i=1}^{n} p_i \prec e_i - g, e_i - g \succ_M$$

L'inertie en un point a quelconque $a \in \mathbb{R}^p$ est définie par :

$$I_a = \sum_{i=1}^{n} p_i \parallel e_i - a \parallel_M^2 = \sum_{i=1}^{n} p_i (e_i - a)^t M \ (e_i - a)$$

On a la relation de **Huyghens :**

$$I_a = I_g + (g-a)^t M \ (g-a) = I_g + \parallel g - a \parallel_M^2$$

Remarque 1.2.2 *1. si $g = 0$ (X centré) alors*

$$I_g = \sum_{i=1}^{n} p_i \ {e_i}^t M \ e_i = \sum_{i=1}^{n} p_i \prec e_i, e_i \succ_M$$

2. On définit l'inertie totale par la relation suivante :

$2I_g = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (e_i - e_j)^t M \ (e_i - e_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \parallel e_i - e_j \parallel_M^2$, *où I_g la moyenne des carrées de to*

[8]

Proposition 1.2.1 *L'inertie totale est la trace de la matrice MV ou VM est définie par : $I_g = tr(MV) = tr(VM)$. [8]*

Preuve. 1.On démontre que $I_g = tr(MV)$:

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

2.On a $tr(A) = tr(A^t)$ donc

$$I_g = tr(MV) = tr((MV)^t) = tr(V^t M^t) = tr(VM).$$

■

Remarque 1.2.3 *1. si $M = \mathbf{I}_p$ l'inertie est égale à la somme des variances des p variables c-â-d :*

$$I_g = tr(I_p V) = tr(V) = \sum_{j=1}^{p} \delta_j^2.$$

2. si $M = D_{\frac{1}{\delta^2}}$ alors $tr(MV) = tr(D_{\frac{1}{\delta^2}} V) = tr(D_{\frac{1}{\delta}} V D_{\frac{1}{\delta}})$, ce qui est égal à $tr(R) = p$. L'inertie est donc égale au nombre des variables et ne dépend pas de leurs valeurs. [8]

1.3 Nuage des variables

Chaque variable x_j peut alors être représentée par un vecteur de $\mathbb{R}^n$ appelé espace vectoriel des variables et chaque dimension de $\mathbb{R}^n$ est représentée par individu. L'ensemble des p

variables constitue un nuage de points appelé nuage des variables.

1.3.1 Métrique des poids

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

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

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

2. Si les deux variables sont centrées alors :

$$\prec x_j, x_{j'} \succ_D = \text{cov}(x_j, x_{j'}) = \delta_{jj'}.$$

3. La norme d'une variable $\| \ x_j \ \|_D$(i.e. la longueur d'une variable) est égale à son écart-type :

$$\| \ x_j \ \|_D^2 = \delta_j^2$$

4. L'angle $\theta_{jj'}$ entre deux variables centrées est donnée par :

$$\cos\theta_{jj'} = \frac{\prec x_j, x_{j'} \succ_D}{\| \ x_j \ \|_D \| \ x_{j'} \ \|_D} = \frac{\delta_{jj'}}{\delta_j \delta_{j'}} = r_{jj'} = \frac{\text{cov}(x_j, x_{j'})}{\delta_j \delta_{j'}}$$

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 aux l'angle entre les vecteurs. [8]*

Chapitre 2

Analyse en composantes principales

L'analyse en composantes principales que nous notons par ACP, est un la "mère" de la plupart des méthodes descriptives multidimensionnelles appelées méthode factorielles. Elle a été conçue par **Karl Pearson**(1901) et intégrée à la statistique par **Harold Hotelling**(1933) [4] . Elle cherche à représenter graphiquement les relations entre individus par l'évaluation de leurs ressemblances, ainsi que les relations entre variables par l'évaluation de leurs liaisons, l'étude doit se faire simultanément. Le but final de ces représentations est l'interprétation par une analyse des résultats.

Dans le cas de l'ACP les données doivent être quantitatives, continues, elles peuvent être homogènes ou non et sont a priori corrélées entre elles.

2.1 Principe de l'ACP

Si $p = 3$ on peut représenter les individus mais lorsque la dimension est plus grand que 3, il est possible de les visualiser. Dans ce cas, on cherche une représentation approché du nuage des n individus dans un sous espace F_k de $\mathbb{R}^p$ de dimension $k \prec p$ (dimension faible). Autrement dit on cherche à définir k nouvelles variables dites combinaison linéaire des p variables initiales contenant le plus d'informations possible. [3]

2.2 Projection des individus sur un sous-espace

Le choix de l'espace de projection s'effectue pour déformer le moins possible les distances en projection. Le sous-espace de projection F_k de dimension $k < p$ recherché est tel que la moyenne des carrés des distances entre projections soit la plus grande possible. En d'autres termes il faut que l'inertie du nuage projeté sur le sous-espace F_k soit maximale.

Définition 2.2.1 *soit P la matrice de projection (operateur) $M-$orthogonale sur le sous espace F_k, elle vérifie les deux conditions suivantes :*

1. $P^2 = P$ (P est idompotente).
2. $MP = P^tM$ (P est $M-$symétrique). [8]

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

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

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

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

Remarque 2.2.1

1. *f_i^t est la $i^{éme}$ ligne du tableau X_{proj} ainsi que e_i est la $i^{éme}$ ligne du tableau initial X.*
2. *La matrice de covariance associée au nuage projeté :*

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

3. *L'inertie du nuage projeté :*

$$I_{proj} = tr(VMP).$$

4. Le centre de gravité projeté :

$$g_{proj} = Pg.$$

Preuve. On a :

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

1. On a :

$$\begin{aligned} I_{proj} &= tr(V_{proj} M) = tr(P V P^t M) \\ &= tr(PVMP), \text{ car } P^t M = MP, \\ &= tr(VMP^2), \text{ car } tr(AB) = tr(BA) \\ &= tr(VMP), \text{ car } P \text{ est idempotente.} \end{aligned}$$

2. On a :

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

■

-Le problème est donc de trouver la matrice de projection P M-orthogonale de rang k qui maximise $tr(VMP)$ pour construire le sous-espace F_k.

2.2.1 Construction de F_k

Pour obtenir F_k on pourra donc procéder de proche en proche en cherchant d'abords le sous espace Δ_1 de dimension 1 d'inertie maximal puis le sous espace Δ_2 de dimension 1 M-orthogonale à Δ_1 et d'inertie maximal, ...etc. La somme directe de ces sous espaces de dimension 1 est F_k tel que :

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

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

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

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

$$\begin{aligned} I_{\Delta_1} &= tr(VMP_1) \\ &= tr(VM\frac{a_1 a_1^t M}{a_1^t M a_1}) \\ &= \frac{1}{a_1^t M a_1} tr(VMa_1 a_1^t M) \\ &= \frac{1}{a_1^t M a_1} tr(a_1^t MVMa_1), \text{ car } tr(AB) = tr(BA) \\ &= \frac{a_1^t MVMa_1}{a_1^t M a_1}. \end{aligned}$$

donc

$$I_{\Delta_1} = \frac{a_1^t MVMa_1}{a_1^t M a_1}, \text{ tel que } (a_1^t MVMa_1) \in \mathbb{R}.$$

-Pour obtenir le maximum de $\frac{a_1^t MVMa_1}{a_1^t Ma_1}$, 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 $\frac{a_1^t MVMa_1}{a_1^t Ma_1} = \lambda \in \mathbb{R}$ alors

$$VMa_1 = \lambda a_1.$$

Remarque 2.2.2 *Le a_1 est un vecteur propre de VM avec M matrice régulière et λ est la plus grand valeur propre associé 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). Le droit (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.

Théorème 2.2.1 *Le sous-espace F_k de dimension k est engendré par les k vecteur propres de VM associés aux k plus grandes valeurs propres.[8]*

2.3 Eléments principaux

L'ACP repose essentiellement sur les trois éléments suivants :"axes principaux", "facteurs principaux", "composantes principales".

2.3.1 axes principaux

Ce sont les p vecteurs propres a_1, a_2, ..., a_p de la matrice VM, M-normés à 1 ie :

$$\begin{cases} VMa_j = \lambda_j a_j \quad j = \overline{1,p} \\ \quad \mid\mid a_j \mid\mid_M^2 = 1. \end{cases}$$

Remarque 2.3.1 *Les axes principaux a_j sont V^{-1} orthogonaux.*

2.3.2 Facteurs principaux

soit a_j un axe principal, le facteur principal u_j est un vecteur propre de MV tellque :

$$\begin{cases} MVu_j = \lambda_j a_j \quad \forall\ j = \overline{1,p} \\ \quad \|u_j\|_{M^{-1}}^2 = 1. \end{cases}$$

où $u_j = Ma_j \in \mathbb{R}^p$.

Propriété 2.3.1 *1. u_j sont $V-$orthogonaux.*

2. u_j sont $M^{-1}-$orthonormé.

2.3.3 Composantes principales

Ce sont les variables notées $c_j = (c_{1j}, c_{2j}, ..., c_{nj}) \in \mathbb{R}^n$ définies par les facteurs principaux comme suit :

$$c_j = Xu_j = XMa_j \qquad j = \overline{1,p.} \tag{2.1}$$

Et on a la formule suivante :

$$c_{ij} =< e_i, a_j >_M \text{ telque } \|a_j\| = 1.$$

Propriété 2.3.2 *1. Les composantes principales sont non corrélées deux à deux i.e :*

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

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

$$\text{var}(c_j) = \lambda_j. \tag{2.2}$$

3. Les composantes principales sont des combinaisons linéaires des variables initiales.

4. les composantes principales sont les vecteur propres de la matrice $X\ M\ X^tD$, qu'ils sont $D-$orthogonaux :

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

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 utilisé 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 matrice des corrélation R donc, les facteurs et les axes principaux sont les mêmes :

$$u_j = Ma_j = I_pa_j = a_j.$$

qui sont les vecteurs propres de la matrice de corrélation R associées aux valeurs propres où ces valeurs propres sont d'ordre décroissant ($\lambda_1 > \lambda_2 > ... > \lambda_p$) :

$$Ru_j = \lambda u_j \qquad \text{avec } \|u_j\|^2 = 1.$$

Donc, les composantes principales sont données par :

$$c_j = Zu_j \qquad \forall\ j = \overline{1,p}.$$

On peut remarquer que les composantes principales deviennent une combinaison linéaire des variables z_j (1.7)et $c_1 = Zu_1$ est la variance maximale.

Propriété 2.4.1 *Les composantes principales c_l sont les plus liées aux variables initiales x_1, x_2, ..., x_p, au sens de la somme des carrés des corrélations :*

$$\sum_{j=1}^{p} r^2(c_l, x_j) \text{ est maximal avec } l = \overline{1,k}.$$

Remarque 2.4.1 *L'ACP revient à remplacer les variables initiales x_1, x_2, ..., x_p qui sont corrélées entre elles, par des nouvelles variables c_1, c_2, ..., c_l appelées composantes principales qui sont des combinaisons linéaires des x_j non corrélées et de variance maximale.*

2.5 Interprétation et qualité de représentation

Le but de l'ACP est de construire de nouvelles variables, artificielles et fournit des représentations graphiques permettant de visualiser les relations entre variables ainsi que l'existence éventuelle de groupes d'individus et de groupes de variables et obtenir une représentation des individus dans un espace de dimension k plus faible que la dimension p. [7]

L'interprétation des résultats est une phase délicate qui doit se faire en respectant une démarche dont les éléments sont les suivants.

2.5.1 Interprétation des individus

On présente quelque définition sur l'interprétation des résultats pour les 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 :

$$QLT(F_k) = \frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} = \frac{\text{inertie projetée}}{I_g}.$$

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

Remarque 2.5.1 *Si* $QLT(F_k)$ *est proche de* 1, *la représentation sur* F_k *est bonne.*

Qualité de représentation d'un individu par rapport à l'axe : La qualité de représentation d'un individu e_i sur un axe l qui est donnée par :

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

où θ_{il} est l'angle formée entre le vecteur z_i et l'axe l.

Remarque 2.5.2

1. *-La qualité de représentation d'un individu* e_i *sur le plan* (l, l') *qui est donné par :*

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

Et on a aussi :

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

2. *-Si la valeur du* $\cos^2$ *est proche de* 1, *alors la représentation graphique de l'individu est de meilleure qualité.*

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

$$CTR_l(e_i) = \frac{p_i \ c_{il}^2}{\sum_{i=1}^{n} p_i \ c_{il}^2} = \frac{p_i \ c_{il}^2}{\lambda_l} \qquad \text{tel que } l = \overline{1,k}. \tag{2.3}$$

avec c_{il} : est la valeur de la composante principale l pour l'indice i.

Remarque 2.5.3 *1. La contribution d'un individu e_i est importante si :*

$$CTR_l(e_i) > p_i.$$

2. La contribution d'un groupe d'individu est égale à la somme des contributions des individus e_i et $e_{i'}$:

$$CTR_l(e_i, e_{i'}) = \frac{p_i \ c_{il}^2 + p_{i'} \ c_{i'l}^2}{\sum_{i=1}^{n} p_i \ c_{il}^2}. \qquad \forall i, i' = \overline{1,n} \ \textit{et} \ l = \overline{1,k}.$$

2.5.2 Interprétation des variables

On présente quelque définition sur l'interprétation des résultats pour les variables.

Qualité de représentation du nuage des variables : La méthode la plus naturelle pour donner une signitication à une composante principale c 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. [7]

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

$$r(c_l, x_j) = \sqrt{\lambda_l u_j}. \tag{2.4}$$

telle que λ_l est la valeur propre associée à c_l et, u_j la l ^{éme} composante principale.

Et on a aussi :

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

Remarque 2.5.4 *Les corrélations d'une variable x_j avec un couple de composantes principales c_1 et c_2 sont exprimées sur une cercle appelée cercle des corrélations*
de rayon 1

Contribution d'une variable : La contribution de la variable x_j à la composante c_l est donnée par la formule suivante :

$$CTR_l(x_j) = \frac{r^2(c_l, x_j)}{\sum_{j=1}^{p} r^2(c_l, x_j)} = \frac{r^2(c_l, x_j)}{\lambda_l}. \tag{2.5}$$

Et on a aussi :

$$CTR_l(x_j) = u_j^2.$$

2.5.3 Représentation d'élément supplémentaire

Représentation des individus supplémentaire : On note par $w = (w_1, w_2, ..., w_p)^t \in \mathbb{R}^p$ un nouvel individu appelé individu supplémentaire.Alors pour faire la représentation de cet individu supplémentaire sur le sous espace de projection F_k, il suffit de calculer les coordonnées de cet individu w dans le système des axes principaux (les combinaisons linéaires) comme suit :

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

Représentation d'une variable supplémentaire : On note par $t = (t_1, t_2, ..., t_p)^t \in \mathbb{R}^n$ un nouvel variable appelé variable supplémentaire.Alors pour faire la représentation de cet variable supplémentaire sur le sous espace de projection F_k, il suffit de calculer les

coordonnées de cet variable t dans le système des composantes principales. comme suit :

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

Exemple 2.5.1 *Le tableau ci-dessous représente les notes de* 9 *élèves dans* 5 *matières différentes :*

Sujet	math	Science	Français	latin	musique
Jean	6	6	5	5.5	8
Aline	8	8	8	8	9
Annie	6	7	11	9.5	11
Monique	14.5	14.5	15.5	15	8
Didier	14	14	12	12	10
André	11	10	5.5	7	13
Pierre	5.5	7	14	11.5	10
Brigitte	13	12.5	8.5	9.5	12
Evelyne	9	9.5	12.5	12	18

TAB. 2.1 – Notes des étudiants

On remarque que le nombre des variables p est égal à 5 *et le nombre des individus n est égale à* 9 *ce qui implique qu'on ne peut pas représenter les données. C'est pourquoi on va appliquer l'ACP sur ces données pour les représenter graphiquement dans un sous-espace de dimension* 2 *ou* 3.

Matrice de corrélation : *on peut construire la matrice de corrélation R suivant utilisant la formule (1.8) :*

$$R = \begin{pmatrix} 1.0000 & 0.9825 & 0.2267 & 0.4905 & 0.1112 \\ 0.9825 & 1.0000 & 0.3967 & 0.6340 & 0.0063 \\ 0.2267 & 0.3967 & 1.0000 & 0.9561 & 0.0380 \\ 0.4905 & 0.6340 & 0.9561 & 1.0000 & 0.0886 \\ 0.1112 & 0.0063 & 0.0380 & 0.0886 & 1.0000 \end{pmatrix},$$

telle que les lignes et les colonnes de cette matrice sont :math,science,français,latin et musique.

Valeurs propres et inerties *: Utilisant la formules (2.2) on peut calculer les valeurs propres et pour illustrer mieux l'importance et la qualité des notre valeurs*

propres, il va falloir leurs pourcentages et les pourcentages cumulées qui sont afficher dans le tableau suivant :

	val.prop	pourcentages(%)	pourcentages cumulés(%)
1	2.8618	57.24	57.24
2	1.1507	23.01	80.25
3	0.9831	19.66	99.91
4	0.0039	0.08	99.99
5	0.004	0.01	100

TAB. 2.2 – Valeurs propres et inerties

Commentaire : *D'après la table (2.2) et la figure (2.1) on remarque que les deux premiers axes traduisent* 80.25 % *de l'information disponible*

Compte ici qu'on pouvait s'en tenir uniquement au premier facteur. Mais c'est moins pratique pour les graphiques. c.à.d. une bonne qualité sur ce plan.

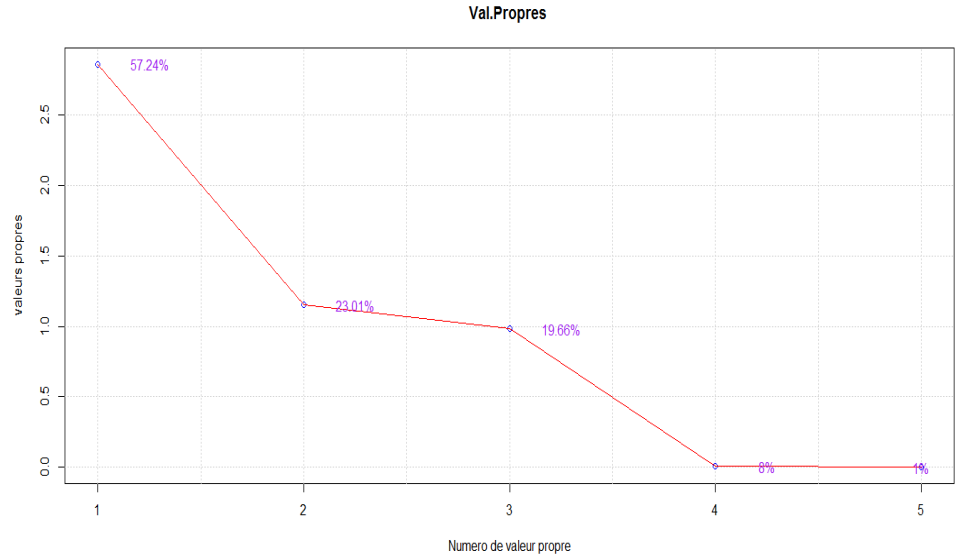

FIG. 2.1 – Représentation graphique de valeurs propres.

Nuage des individus : *Le tableau suivant fournit les composantes principales des individus* (c_1, c_2) *calculées par (2.1) et les Contributions des individus exprimées en pourcentages*(CTR_1, CTR_2) *qui sont calculées par(2.3)* :

Etudiants	c_1	c_2	CTR_1	CTR_2
Jean	-2.7857	0.6765	30.13	4.42
Aline	-1.2625	0.3303	6.19	1.05
Annie	-1.0167	-1.0198	4.01	10.04
Monique	3.1222	0.1659	37.85	0.27
Didier	1.9551	0.7879	14.84	5.99
André	-0.9477	1.2014	3.49	13.94
Pierre	-0.3250	-1.7548	0.41	29.73
Brigitte	0.6374	1.1298	1.58	12.33
Evelyne	0.6231	-1.5173	1.51	22.23

TAB. 2.3 – Composantes et Contribution des individus.

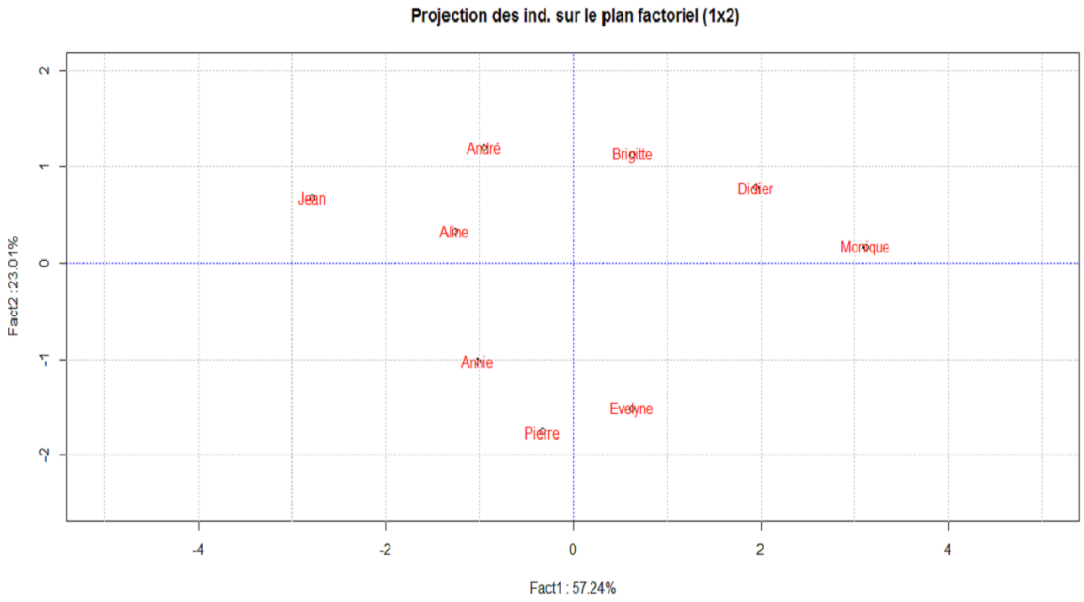

FIG. 2.2 – Représentation de nuage des individus.

Qualités de la représentation des individus :

	QLT_1	QLT_2
Jean	0.8855	0.0522
Aline	0.7920	0.0542
Annie	0.4784	0.4813
Monique	0.8786	0.0025
Didier	0.8515	0.1383
André	0.2465	0.3962
Pierre	0.0263	0.7671
Brigitte	0.1877	0.5898
Evelyne	0.0583	0.3458

TAB. 2.4 – Qualité de la réprésentation des individus.

Commentaire : *On remarque que "Didier" est bien représenté sur le plan avec une qualité de représentation égale à :*

$$QLT_{(1,2)}(Didier) = 0.85 + 0.13 = 0.98.$$

par contre "Evelyne" est très mal représenté :

$$QLT_{(1,2)}(Evelyne) = 0.39.$$

Nuage des variables : *Le tableau suivant fournit les composantes principales des variables (c_1, c_2) calculées par (2.4)et les Contributions des variables exprimées en pourcentages(CTR_1, CTR_2) qui sont calculées par (2.5)* :

Matières	c_1	c_2	CTR_1	CTR_2
Math	$0,8059$	0.5714	22.69	28.37
Sience	$0,8970$	0.4308	28.12	16.13
Français	$0,7581$	-1.6110	20.08	32.45
Latin	$0,9103$	-0.3975	28.95	13.73
Musique	$0,0667$	-0.3275	0.16	9.32

TAB. 2.5 – Composantes et Contribution des variables

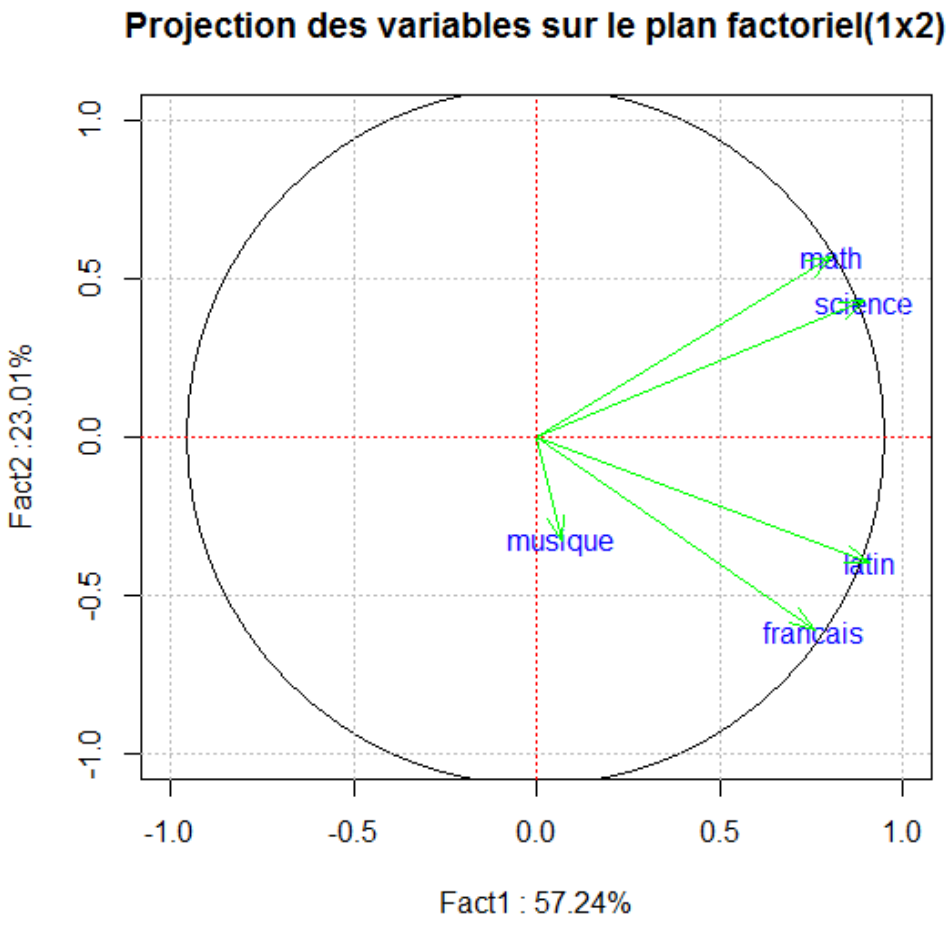

FIG. 2.3 – Répresentation des variables

Interprétation des résultats :

Individus :

axe2 *: On remarque que Monique et Didier sont en opposition avec Aline et Jean. Si on observe dans le tableau des données les notes des élèves (2.1) on trouve que Monique et Didier ont eu les meilleurs notes tandis que Jean et Aline ont eu des résultats très faibles.*

axe1 *: André et Brigitte sont en opposition avec Evelyne et Pierre. En effet les deux étudiants André et Brigitte ont presque le même niveau, ils maitrisent les matières scientifique plus que les autre matières ce qui n'est pas vrai pour Evelyne et Pierre qui travaillent beaucoup plus dans les matières de littératures. Pour les deux axes, Annie est toujours en opposition avec les bons élèves et d'après le graphe (2.2) on remarque aussi qu'il est littéraire plus que scientifique (il est proche de Evelyne et Pierre).*

Variables : *D'aprés le graph dans la figure 2.3 :*

axe1 *: On voit que les variables sont corrélées positivement et assez fortement entre elles, plus un élève obtient de bonnes notes dans une des matières plus il a une coordonnée importante sur l'axe 1.*

axe2 *: La variable musique n'est pas bien représentée, l'axe 2 oppose les matières littéraires aux matières scientifiques.*

Conclusion

En conclusion, dans ce mémoire nous avons étudie l'Analyse en Composantes Principales (ACP) qui est une des premières analyses factorielles, et certainement aujourd'hui l'une des plus employées. Elle est sent doute a la base de la compréhension actuelle des analyses factorielles .Son utilisation a cependant été plus tardive avec l'essor des capacités de calculs.

L'ACP est une méthode puissante pour synthétiser et résumer de vastes populations décrites par plusieurs variables quantitatives. 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 et son avantage le plus intéressant est de faire un traitement de façon simultanée d'un grand nombre des données.

Finalement, on dit que les représentations graphiques fournies par l'ACP sont simples et riches d'informations.

Bibliographie

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

[2] Boulet, H. (2013). Techniques Quantitatives. http ://www.educatim.fr/res/TQ_cours.

[3] Bounkhala, A. (2017). Méthodes ACP et AFC en statistiques et leurs applications, Tlemcen.

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

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

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

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

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

Annexe A : Logiciel *R*

2.6 Qu'est-ce-que le langage R ?

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

• **R** a été créé par Ross Ihaka et Robert Gentleman en 1996 du département de statistique de l'Université d'Auckland, en Nouvelle Zélande, et est maintenant développé par la R déve- loppement Core Team. Il est conçu pour pouvoir être utilisé avec les système d'exploitation Unix, Linux, Windows et MacOS.[5]

Le **R** est un application n'offrant qu'une invite de commande il basé sur la notion de vecteur, ce qui simpli. . . e les calculs mathématique et réduit considérablement le recours aux struc-

tures itératives (boucles for, ...ect). Programmes courts, en général quelques lignes de code seulement. Temps de développement très court.

Le logiciel R contient des packages de base trouvées dans toutes les versions, et des packages correspond avec quelques version.

Les packages et les fonctions utiles dans la réalisation de notre travail est :

Les packages : ade4.

Les fonctions : cov, cor, scale, dudi.pca, abline, symbols,....

Programmation :

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

On fait entrer les données da la façon suivante :

math<-c(6,8,6,14.5,14,11,5.5,13,9)

latin<-c(5.5,8,9.5,15,12,7,11.5,9.5,12)

science<-c(6,8,7,14.5,14,10,7,12.5,9.5)

français<-c(5,8,11,15.5,12,5.5,14,8.5,12.5)

musique<-c(8,9,11,8,10,13,10,12,18)

X<-data.frame(math,science,français,latin,musique)

V = cov(X) # Matrice de covariance V.

R =cor(X)# Matrice de corrélation R.

Z =scale(X)# Tableau standard Z.

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

vp=acp$eig # Valeurs propres λ.

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

plot(vp,type="n",ylab="valeurs propres",xlab="Numero de valeur propre",lwd=5,main="Val.

Propres (matrice de corr■l) ") # Graphique des vp.

Interprétation des axes et les graphes sur le plan factoriel (1 ; 2) :

Nuage des variables :

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

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

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

plot(c1,c2,type="n",ylab="Fact2 :23.01%",xlab="Fact1 : 57.24%",main="Projection des variables sur le plan factoriel(1x2)",xlim=c(-1,1),ylim=c(-1,1),col=1)

abline(h=0,v=0,lty=3,col=2)

text(c1,c2,row.names(acp$co),col="blue")

symbols(0,0,circles=2,col="red",ylab="Fact2 :23.01%",xlab="Fact1 : 57.24%",inches=2,add=T)

for(i in 1 :5){

arrows(0,0,c1[i],c2[i],angle=20,length=0.15,col="green")}

abline(h=-1 :1,v=-1 :1,lty=3,col="gray")

abline(h=-0.5 :0.5,v=-0.5 :0.5,lty=3,col="gray")

abline(h=0,v=0,lty=3,col=2)

Nuage des individus :

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

c_2=acp$li[,2] .# $1^{ére}$ composante principale c_1

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

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

plot(c_1,c_2,ylab="Fact2 :23.01%",xlab="Fact1 : 57.24%",main=" Projection des ind. sur le plan factoriel (1x2)",xlim=c(-5,5),ylim=c(-2.5,2),col=1)

abline(h=0,v=0,col="gray")

text(c_1,c_2,row.names(acp$li),col="red",cex=1) # Tracer le graphe des éleves celons les 2 axes.

Annexe B : 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$^{\text{ème}}$variable.
e_i	:	i$^{\text{ème}}$individu.
$\overline{x_j}$	:	Moyenne de la variable x_j.
$\mathbb{R}^p$	:	Espace des nombres réels de dimension p.
$\mathbb{R}^n$	:	Espace des nombres réels de dimension n.
D	:	Matrice de poids.
p_i	:	Poids.

I_n	:	Matrice d'identité de taille n.
$\mathbf{1}_n$	:	Vecteur unitaire d'identité de taille n.
g	:	Centre de gravité.
R	:	Matrice de corrélation.
$r_{jj'}$	:	Coefficient de corrélation.
V	:	Matrice de variance-covariance.
Z	:	Tableau des données centrés reduites.
Y	:	Tableau des données centrés.
$\text{cov}(.,.)$	:	Covariance.
$\text{var}(.,.)$	:	Variance.
$d(e_i, e_{i'})$	:	Distance entre e_i et $e_{i'}$.
M	:	Métrique.
I_g	:	Inertie totale.
tr	:	Trace d'une matrice.
$i.e$	:	C'est -à-dire.
f_i	:	Projection de l'individu e_i.
a	:	Axe principal.
u	:	Facteur principal.
c	:	Composante principale.
V_{proj}	:	Matrice de variance de nuage projeté.
F_k	:	Sous-espace de dimension k.
I_{proj}	:	Inertie de nuage projeté.
λ	:	Valeur propre.
$QLT(F_k)$	:	Qualité sur F_k.
$QLT_l(e_i)$	:	Qualité de e_i sur l'axe l.
$QLT_{l,l'}(e_i)$	:	Qualité de e_i sur plan(l, l').

$CTR_l(e_i)$: Contribution sur l'axe l de e_i.

$CTR_l(e_i, e_{i'})$: Contribution sur l'axe l de couple$(e_i, e_{i'})$.

$CTR_l(x_j)$: Contribution sur l'axe l de x_j.

Résumé

L'objectif principal de ce travail est d'étudier l'Analyse en Composantes Principales (ACP), qui est un 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 et interpréter. Il permet également de voir les corrélations entre les variables et les ressemblances entres les individus et obtenir une représentation simple du nuage des données plus proche de la réalité dans un espace de dimension faible, nous avons appliqué cette méthode a un exemple sur les notes de 9 élèves dans 5 matières différentes.

Mots clés : ACP, individus, variables, ressemblance et corrélation.

Abstract

The main objective of this work is to study Principal Component Analysis (PCA), which is an extremely powerful tool for synthesizing information, very useful when there is a large amount of quantitative data to be processed and interpreted. It also allows to see correlations between variables and similarities between individuals and to obtain a simple representation of the data cloud closer to reality in a small space; we applied this method to an example on the grades of 9 students in 5 different subjects.

Key words: PCA, individuals, variables, resemblance and correlation.

الملخص

ان الهدف الرئيسي من هذا العمل هو دراسة تحليل المكونات الرئيسية، وهي أداة قوية للغاية لتجميع المعلومات ومفيدة للغاية عندما يكون هناك كمية كبيرة من البيانات الكمية التي يتعين معالجتها وتفسيرها. كما يسمح لنا برؤية الارتباطات بين المتغيرات وأوجه التشابه بين الأفراد والحصول على تمثيل بسيط لسحابة البيانات أقرب إلى الواقع في مساحة صغيرة، قمنا بتطبيق هذه الطريقة على مثال على درجات 9 طلاب في 5 مواد مختلفة.

الكلمات المفتاحية: تحليل المكونات الرئيسية, الأفراد , المتغيرات , التشابه والارتباط.