République Algériennne 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 l'obtention du Diplôme :

MASTER en "Mathematiques"

Option : **Statistique**

Par

Settouti Fatima Zohra

Titre :

Analyse en composantes principales et applications

Membres du Comité d'Examen :

Pr.	BENATIA Fatah	UMKB	Encadreur
Pr.	MERAGHNI Djamel	UMKB	Président
Dr.	ROUBI Afaf	UMKB	Examinateur

28/06/2022

Dédicace

Aprés avoir terminé cette recherche, si **"Dieu"** le veut.

je dédie ce modeste travail à

Deux êtres chers à mon cour "**Mes parents**".

La lumière de ma vie qui m'ont tout donné et offert leur amour, encouragement, soutien aide ainsi que leur confiance et compréhension

pour faire de moi ce que je suis.

Aucune dédicace ne saurait être assez éloquente pour exprimer ce que vous méritez pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte.

A qui la vie nous a réunis, mes chères frères et soeurs.

Aux chers membres de la famille **"Settouti"**.

Atous ceux qui m'ont aidé de prés ou de loin à accomplir cette mémoire.

Et nous demandons à **"Dieu"** d'en faire un phare pour chaque étudiant de la connaissance.

A tout la promotion 2eme Master mathématique **2021-2022**.

Merci

$\mathcal{R}$emerciements

Louanges et remerciements à **"Dieu"**

J'adresse mes sincéres remerciements à mon superviseur, **Pr.BENATIA Fatah,** je remercie de m'avoir encadré, conseillé et aidé. Qu'**Allah**le récompense de tout le meilleur. Il a tout mon appréciation et respect.

J'aimerais présenter mes remerciements aux membres du jury, **Pr.MERAGHNI Djamel** et **Dr.ROUBI Afaf** pour le grand honneur qu'ils nous font en acceptant de juger ce travail.

Aussi, je souhait adresser mes sincéres remerciements à tous les professeurs qui m'ont aidé avec leurs conseils et orientations.

Un grand merci particulier à mes collègues et mes amies, pour les sympathiques moments qu'on a passés ensemble, je les remercie pour leur confiance, et leurs soutien moral au cours de ces années.

Je remercie mes trés chers parents pour leurs encouragements et leur soutien.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'analyse des données est un sous domaine des statistiques et un ensemble de techniques pour comprendre la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de le traduire par une structure plus simple [4], par utilisation de plusieurs méthodes, dont la plus importante est : l'analyse en composantes principales (ACP). Cette méthode est plusieurs utilisée dans un grand nombre de domaines domaines scientifiques, industriels et aussi en marketing et en méorologie...etc.

L'analyse en composantes principales (ACP), ou principal component analysis (PCA) en anglais, fait partie du groupe de méthodes descriptives multidimensionnelles appelées méthodes factorielles. Ces méthodes sont apparues au début des années 30, et ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzécri qui a introduit une approche géométrique et exploité les représentations graphiques [3]. L'ACP est utilisée pour extraire et visualiser les informations importantes contenues dans le tableau de données multivariées. L'objectif de l'ACP est alors de fournir des représentations synthétiques resumant les vastes ensembles de donnéees numériques essentiellement sous forme de visualisations graphiques planes et aussi de visualiser le plus fidélement possible, dans un espace de faible dimension, en déformant le moins possible la réalité [10]. C'est-à-dire qu'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 [6].

Le but de ce mémoire est de présenter et faire une description de l'ACP, de savoir comment résoudre le problème de la représentation des données multivariées, étudier la ressemblance entre les individus et la relation entre les variables par l'évaluation de leurs liaisons. Ce travail est organisé en deux parties : théorique et pratique (La partie théorique se compose en deux chapitres) :

Chapitre1 : On présente ic quelques définitions, propositions et propriétés de bases concernant ce domaine. En d'autres termes, on va faire une description des données et leurs caractéristiques.

Chapitre2 : Le deuxième chapitre est consacré à la représentation du principe et méthode de l'ACP ainsi que l' Interprétation des résultats obtenus.

Chapitre3 : Dans le dernier chapitre et à l'aide du logiciel R ,nous avons effectués l'ACP sur un table de données réels qui est "Les taux de différent délits dans 20 etats des Etats-unis", les commentaires des résultats obtenus par cette méthode pour enfin présentés à la fin de ce chapitre.

Chapitre 1

Préliminaires

L'analyse de données est un domaine issu du monde des statistiques qui vise à faire le lien entre les différentes données statistiques pour les classer, les décrire et les analyser de manière succincte.

L'objectif de l'analyse des données est d'extraire une information statistique qui permet de cerner plus précisément le profil de la donnée.

Dans ce chapitre, nous étudierons les données et leurs propriétés telles que 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

Les données sont généralement sous la forme d'un tableau rectangulaire à n lignes représentant les individus et à p colonnes correspondant aux variables.

1.1.1 Tableau des données et types de variables

On note X la matrice de dimension (n, p) contenant les observations : [10]

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Où x_{ij} est la valeur de l'individu i pour la variable j.

- L'individu e_i :La description du $i^{ème}$ individu (ligne de X).

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

- La variable X_j :La description du $j^{ème}$ variable (colonne de X).

$$X_j = (x_{1j}, ..., x_{nj})^t \, \epsilon \mathbb{R}^n, \quad \text{pour } j = \overline{1, p}$$

Types de variables

On distingue les variables qualitatives et quantitatives : les premières peuvent être nominales ou ordinales, et les secondes peuvent être discrètes ou continues. il est important, pour récolter et analyser les résultats d'une expérimentation, de connaître les types des variables qui y sont associées. Cela détermine notamment les analyses statistiques qu'il est permis d'effectuer.

1.1.2 Matrice des poids et centre de gravité

Matrice des poids

Si les données ont été recueillies à la suite d'un tirage aléatoire à probabilités égales, les n individus ont tous la même importance $1/n$, dans le calcul des caractéristiques de l'échantillon. Il n'en est pas toujours ainsi et il est utile pour certaines applications de travailler avec des poids p_i éventuellement différents d'un individu à l'autre (échantillons redressés, données regroupées...).

Ces poids, qui sont des nombres positifs de somme 1 comparables à des fréquences, sont regroupés dans une matrice diagonale D de taille n : [11]

$$D = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_n \end{pmatrix}, \quad \text{avec} \sum_{i=1}^{n} p_i = 1 \text{ et } p_i \geq 0.$$

Dans le cas le plus usuel des poids égaux,

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

tel que I_n est la matrice d'identité de taille n.

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

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} p_1 = p_1 \sum_{i=1}^{n} 1 = p_1 n = 1$$

Donc

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

Alors

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

■

centre de gravité

On définit le centre de gravité du nuage des individus par : [8]

$$g = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)^t = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ip} \end{pmatrix} \epsilon M(p, 1).$$

Le centre de gravité ou Point moyen g est le vecteur dont la $j^{éme}$ coordonnée g_j correspond à la valeur moyenne de la variable j sur les n individus.

Et on a la forme matricielle de g :

$$g = X^t D 1_n. \tag{1.1}$$

tel que 1_n est le vecteur unitaire de taille n.

Preuve.

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

$$= \begin{bmatrix} \sum_{i=1}^{n} p_i x_{i1} \\ \vdots \\ \sum_{i=1}^{n} p_i x_{ip} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = g.$$

■

1.1.3 Transformation et Matrice de variance-covariance-corrèlation

Types de transformation utilisées sur les données

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 :

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

la forme matricielle associée est alors donnée par :

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

Preuve. on a

$$1_n g^t = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} g_1 & g_2 & \cdots & g_p \end{pmatrix} = \begin{pmatrix} g_1 & \cdots & g_p \\ \vdots & \ddots & \vdots \\ g_1 & \cdots & g_p \end{pmatrix} \epsilon M(n,p)$$

Donc

$$X - g^t = \begin{pmatrix} x_{11} - g_1 & \cdots & x_{1p} - g_p \\ \vdots & \ddots & \vdots \\ x_{n1} - g_1 & \cdots & x_{np} - g_p \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} = Y \epsilon M(n,p).$$

■

Les colonnes de la matrice centrée Y sont de moyenne nulle :

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij} = 0.$$

Tableau centré réduit Il est souvent plus commode de travailler avec des variables réduites(c'est-à-dire dont l'écart-type vaut 1), afin de les rendre comparables entre elles et de gommer les effets d'échelle. En d'autres termes, on divise les coordonnées de chaque colonne y_{ij} par l'écart-type correspondant, on construit un tableau standard noté par Z de terme général :

$$z_{ij} = \frac{y_{ij}}{s_j}$$

Avec :$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - g_j\right)^2, j = 1, ..., p.$

La forme matricielle :

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

Où $D_{1/s}$ est la matrice poids définit par :

$$D_{1/s} = \begin{pmatrix} 1/s_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1/s_p \end{pmatrix} \epsilon M\left(p \times p\right).$$

Preuve.

$$YD_{1/s} = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} \begin{pmatrix} 1/s_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1/s_p \end{pmatrix}$$

$$= \begin{pmatrix} y_{11}/s_1 & \cdots & y_{1p}/s_p \\ \vdots & \ddots & \vdots \\ y_{n1}/s_1 & \cdots & y_{np}/s_p \end{pmatrix} = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{pmatrix} = Z\epsilon M\left(n, p\right).$$

■

Les colonnes de la matrice centrée-réduite Z sont de moyenne 0 et de variance égale 1 :

$$\bar{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij} = 0,\ var\left(z_j\right) = \frac{1}{n}\sum_{i=1}^{n}\left(z_{ij} - \bar{z}_j\right)^2 = 1.$$

Matrice de variance-covariance-corrélation

Définition 1.1.1 ***(Matrice de variance-covariance)***

Une matrice de variance-covariance est une matrice carrée caractérisant les inter-

actions (linéaires) entre p variables aléatoires notée par V :

$$V = \begin{pmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_p^2 \end{pmatrix}$$

Où s_j^2 est la variance de la variable x_j telle que :

$$s_j^2 = s_{jj} = var(x_j) = \sum_{i=1}^{n} p_i (x_{ij} - \bar{x}_j)^2.$$

Et $s_{jj'}$ est la covariance des variables x_j et $x_{j'}$ tel que :

$$s_{jj'} = cov\,(x_j, x_{j'}) = \sum_{i=1}^{n} p_i (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}), \text{ pour } j, j' = \overline{1, p}$$

La forme matricielle Matrice de variance-covariance :

$$V = Y^t D Y = X^t D X - g g^t. \tag{1.2}$$

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

$$\begin{aligned} V = Y^t D Y &= \left(X - I_n g^t\right)^t D \left(X - I_n g^t\right) \\ &= X^t D X - \left(X^t D I_n\right) g^t - g \left(I_n^t D X\right) + g I_n^t D I_n g^t \\ &= X^t D X - g g^t - g g^t + g g^t, \text{ car } I_n^t D I_n = \sum_{i=1}^{n} p_i = 1. \\ &= X^t D X - g g^t. \end{aligned}$$

■

Définition 1.1.2 ***(Matrice de corrélation)***

La matrice de corrélation est la matrice regroupant tous les coeffcients de corrélation entre les p variables prises deux à deux, on la note par R :

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & & \vdots \\ \vdots & & \ddots & r_{p-1p} \\ r_{p1} & \dots & r_{pp-1} & 1 \end{pmatrix}$$

Tel que R est symétrique et :

$$r_{jj'} = \frac{cov\,(x_j, x_{j'})}{s_j s_{j'}} = \frac{s_{jj'}}{s_j s_{j'}}$$

La forme matricielle de corrélation :

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

Preuve.

$$D_{\frac{1}{s}} V D_{\frac{1}{s}} = \begin{pmatrix} \frac{1}{s_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{s_p} \end{pmatrix} \begin{pmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_p^2 \end{pmatrix} \begin{pmatrix} \frac{1}{s_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{s_p} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \cdots & \frac{s_{1p}}{s_1 s_p} \\ \vdots & \ddots & \vdots \\ \frac{s_{p1}}{s_p s_1} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix} = R.$$

■

1.2 Nuage de points (individus et variables)

1.2.1 Nuage de points des individus

Chaque individu e_i étant un point défini par p coordonnées est considéré comme un élément d'un espace vectoriel $\mathbb{R}^p$ appelé l'espace des individus. L'ensemble des n individus est alors un « nuage» de points dans $\mathbb{R}^p$ et g en est le centre de gravité. [11]

L'espace $\mathbb{R}^p$ est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus.

Le rôle de la métrique

La distance euclidienne entre deux individus (e_i, e_j) dans l'espace $\mathbb{R}^p$ se calcule facilement par la formule de pythagore(le carré de la distance est la somme des différences des coordonnées), et s'écrit alors :

$$d^2(e_i, e_{i'}) = (x_{i1} - x_{i'1})^2 + ... + (x_{ip} - x_{i'p})^2 = \sum_{j=1}^{p} (x_{ij} - l : x_{i'j})^2 , \text{ pour } i, i' = \overline{1, n}$$

Cette définition suppose que les dimensions sont de même nature, i.e que les mesures sont faites dans la même unité.

Comment pouvons-nous résoudre le problème de l'unité ?

Pour le résoudre, on choisit de transformer les données en données centrées et réduites.

De plus, la formule de Pythagore n'est valable que si les axes sont perpendicu-

laires, ce que l'on conçoit aisément dans l'espace physique. Mais en statistique ce n'est que par pure convention que l'on représente les caractères par des axes perpendiculaires.

On utilisera donc la formulation générale suivante : la distance entre deux individus e_i et $e_{i'}$ est définie par la forme quadratique : [11]

$$d^2\left(e_i, e_{i'}\right) = \left(e_i - e_{i'}\right)^t M \left(e_i - e_{i'}\right)$$

Où M est une matrice symétrique de taille p définie positive. L'espace des individus est donc muni du produit scalaire :

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

Quand on travaille sur le tableau y on utilise la métrique $M = D_{\frac{1}{s^2}}$ mais quand on travaille sur le tableau Z on utilise la métrique $M = I_p$ telle que I_p est la matrice identité d'ordre p [7].

L'inertie

On définit l'inertie totale d'un nuage de points par la moyenne pondérée des carrés des distances des points $(e_i)_{1\leq i\leq n}$ du centre de gravité g : [10]

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

L'inertie dans un point α quelconque est définie par :

$$I_\alpha = \sum_{i=1}^{n} p_i d^2\left(e_i, \alpha\right) = \sum_{i=1}^{n} p_i \left(e_i - \alpha\right)^t M \left(e_i - \alpha\right) = \sum_{i=1}^{n} p_i \left\|e_i - \alpha\right\|^2$$

Il existe un autre terme de l'inertie définie par :

$$I_g = tr\left(MV\right). \tag{1.3}$$

Si $M = I_p : I_g = tr\left(I_P V\right) = tr\left(V\right) = \sum_{j=1}^{p} s_j^2$.

si $M = D_{1/S^2} : I_g = tr\left(D_{1/S^2} V\right) = tr\left(D_{1/S} V D_{1/S}\right) = tr\left(R\right) = p$.

Inertie par rapport à un sous-espace vectoriel

Notons F_k un sous-espace de projection et Δ le sous-espace vectoriel passant par le centre de gravité g. L'inertie du nuage de points par rapport à ce sous-espace vectoriel F_k est définie par [8] :

$$I_\Delta = \sum_{i=1}^{n} p_i d^2\left(e_i, f_i\right), \text{ pour } i = \overline{1,n}.$$

Où f_i désigne la projection orthogonale de e_i sur F_k.

Décomposition de l'inertion

Considérons une décomposition de F_k en deux sous-espaces vectoriels supplémentaires, orthogonaux, et passant par g ; c'est à dire :

$$F_k = \Delta \oplus \Delta^{\perp}.$$

Où $\Delta^{\perp}$ désigne l'espace orthogonal à Δ.

$$I_{F_k} = I_\Delta + I_{\Delta^{\perp}} = I_g.$$

Si $F_k = \Delta_1 \oplus \Delta_2 \oplus \ldots \oplus \Delta_k$ (Théorème de Huygens) alors :

$$I_{F_k} = I_{\Delta_1^{\perp}} + I_{\Delta_2^{\perp}} + \ldots + I_{\Delta_k^{\perp}}.$$

1.2.2 Nuage de points des variables

Chaque variable x_j est définie par n coordonnées, on la considère alors comme un vecteur d'un espace à n dimensions appelé espace des variables. L'ensemble des p variables constitue un nuage de points dans $\mathbb{R}^n$ appelé nuage des variables [10].

Métrique des variables

Pour le calcul des vecteur d'un espace à n dimensions appelé espace des variables. Pour le calcul des « distances » entre variables, on utilise la métrique D_p diagonale des poids qui possède, lorsque les variables sont centrées, les propriétés suivantes :

– Le produit scalaire de deux variables x_j et $x_{j'}$ [10] :

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

Si les deux variables sont centrées alors :

$$\langle x_j, x_{j'} \rangle_D = cov\,(x_j, x_{j'}) = s_{jj'}, \text{ pour } j, j' = \overline{1, p}.$$

– Le carré de la norme d'une variable centrée est égal alors à sa variance :

$$\|x_j\|_D^2 = s_j^2, \text{ pour } j, j' = \overline{1, p}.$$

– Dans un espace euclidien en notant $\theta_{jj'}$ l'angle entre deux variables centrées on

a :

$$\cos(\theta_{jj'}) = \frac{\langle x_j, x_{j'} \rangle_D}{\|x_j\|_D \|x_{j'}\|_D} = \frac{s_{jj'}}{s_j s_{j'}} = r(x_j, x_{j'}), \text{ pour } j, j' = \overline{1, p}.$$

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

Chapitre 2

Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (corrélées) en nouvelles variables indépendantes les unes des autres (non corrélées). Ces nouvelles variables sont nommées "composantes principales" ou "axes".

Le principe de la méthode est d'obtenir une représentation approchée du nuage des n individus dans un sous-espace de dimension faible [2].

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 difficile à mettre en évidence les relations globales existant entre les variables ,car impossibles à visualiser.

On cherche une représentation des n individus , dans un sous-espace F_k de $\mathbb{R}^p$ de dimension k (k petit égal $1, 2$ ou 3, par exemple un plan).

Autrement dit, on cherche à définir k nouvelles variables combinaisons linéaires des p variables initiales qui feront perdre le moins d'information possible [5].

2.1.1 Méthode de l'ACP

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

Définition 2.1.1 *[5]*

1. On définit P une matrice (operateur de projection) M-orthogonal sur l'espace Δ_k tel que :

$$P = \alpha \left(\alpha^t M \alpha\right)^{-1} \alpha^t M, \tag{2.1}$$

où $\alpha \in R^p$ est un vecteur directeur de Δ.

Et vérifie les deux condition suivantes :

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

. $MP = P^t M$ (P est M-symétrique).

2. Le nuage projeté associ au tableau sera donnée par :

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

- Le centre de gravité projeté est :

$$g_{proj} = Pg.$$

Preuve. D'après (1.1) :

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

■

- La matrice de covariance du tableau X_{proj} est définit par :

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

Preuve. D'après (1.2) :

$$\begin{aligned} V_{proj} &= X_{proj}^t D X_{proj} - g_{proj} g_{proj}^t \\ &= PX^t DXP^t - Pgg^t P^t \\ &= P\left(X^t DX - gg^t\right) P^t \\ &= PVP^t \end{aligned}$$

■

- L'inertie du nuage projeté est définit par :

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

Preuve. D'après (1.3) :

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

■

Projection des individus sur ce nouveau sous espace

D'après le premier chapitre on a :

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

Comment construire le sous espace F_k ?

On construit F_k de proche en proche en cherchant d'abord le sous espace Δ_1 de dimension 1 et $I_{\Delta_1^{\perp}}$ soit maximale, puis le sous espace Δ_2 de dimension 1 M-orthogonal à Δ_1 et tel que $I_{\Delta_2^{\perp}}$ soit maximale, et ainsi de suite. La somme directe de ces sous espaces nous donne F_k [10].

2.1.2 Eléments de l'ACP et propriétés

L'ACP repose essentiellement sur les trois éléments suivants [5] :

- Les axes qu'elles déterminent : « axes principaux».
- Les formes linéaires associées : « facteurs principaux».
- Les variables associées «composantes principales ».

Axes principaux

Nous devons chercher la droite de $\mathbb{R}^p$ passant par g maximisant l'inertie du nuage projeté sur cette droite.

Recherche du premier axe principal On cherche donc dans $\mathbb{R}^p$ un axe Δ_1, passant par g, et tel que I_{Δ_1} soit minimum, ou de façon équivalente tel que $I_{\Delta_1^\perp}$ soit maximum.

Notons α_1 le vecteur directeur unitaire de l'axe Δ_1 : On cherche donc α_1 tel que $I_{\Delta_1^\perp}$ soit maximum sous la contrainte $\|\alpha_1\|^2 = 1$. On obtient alors le problème d'optimisation sous contrainte suivant [8] :

$$\begin{cases} \max_{u_1} I_{\Delta_1^\perp} \\ \|\alpha_1\|_M^2 = 1 \end{cases} \iff \begin{cases} \max_{u_1} tr\left(VMP_1\right) \\ \alpha_1^t M \alpha_1 = 1 \end{cases}.$$

En remplaçant le projecteur $P_1 = \alpha_1 \left(\alpha_1^t M \alpha_1\right)^{-1} \alpha_1^t M$ dans la définition de l'inertie du nuage projeté, on obtient :

$$\begin{aligned} I_{\Delta_1^\perp} &= tr\left(VMP_1\right) \\ &= tr\left(VM\left(\alpha_1\alpha_1^t M\ /\alpha_1^t M\alpha_1\right)\right), \text{ car } P_1 = \alpha_1\alpha_1^t M\ /\alpha_1^t M\alpha_1 \text{ et } \left(\alpha_1^t M\alpha_1\right) \in \mathbb{R} \\ &= tr\left(VM\alpha_1\alpha_1^t M\ \right)/\alpha_1^t M\alpha_1 \\ &= tr\left(\alpha_1^t MVM\ \alpha_1\right)/\alpha_1^t M\alpha_1 \\ &= \alpha_1^t MVM\ \alpha_1/\alpha_1^t M\alpha_1 \end{aligned}$$

L'inertie du nuage projeté sur Δ_1 est

$$I_{\Delta_1^\perp} = \frac{\alpha_1^t MVM\ \alpha_1}{\alpha_1^t M\alpha_1}$$

Pour résoudre le problème de maximisation sous contraintes, il suffit d'annuler la dérivée de cette expression par rapport à α_1. En appliquant la régle de dérivation d'une forme quadratique par rapport à un vecteur, on obtient : [11]

$$VM\alpha_1 = \frac{\alpha_1^t MVM\ \alpha_1}{\alpha_1^t M\alpha_1}\alpha_1.$$

On pose $\frac{\alpha_1^t MVM\ \alpha_1}{\alpha_1^t M\alpha_1} = \lambda_1 \in \mathbb{R}$.

Alors

$$VM\alpha_1 = \lambda_1\alpha_1$$

α_1 est donc vecteur propre de matrice VM associée à la plus grande valeure propre (λ_1).

Recherche des axes suivants Une fois que le premier axe a été identifié, on cherche l'axe Δ_2, orthogonal à Δ_1 et tel que l'inertie I_{Δ_2} soit minimale, ou de façon équivalente telle que $I_{\Delta_2^\perp}$ soit maximale. En notant α_2 le vecteur directeur unitaire de Δ_2 on doit alors résoudre le système suivant [8] :

$$\begin{cases} \max_{u_2} tr\,(VMP_2) \\ \|\alpha_2\|_M^2 = 1 \\ \alpha_2 \perp \alpha_1 \iff \langle \alpha_1, \alpha_2 \rangle = \alpha_2^t \alpha_1 = 0 \end{cases}$$

on obtient

$$VM\alpha_2 = \lambda_2\alpha_2$$

α_2 est donc vecteur propre de matrice VM associée à la deuxième plus grande valeure propre (λ_2).

On raisonne de même pour trouver les axes suivants, dont les vecteurs directeur

unitaires sont tous des vecteurs propres de la matrice VM, associés aux valeurs propre ordonnées par ordre décroissant ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$).

La matrice VM étant M-symétrique, elle possède bien p vecteurs propres qui forment une base orthogonale de R^P ($\alpha_1 \perp \alpha_2 \perp ... \perp \alpha_p$).

Les axes Δ_1, Δ_2, ..., Δ_p sont deux à deux M-orthogonaux appelées axes factoriels, ou axes principaux d'inertie ($\Delta_1 \perp \Delta_2 \perp ... \perp \Delta_p$).

donc

$$\begin{cases} VM\alpha_j = \lambda_j \alpha_j, \quad j = \overline{1, P}. \\ \|\alpha_j\|_M^2 = 1. \end{cases}$$

1. Les axes principaux α_j sont V^{-1}-orthogonaux.
2. Les axes principaux α_j sont M-orthonormé.

Facteurs principaux

Le facteur principal noté u_j associé à l'axe principal α_j est définit par [11] :

$$u_j = M\alpha_j \in \mathbb{R}^p.$$

u_j est un vecteur propre de la matrice MV associé à la valeurs propre λ_j tel que

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

1. Les facteurs principaux u_j sont V-orthogonaux.
2. Les facteurs principaux u_j sont M^{-1}-orthonormés.

Composantes principales

Les Composantes principales notées c_j sont "les nouvelles variables" définies par les facteurs principaux :

$$c_j = XM\alpha_j = Xu_j \text{ pour } j = \overline{1,p}.$$

Et c_j est le vecteur renfermant les coordonnées des projection M-orthogonales des n individus sur l'axe défini par α_j avec α_j unitaire :

$$c_{ij} = \langle e_i, \alpha_j \rangle_M \text{ tel que } \|\alpha_j\| = 1.$$

Les composantes principales vérifient :

1. $Var[c_j] = \lambda_j$.
2. $Cov[c_j, c_{j'}] = 0, j \neq j'$.
3. $XMX^tDc_j = \lambda_j c_j$. ($c_j$ est le vecteur proupre de la matrice XMX^tDc associé à la valeurs propre λ_j).

   Preuve. 1.

$$\begin{aligned} Var[c_j] &= c_j^t D c_j - g_{c_j} g_{c_j^t} \\ &= u_j^t X^t D X u_j - u_j^t X^t D 1_n 1_n^t D X u_j \\ &= u_j^t \left(X^t D X - X^t D 1_n 1_n^t D X \right) u_j \\ &= u_j^t \left(X^t D X - g g^t \right) u_j \\ &= u_j^t V u_j \text{ (d'après (1.2))} \end{aligned}$$

   On a $MVu_j = \lambda_j u_j \iff Vu_j = M^{-1}\lambda_j u_j$ donc :

$$\begin{aligned} Var\left[c_j\right] &= \lambda_j u_j^t M^{-1} u_j \\ &= \lambda_j \text{ car } \left(u_j^t M^{-1} u_j = \|u_j\|^2_{M^{-1}} = 1\right). \end{aligned}$$

2. Les composantes principales sont non corrélées deux à deux, car les axes associés sont orthogonaux.

3. On a

$$MVu_j = \lambda_j u_j \iff MX^t DXu_j = \lambda_j u_j.$$

En multipliant à gauche par X et on remplaçant Xu_j par c_j on trouve :

$$XMX^t Dc_j = \lambda_j c_j$$

■

Pour résumer :

Axes principaux α	$VM\alpha = \lambda\alpha$	M-orthonormés
Facteurs principaux u	$VMu = \lambda u$	M^{-1}-orthonormés
Composantes principales c	$XMX^tDc = \lambda c$	D-orthogonales

TAB. 2.1 – Eléments de l'ACP

Le choix de la métrique est toujours délicat en générale, or en pratique on va travailler avec un tableau centré réduit Z ce qui implique que $M = I_p$ et donc la matrice de variance covariance ne sera d'autre que la matrice de corrélation R. Dans ce cas $C_j = Zu_j$ sera une combinaison linéaire des variables centrés réduites, et la première Composante principale ayant une variance maximale [11].

2.2 Interprétation des résultats

L'interprétation des résultats est une phase délicate et important qui doit se faire.

2.2.1 Contribution des axes à l'inertie totale

Par le théorème de Huygens, on a :

$$I_g = I_{\Delta_1^\perp} + I_{\Delta_2^\perp} + \ldots + I_{\Delta_p^\perp} = \lambda_1 + \lambda_2 + ... + \lambda_p.$$

L'inertie expliquée par l'axe Δ_j est :

$$I_{\Delta_j^\perp} = \lambda_j$$

et le pourcentage d'inertie expliquée par cet axe, aussi appelé "contribution relative", est égale à [7] :

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p}.$$

De la même façon, on peut définir le pourcentage d'inertie expliquée par le sous espace

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

par :

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{I_g}.$$

Ce pourcentage nous permet de déterminer le nombre d'axes retenus on calcul.si les dernières valeurs propres sont faibles, les axes associés expliqueront une faible part de l'inertie totale et pourront être négligés. Dans la pratique,on représentera alors le nuage de points dans un sous-espace de dimension $K < p$; où K est choisi

tel que l'inertie expliquée par les K premiers axes soit proche de 1 ; c'est à dire proche de 100/100 [8].

2.2.2 Interprétation des individus

Nous allons dans cette partie présenter une Interprétation des résultats pour les individus :

Qualité de représentation des individus

Pour interpréter correctement la proximité entre deux points projetés il faut s'assurer au prèalable que ces points sont bien représentés dans le sous-espace sur lequel on les a projetés. Pour cela, on s'intéresse à l'angle entre le point e_i et le sous-espace sur lequel on projette, et plus précisément au cosinus carré de cet angle. Ainsi, si ce cosinus carré est proche de 1, cela veut dire que l'individu est bien représenté par sa projection sur le sous-espace. Par suite. Le cosinus au carrée de l'angle θ_{ij} entre e_i et Δ_j est donné par [8] :

$$\cos^2(\theta_{ij}) = \frac{\langle e_i, \alpha_j \rangle_M^2}{\|e_i\|^2} = \frac{c_{ij}^2}{\|e_i\|^2}$$

avec c_{ij} est la valeur de la composante principale j pour le $i^{éme}$ individu.

De même, on peut calculer le cosinus au carrée de l'angle $\theta_{ijj'}$ entre e_i est sous-espace engendré par les axes Δ_j et $\Delta_{j'}$ par :

$$\cos^2(\theta_{ijj'}) = \cos^2(\theta_{ij}) + \cos^2(\theta_{ij'}).$$

Interprétation des nouveaux axes en fonction des individus

On rappelle que l'inertie expliquée par un axe Δ_j est :

$$I_{\Delta_j^{\perp}} = \sum_{i=1}^{n} p_i d^2\left(e_i, f_{ij}\right) = \sum_{i=1}^{n} p_i \left\langle e_i, \alpha_j \right\rangle_M^2 = \sum_{i=1}^{n} p_i c_{ij}^2 = \lambda_j, \ j = \overline{1,k}.$$

Si c_1 est trés corrélé avec X_j cela veut dire que les individus ayant une forte coordonnée positive sur le premier axe sont caractérisés par une valeur de X_j nettement supérieur à la moyenne donc il est très utile de calculer pour chaque axe la contribution apportée par les divers individus [11].

Contribution absolue La contribution absolue est définit par [5] :

$$Ca\left(i,j\right) = p_i c_{ij}^2.$$

Contribution relative La contribution relative est donnée par [5] :

$$Cr\left(i,j\right) = \frac{ca\left(i,j\right)}{\lambda_j} = \frac{p_i c_{ij}^2}{\lambda_j}, j = \overline{1,k}. \tag{2.2}$$

2.2.3 Interprétation des variables

Passant maintenant : l'Interprétation des résultats pour les variables.

Qualité de représentation du nuage des variables

Pour donner une signification à la composante principale il faut la reliée aux variables initiales X_j, en calculant le coefficient de corrélation $r(c, X_j)$ est on

s'intéresse au plus fort coefficiont en valeur absolue [7].

$$r(c, X_j) = r(c, z_j) = \frac{cov\,(c, z_j)}{s_c s_{z_j}} = \frac{c^t D z_j}{\sqrt{\lambda}}.$$

Car z_j sont centrées réduites et $Var\,(c) = \lambda \Longrightarrow s_c = \sqrt{\lambda}$.

et comme on a $c = Zu$ avec u le facteur principal associé à c et le vecteur propre de R matrice de corrélation associé à la valeur propre λ

$$r(c, z_j) = \sqrt{\lambda} u_j.$$

Preuve.

$$\begin{aligned} r(c, z_j) &= \frac{c^t D z_j}{\sqrt{\lambda}} = \frac{1}{\sqrt{\lambda}} u^t z_j^t D z_j \\ &= \frac{1}{\sqrt{\lambda}} z_j^t D z_j u \\ &= \frac{1}{\sqrt{\lambda}} R u \\ &= \frac{1}{\sqrt{\lambda}} \lambda u \\ &= \sqrt{\lambda} u \end{aligned}$$

■

1. Les corrélations d'une variable X_j avec un couple de composantes principales c_1 et c_2 sont exprimées sur un cercle appelé cercle des corrélations de rayon 1.

2. Si

$$r^2(c_1, X_j) + r^2(c_2, X_j) \simeq 1$$

alors

$$\sum_{k=3}^{p} r^2(c_k, X_j) \simeq 0,$$

par conséquent $r(c_k, X_j), k = 3, ..., p$ sont aussi prochent de zéro. Ce ci explique que la variable X_j a une forte corrélation avec les premieres composantes principales et non correlée avec le reste des composantes.

La quantité $r(c_k, X_j)$ donne la qualité de représentation de la variable j sur l'axe Δ_k. Plus elle est proche de 1 en valeur absolue, plus la variable est bien représentée par l'axe k. Le signe de la corrélation permet de savoir si la variable contribue positivement ou négativement à la définition de l'axe k.

Plus une variable est proche du cercle de corrélation, mieux elle est représentée par le plan considéré.

Contribution d'une variable

La contribution de la variable X_j à la composante principale c_k est donne par la formule suivante [11] :

$$Cv = \frac{r^2\left(c_k, X_j\right)}{\sum_{j=1}^{p}\left(c_k, X_j\right)} = \frac{r^2\left(c_k, X_j\right)}{\lambda_k},$$

on peut aussi définir la contribution comme suit :

$$Cv = u_{jk}^2$$

2.3 En pratique

Dans la pratique, on retient un nombre $k < p$ d'axes principaux, sur lesquels on va projeter notre nuage de points. On doit alors proposer une interprétation des nouveaux axes obtenus, ou de façon équivalente des composantes principales. 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 :[8]

1. centrage de la matrice de données.
2. réduction de la matrice de données si nécessaire.
3. calcul des valeurs propres de V, et choix du nombre d'axes à retenir en fonction du pourcentage d'inertie que l'on souhaite conserver.
4. interprétation des nouvelles variables, à l'aide des cercles de corrélations (attention, 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.

Chapitre 3

Application sous R

Ce chapitre traite la mise en oeuvre dans l'environnement R, ce logicial utilisé pour le traitement de données, nous sera utile pour montrer comment effectuer une ACP sur des données réelles à l'aide de ce logiciel.

3.1 Les packages

Pour effectuer l'analyse, on utilise plusieurs packages de ce logiciel R :

- Le package FactoMineR (analyse de données exploratoire multivariée et fouille de données).

-Le package ade4 (analyse de data écologique : méthodes exploratoires et euclidiennes en sciences de l'environnement).

3.1.1 Les fonctions

Dans cette partie, on décrit la plupart des fonctions utilisées dans l'application ACP :

- PCA() : Package FactoMineR.

- dudi.pca() :Package ade4.

- cov,cor,scale,plot,abline,symbols,....

3.1.2 Installation des packages

La première étape consiste à installer et charger ces packages comme suit :

- install.packages("FactoMineR") et library("FactoMineR").

- install.packages("ade4") et library("ade4").

3.2 Exemple sur le Tableau des données

Le tableau ci-dessous représente les taux de différents délits (qui sont au nombre de six) commis pour 100000 habitants dans 20 Etats des Etats-unis. Ces données peuvent être mises dans un tableau individu-variable (le nombre des variables p est égal à 6 et le nombre des individus n est égale à 20) [9].

3.2.1 Tableau des données

1 : Alabma	2 : Alaska	3 : Arizona	4 : Arkansas	5 : California
6 : Colorado	7 : Connecticut	8 :Delaware	9 : Florida	10 : Georgia
11 : Hawaii	12 : Idaho	13 : Illinois	14 : Indiana	15 : Iowa
16 : Kansas	17 : Kentucky	18 : Louisiana	19 : Maine	20 : Maryland

ETAT	Meurtre	Rapt	Vol	Attaque	Viol	Larcin
1	14.2	25.2	96.8	278.3	1135.5	1881.9
2	10.8	51.6	96.8	284	1331.7	3369.8
3	9.5	34.2	138.2	312.3	2346.1	4467.4
4	8.8	27.6	83.2	203.4	972.6	1862.1
5	11.5	49.4	287	358	2139.4	3499.8
6	6.3	42	170.7	292.9	1935.2	3903.2
7	4.2	16.8	129.5	131.8	1346	2620.7
8	6	24.9	157	194.2	1682.6	3678.4
9	10.2	39.6	187.9	449.1	1859.9	3840.5
10	11.7	31.1	140.5	256.5	1351.1	2170.2
11	7.2	25.5	128	64.1	1911.5	3920.4
12	5.5	19.4	39.6	172.5	1050.8	2599.6
13	9.9	21.8	211.3	209	1085	2828.5
14	7.4	26.5	123.2	153.5	1086.2	2498.7
15	2.3	10.6	41.2	89.8	812.5	2685.1
16	6.6	22	100.7	180.5	1270.4	2739.3
17	10.1	19.1	81.1	123.3	872.2	1662.1
18	15.5	30.9	142.9	335.5	1165.5	2469.9
19	2.4	13.5	38.7	170	1253.1	2350.7
20	8	34.8	292.1	358.9	1400	3177.7

TAB. 3.1 – Les taux de différent délits dans 20 etats des Etats-unis

Programmation :

On fait entrer les données de la façon suivante :

> Meurtre <-c(14.2,10.8,9.5,8.8,11.5,6.3,4.2,6,10.2,11.7,7.2,5.5,9.9,7.4,2.3,6.6,

10.1,15.5,2.4,8)

> Rapt <-c(25.2,51.6,34.2,27.6,49.4,42,16.8,24.9,39.6,31.1,25.5,19.4,21.8,26.5,

10.6,22,19.1,30.9,13.5,34.8)

> Vol <-c(96.8,96.8,138.2,83.2,287,170.7,129.5,157,187.9,140.5,128,39.6,211.3,

123.2,41.2,100.7,81.1,142.9,38.7,292.1)

> Attaque <-c(278.3,284,312.3,203.4,358,292.9,131.8,194.2,449.1,256.5,64.1,172.5,

209,153.5,89.8,180.5,123.3,335.5,170,358.9)

> Voil <-c(1135.5,1331.7,2346.1,972.6,2139.4,1935.2,1346,1682.6,1859.9,1351.1,

1911.5,1050.8,1085,1086.2,812.5,1270.4,872.2,1165.5,1253.1,1400)

> Larcin <-c(1881.9,3369.8,4467.4,1862.1,3499.8,3903.2,2620.7,3678.4,3840.5,2170.2,

3920.4,2599.6,2828.5,2498.7,2685.1,2739.3,1662.1,2469.9,2350.7,3177.7)

> X <-data.frame (Meurtre, Rapt, Vol, Attaque, Voil, Larcin)

Centre de gravité Centre de gravité est obtenue par les commandes suivantes :

$> g = colMeans(X)$ # Centre de gravité g.

$> round(g, 3)$

Meurtre	Rapt	Vol	Attaque	Viol	Larcin
8.405	28.325	134.320	230.880	1400.365	2911.300

Tableau centré réduit Est obtenue par les commandes suivantes :

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

$> round(Z, 3)$

$$Z = \begin{bmatrix} 1.647 & -0.282 & -0.530 & 0.469 & -0.610 & -1.311 \\ 0.681 & 2.098 & -0.530 & 0.525 & -0.158 & 0.584 \\ 0.311 & 0.530 & 0.055 & 0.804 & 2.179 & 1.981 \\ 0.112 & -0.065 & -0.722 & -0.272 & -0.985 & -1.336 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Matrice de covariance On peut construire la matrice de covariance V en utilisant les commandes suivantes :

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

$> round(V, 3)$

$$V = \begin{bmatrix} 12.386 & 20.862 & 88.115 & 208.721 & 154.280 & -292.633 \\ 20.862 & 123.065 & 463.940 & 842.182 & 2878.246 & 4507.631 \\ 88.115 & 463.940 & 5017.466 & 4439.163 & 16297.411 & 25135.534 \\ 208.721 & 842.182 & 4439.163 & 10244.503 & 20565.221 & 28026.129 \\ 154.280 & 2878.246 & 16297.411 & 20565.221 & 188452.172 & 294978.875 \\ -292.633 & 4507.631 & 25135.534 & 28026.129 & 294978.875 & 616933.979 \end{bmatrix}$$

Matrice de corrélation Est obtenue par les commandes suivantes :

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

$> round(V, 3)$

$$R = \begin{bmatrix} 1.000 & 0.534 & 0.353 & 0.586 & 0.101 & -0.106 \\ 0.534 & 1.000 & 0.590 & 0.750 & 0.598 & 0.517 \\ 0.353 & 0.590 & 1.000 & 0.619 & 0.530 & 0.452 \\ 0.586 & 0.750 & 0.619 & 1.000 & 0.468 & 0.353 \\ 0.101 & 0.598 & 0.530 & 0.468 & 1.000 & 0.865 \\ -0.106 & 0.517 & 0.452 & 0.353 & 0.865 & 1.000 \end{bmatrix}$$

3.3 Interprétation des résultats

On va appliquer l' ACP sur les données précédentes pour les représenter graphiquement dans un sous-espace de dimension 2 ou 3.

3.3.1 Valeurs propres et inerties

Les valeurs propres et le pourcentage d'inertie expliqué (contribution relative),nous donnent une idée sur la quantité d'information retenue par chaque axe, et s'ob-

tiennent par les commandes suivantes :

$> library(FactoMineR)$

$> ACP = PCA(X)$ # Utilisation de l'ACP.

$> vp = ACP\$eig$ # Valeurs propres λ.

$> round(vp, 3)$

Valeurs propres	3.486	1.422	0.473	0.294	0.219	0.106
pourcentages d'inertie(%)	58.099	23.698	7.880	4.906	3.644	1.774
pourcentages cumulés(%)	58.099	81.79	89.676	94.582	98.226	100.000

TAB. 3.2 – Valeurs propres et inerties

Histogramme : En utilisant les commandes suivantes pour trace histogramme des Valeurs propres (figures (3.1)) :

$> barplot(vp[,2], ylab = "\%d'inertie", names.arg = (round(vp[,2],3)), col = 3)$

Histogramme des vps

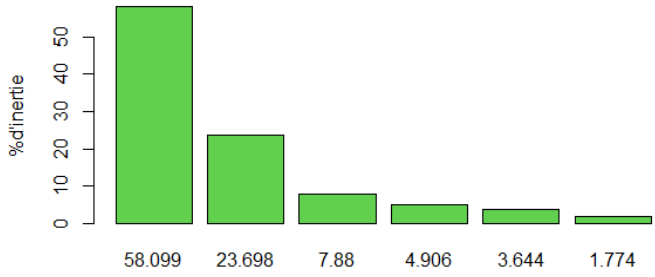

FIG. 3.1 – pourcentage des valeurs propres

Commentaire :

D'aprés la table (3.3) et l'histogramme (1.3) on remarque que le $1^{ér}$ axe représente environ 58.099% de valeur propre λ_1, le $2^{éme}$ axe représente environ 23.698% de

valeur propre λ_2...ect.

pour cela, on ne prend que les deux premiers axes principaux, parce qu'ils traduisent a eux seuls 81.797% de l' information disponible. Donc l'analyse est de bonne qualité sur les deux premiér dimensions (plan).

3.3.2 Représentation des individus et variables dans les nouveaux axes

Analyse du nuage des individus

Le tableau (3.3) fournit les coordonnées (composantes principales), les contributions et qualité de représentation (cosinus carré) des individus qui sont calculées à l'aide des commandes suivantes :

$> coordl = ACP\$ind\$coord$ # Coordonnées.

$> cos2l = ACP\$ind\$cos2$ # Cosinus Carré.

$> contribl = ACP\$ind\$contrib$ # Contribution.

$> cl1 = coordl[, 1]$ # $1^{ér}$ composant principale.

$> cl2 = coordl[, 2]$ # $2^{éme}$ composant principale.

$> round(cl1, 3)$

$> round(cl2, 3)$

$> round(contribl[, 1], 3)$

$> round(contribl[, 2], 3)$

$> round(cos2l[, 1], 3)$

$> round(cos2l[, 2], 3)$

On peut construire le graphe (3.2) des individus dans $\mathbb{R}^2$ pour utilise les commandes :

$> round(range(cl1), 3)$ # Borne du $1^{ére}$ axe.

$-3.170 \quad 3.714$

$> round(range(cl2), 3)$ # Borne du $2^{iéme}$ axe.

$-1.981 \quad 2.233$

Tracer le graphe des états selonx les 2 axes.

$> plot(cl1, cl2, ylab = "axe2 = 23.698\%", xlab = "axe1 = 58.099\%", xlim = c(-4, 4), ylim = c(-2.5, 2.5), main = "Projection des individus sur leplan(ACP)", col = 5)$

$> abline(h = 0, v = 0, col = 1)$

$> text(cl1, cl2, row.names(coordl), col = "red", cex = 1)$

$> abline(h = -4 : 4, v = -4 : 4, lty = 3, col = 4)$

	Coordonnées		Contribution		cos2	
	c_1	c_2	c_1	c_2	c_1	c_2
1	-0.474	2.173	0.323	16.604	0.040	0.834
2	1.374	0.610	2.710	1.306	0.310	0.061
3	2.461	-1.525	8.688	8.175	0.593	0.228
4	-1.388	1.127	2.764	4.465	0.544	0.358
5	3.714	0.174	19.781	0.107	0.931	0.002
6	1.969	-1.260	5.559	5.586	0.654	0.268
7	-1.510	-0.969	3.269	3.300	0.598	0.246
8	0.299	-1.419	0.128	7.082	0.039	0.881
9	2.872	0.033	11.831	0.004	0.867	0.000
10	0.118	1.320	0.020	6.130	0.007	0.865
11	0.005	-1.981	0.000	13.801	0.000	0.627
12	-1.970	-0.328	5.564	0.377	0.866	0.024
13	-0.159	0.555	0.036	1.082	0.010	0.127
14	-1.098	0.158	1.730	0.088	0.755	0.016
15	-3.170	-1.113	14.410	4.355	0.856	0.106
16	-1.068	-0.339	1.637	0.403	0.904	0.091
17	-2.234	1.274	7.160	5.705	0.716	0.233
18	0.737	2.233	0.780	17.533	0.089	0.815
19	-2.405	-1.013	8.293	3.606	0.732	0.130
20	1.925	0.289	5.317	0.293	0.501	0.011

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

Commentaire :

Contribution des individu D'après l'équation (2.2)

$$Cr\,(i,j) = \frac{p_i c_{ij}^2}{\lambda_j} = Contribution_j\,(e_i)\,.$$

On retient pour l'interprétation des individus dont la *Contribution* est supérieure à la *Contribution moyenne*$(1/n)$, donc $|c_{ij}| > \sqrt{\lambda_j}$.

`L'axe 1`

On a $\sqrt{\lambda_1} = \sqrt{3.486} = 1.867$, on prend seulement les coordonnées des individus de la $1^{ére}$ composante supérieurs ou égales à $\sqrt{\lambda_1}$, puis on regroupe d'après ces signes.

Le tableau suivant contient neuf etats devisés en 2 groupes qui sont bien représentés sur le premièr axe

−	+
12.Idaho	3.Arizona
15.Iowa	5.California
17.Kentucky	6.Colorado
19.Maine	9.Florida
	20.Maryland

`L'axe 2`

De la même manière on compare les coordonnées des individus par la $2^{ième}$ composant a $\sqrt{\lambda_2} = \sqrt{1.422} = 1.192$, puis on regroupe d'après ces signes.

Le tableau suivant contient huit état devisées sur 2 groupes qui sont bien représentées sur deuxième axe.

−	+
3.Arizona	1.Alabma
6.Colorado	10.Georgia
8.Delaware	17.Kentucky
11.Hawaii	18.Louisiana

- si ce cosinus carré est proche de 1, cela veut dire que l'individu est bien représenté

par sa projection sur le sous-espace.

cosinus carré (qualité) On remarque que "Kansas" est bien représenté sur le plan avec une qualité de représentation égale à :

$$cos2_{(1,2)}\,(Kansas) = 0.904 + 0.91 = 0.995.$$

par contre "Illinois" est très mal représenté :

$$cos2_{(1,2)}\,(Illinois) = 0.010 + 0.127 = 0.137$$

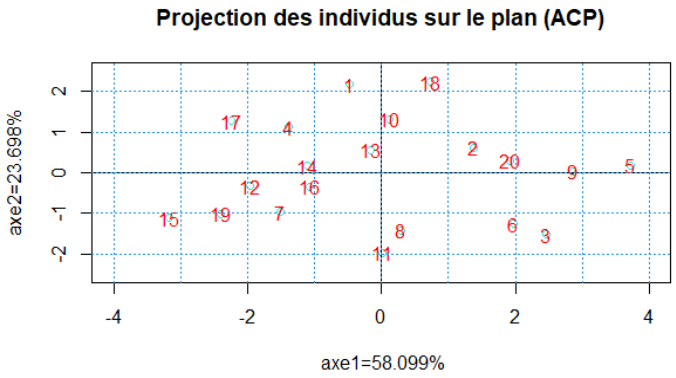

FIG. 3.2 – Représentation de nuage des individus

Analyse du nuage des variables

Les représentations du nuage des variables permettent de visualiser rapidement les corrélations entre les variables. On peut construire le graphe (3.3) des variables et le tableau (3.4) des composantes principales, les contribution et cosinus carré des variables par les commandes suivantes :

$> coord = ACP\$var\$coord$ # Coordonnées.

$> cos2 = ACP\$var\$cos2$ # Cosinus Carré.

$> contrib = ACP\$var\$contrib$ # Contribution.

$> c1 = coord[, 1]$ # $1^{ére}$ composant principale.

$> c2 = coord[, 2]$ # $2^{iéme}$ composant principale.

$> round(c1, 3)$

$> round(c2, 3)$

$> round(contrib[, 1], 3)$

$> round(contrib[, 2], 3)$

$> round(cos2[, 1], 3)$

$> round(cos2[, 2], 3)$

$> round(cor[, 1], 3)$

$> round(cor[, 2], 3)$

	Coordonnées		Contribution		cos2	
	c_1	c_2	c_1	c_2	c_1	c_2
Meurtre	0.501	0.774	7.202	42.104	0.251	0.599
Rapt	0.885	0.161	22.475	1.820	0.783	0.026
Vol	0.788	0.054	17.796	0.203	0.620	0.003
Attaque	0.832	0.343	19.865	8.292	0.692	0.118
Viol	0.802	-0.491	18.474	16.971	0.644	0.241
Larcin	0.703	-0.660	14.188	30.609	0.495	0.435

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

graphe :

$> round(range(c1), 3)$ # Borne du $1^{ére}$ axe.

0.501 0.885

$> round(range(c2), 3)$ # Borne du $2^{iéme}$ axe.

−0.660 0.774

Tracer le graphe de cercle des correlations.

> plot(c1, c2, ylab = "comp1 : 58.099%", xlab = "comp2 : 23.698%", main = "Projection des variables sur le plan", xlim = c(-1,1), ylim = c(-1,1), col=1)

> abline(h = 0, v = 0, lty = 3, col = 1)

> text(c1,c2,row.names(coord),col = "red")

> symbols(0, 0, circles = 1, ylab = "comp1 = 58.099%", xlab = "comp2 = 23.698%", inches = F, add =T)

> for (i in 1 :6) {

> arrows(0,0,c1[i],c2[i],angle = 50,length = 0.05,col="blue")

}

> abline(h = -0.5 :0.5, v= -0.5 :0.5, lty = 3, col = 4)

> abline(h = 0, v = 0, lty = 3, col = "red")

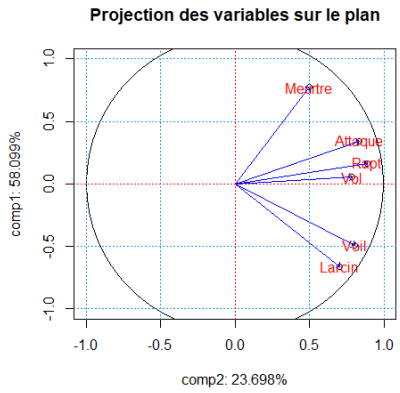

FIG. 3.3 – Représentation de nuage des variables

Commentaire : D'après la table (3.4) et la cercle des corrélations de la figure(3.3) , on observe que :

L'axe 1 Les variables sont corrélées positivement et assez fortement entre elles, et toutes les coordonnées sur le $1^{ér}$ axe proche de 1. Donc les variables sont bien représentons sur cet axe

L'axe 2 Les valeur de corrélation entre les variables et le $2^{éme}$ axe sont faibles surtout "Vol", "Rapt" et "Attaque".

Conclusion

En conclusion, dans ce mémoire nous avons étudiés l'Analyse en Composantes principales (ACP), qui est une méthode de base en statistique multidimensionnelle. L'objectif de cette méthode est de représenter graphiquement les relation entre individus par l'évaluation de leur ressemblance, ainsi que les relations entre variables par l'évaluation de leurs liaisons, en recharchant des espaces de plus faibles qui ajustent au mieux le nuage de points, c'est-à-dire qui respectent le plus possible la configuration initiale.

Nous avons également mené une étude et une interprétation des résultats obtenus par l' ACP sur des données réelles et ce ci en utilisant les différents packages du logiciel R.

ACP certainement aujourd'hui l'une des méthodes les plus employées, vu son importance et son utilisation dans beacuoup de domaines, telsque l'économie, la physique, ingénierie,...etc. Il existe également d'autres méthodes appelées méthodes d'analyse factorielle des correspondances (AFC) et l'analyse des correspondances multiples (ACM), qui se rapporte a ce domaine.

Bibliographie

[1] Antoniadis, A., & Yacrt, B. (5 janvier 2015). Démarrer en R. Université Joseph Fourier, Grenoble.

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

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

[4] Fenelon, J. P. (1981). Qu'est-ce que l'analyse des données. Lefonen, Paris, 3(11).

[5] Gonzalez, P. L. L'analyse en composantes principales (ACP). Tir de.

[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. UABB. Tlemcen.

[8] Necir, A.,(2021) Analyse en Composantes Principales, cours de master 1.UMK.Biskra.

[9] Reboul, L. Maître de conférences en statistique à Aix-Marseille Université. http ://iml.univ-mrs.fr˜reboul/enseignement.html.

[10] Saporta, G., & Niang, N. (2003). Analyse en composantes principales.

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

Annexe A : Logiciel R

Aujourd'hui beaucoup des logiciels peuvent être utilisés à des fins statistiques : Execel, SAS, SPSS et R figurant parmi les plus utilisés. Les possibilités de manipulation de données sous R sont en général largement supérieurs à celles des autres logiciels usuels d'analyse statistique

C'est un logiciel libre et modulaires de nombeux packagrs complémentaires offrent une grande variété de procédures mathématiques ou statistiques, incluant des méthodes de représentation graphiques complexes, le traitement des séries chronologiques, l'analyse des données,... etc. R est de plus utilisé dans tous les secteurs scientifiques, y compris dans le domaine des analyses d'enquêtes et plus généralement, des sciences sociales [1].

Il a été initialement crée, en 1996 par Robert Genteleman et ROss Ihaka. DEPUIS 1997, R est développé par une équipe "R Core Team".

Enfin, est disponible en téléchargement gratuit pour les principaux système d'exploitation à l'adresse : https : //WWW.r-project.org/ [1].

Annexe B : Abréviations et Notations

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

ACP	:	Analyse en composantes principales.
n	:	Nombre des individus.
p	:	Nombre des variables.
X	:	Tableau des données.
x_{ij}	:	Valeur de l'individu i par la variable j.
e_i	:	$i^{éme}$individu.
X_j	:	$j^{éme}$variable.
$\mathbb{R}^p$	:	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.
1_n	:	Vecteur unitaire d'identité de taille n.

n	:	Nombre des individus.
p	:	Nombre des variables.
X	:	Tableau des données.
g	:	Centre de gravité.
$\bar{x}_j$,	:	Moyenne de la variable x_j.
Y	:	Tableau des donnéess centres.
Z	:	Tableau de données centres reduites.
s_j	:	L'écart-type.
s_j^2	:	La variance de lavariable x_j.
$s_{jj'}$	:	La covariance des variables x_j et $x_{j'}$.
Var	:	Variance
Cov	:	Covariance
V	:	Matrice de variance-covariance de tableau X.
R	:	Matrice de corrélation.
$r_{jj'}$	:	Cofficient de corrélation entre x_j et $x_{jj'}$.
$d(e_i, e_{i'})$	:	Distance entre e_i et $e_{i'}$.
M	:	Métrique.
I_g	:	Inertie totale.
tr	:	Trace d'une matrice.
F_k	:	Sous-espace de projection de dimension k.

Δ	:	Sous-espace vectorieil.
f_i	:	Projection de l'individu e_i.
P	:	Operateur de projection.
X_{proj}	:	Tableau de donnée de nuage projeté.
g_{proj}	:	Centre de gravité projeté.
V_{proj}	:	Matrice de variance-covariance de nuage projeté.
α	:	Axe principal.
u	:	Facteur principal.
c	:	Composant principal.
λ	:	Valeur propre.
θ_{ij}	:	L'angle entre e_i et Δ_j.
Ca	:	Contribution absolue.
Cr	:	Contribution relative.
Cv	:	Contribution d'une variable.
$r(.,.)$	:	Cofficient de corrélation.

الملخص

يعد تحليل المكونات الرئيسية أحد أكثر الطرق استخداما لتلخيص الشكل الإحصائي والبياني لأقصى قدر من المعلومات الواردة في جدول واسع من البيانات الكمية، كما يسمح لنا برؤية الارتباطات بين المتغيرات وأوجه التشابه بين الأفراد. نقدم في هذه المذكرة لمحة عامة عن تحليل المكونات الرئيسية، والذي يمكن إجرائه باستخدام البرنامج الإحصائيR ، ونقدم أيضا مثال تطبيقي على بيانات حقيقية.

Résumé

L'Analyse en Composantes Principales (ACP) est l'une des méthodes les plus utilisées pour résumer statistiquement et graphiquement le maximum d'informations contenues dans un large tableau de données quantitatives, il permet également de voir les corrélations existantes entre les variables et les ressemblances entre les individus. Dans ce mémoire, nous donnons un aperçu de l'ACP, qui peut être effectuée à l'aide du logiciel statistique R, nous présentons également un exemple application avec des données réelles.

Abstract

Principal Components Analysis (PCA) is one of the most widely used method for statistically and graphically summarizing the maximum amount of information contained in a broad table of quantitative data, it also allows to see correlations between variables and similarities between individuals. In this memory, we give a general overview of (PCA), which can be performing using the statistical software R; we also present an example application with real data.