MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDERRAHMANE MIRA BEJAIA
FACULTÉ DES SCIENCES EXACTES
DEPARTEMENT DE MATHEMATIQUES


**Polycopié présenté**

**Par**

BELAIDE Karima épouse TIMERIDJINE

**THÈME**

# ANALYSE EN COMPOSANTES PRINCIPALES NORMEES

# Avant propos

Cet ouvrage s'adresse aux étudiants de Statistiques et Traitement Informatique des Données (STID), d'Informatiques ainsi qu'aux étudiants d'économie et gestion et à tous ceux désirant s'initier à l'analyse des données et en particulier l'analyse en composantes principales normées (ACP normée).

Il est conseillé aux lecteurs d'avoir de bonnes connaissances d'Algèbre linéaire ( Calcul matricielle, diagonalisaton, normes matricielles) et quelques notions d'Analyse mathématiques ( espace métrique, produit scalaire, recherche d'extrema).

Il est apparu à travers le monde réel, que l'enseignement de l'analyse des données sans traitements de données réelles avec interprétation demeure un enseignement sans saveur. Pour combler cette lacune, dans ce polycopier, après avoir détailler la méthode d'analyse en composantes principales normée qui est très utilisée dans divers domaines, nous présenterons quelques mini-projets réalisés par les étudiants de la quatrième année Informatiques et les étudiants de la troisième année Licence STID de l'université Abdelrahman Mira sur l'application de la méthode de l'analyse en composantes principales normée dans des problèmes réels.

# Table des matières

# Introduction

Le but des statistiques est de dégager les significations de données, numériques ou non, obtenues lors de l'étude d'un certain phénomène donné. Les statistiques peuvent être vue en fonction de l'objectif fixé. Des méthodes statistiques sont employées pour explorer des données ( approches desriptives et graphiques) ou pour prédire un certain comportement (approches décisionnelles, inferentielles et prédictives).

La statistique mathématiques ou inductive est un ensemble de méthodes permettant de faire des prévisions et des interpolations à partir de caractères observés sur chaque individu de la population à étudier. L'analyse des données est un ensemble de méthodes qui s'inscrivent dans le cadre de la statistique exploratoire multidimentionelle et peuvent également servir la statistique prédictive. Le développement des méthodes d'analyse des données a commencé durant les années 50 avec le développement de l'informatique et le stockage des données, qui depuis n'a cessé de croître. L'analyse des données a été mise en lumière en France, entre autres par ([1, 2]) qui a su par l'analyse des correspondances présenter les données de manières simple et interprétables.

Actuellement, l'analyse des données fais toujours l'objet de recherche pour s'adapter à tous type de données et faire face aux considérations de traitements en temps réel. Les méthodes développées sont souvent, intégré avec des méthodes issues de l'informatique et de l'intelligence artificielle dans le data mining (en français "fouille de données" ou encore extraction de l'information à partir de données) ([3]).

## 0.1 Domaines d'application

Aujourd'hui, les méthodes d'analyse des données sont employées dans de nombreux domaines tel que le marketing dans la gestion de la clientèle, le sondage dans l'analyse d'enquêtes, on peut également, citer la recherche documentaire très utile spécialement dans la recherche avec internet.

Le très grand nombre de données en méthodologie a été une des premières motivations pour développer les méthodes d'analyse des données. En effet, tout domaine scientifique qui doit gérer un grand nombre de données de type varié ont recours à ces approches (comme écologie, linguistique, économie etc.), ainsi que tout domaine industriel( tel que les assurances, banques, téléphonie etc.). Ces dernières sont également utilisées en traitement de signal et image et en ingénierie mécanique.

## 0.2 Les données

Nous considérons tout à bord une population de taille $n < \infty$ sur la quelle se fera l'étude d'un certain phénomène.

La population est décrite par un ensemble de $p$ caractères ( ou variables) qui peuvent être de type quantitatifs ou qualitatifs. Les données se présente sous forme d'un tableau ou de matrices à $n$ lignes et $p$ colonnes.

$$\underset{(n,p)}{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ & x_{ij} & \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Cette représentation de données peut faciliter la lecture des tableaux de petite dimension. Cependant, dés que la taille de la population $n$ est grande ou le nombre de caractères $p$ est élevé les techniques simples des statistiques descriptives ne suffisent plus. En règle générale, il est difficile lorsque le volume des données est important, de tirer des enseignements utiles sans traiter au préalable

## 0.3 Les objectifs

L'objectif principal que se fixe les chercheurs en analyse des données est de répondre aux problèmes posés par les tableaux de grandes dimension. On cherche à visualiser les données dans un meilleurs espace réduit et le regroupement dans tout l'espace.

## 0.4 Les méthodes

L'analyse des données regroupe deux familles de méthodes suivant l'objectif fixé.

1. Une première famille de méthodes cherche à représenter de grands ensembles de données par peu de caractères ie. chercher les dimensions pertinentes de ces données. Parmi ces méthodes nous citons l'analyse en composantes principales ACP et ACP normée qui sera présenté dans cet ouvrage, analyse des correspondances AFC, analyse des correspondances multiples AFC multiples ou encore l'analyse canonique ( [1]).

2. Une deuxième famille de méthodes cherche à classer les données de manière automatique tel que la classification par partition, classification hiérarchique ou analyse discriminante ([7]).

## 0.5 Les logiciels

Dans certains logiciels sont intégrés les méthodes d'analyse de données. On peut citer les logiciels SAS, splus, R, XIStat, UniWinplus, Stalab et SPAD.

Plan

Ce polycopier, est organisé comme suit

. Le chapitre 1 de ce polycopier est consacré aux préliminaires. Nous présentons des définitions et les outils de bases utilisés dans ce travail, ainsi qu'une étude descriptible des données.

. Dans le chapitre 2, nous présentons la méthode de l'analyse en composantes principales normée.

. Le chapitre 3 est consacré à des problèmes réels étudiés par les étudiants de la quatrième année informatique promotion (2008/2009) et les étudiant de troisième année licence Statistiques et Traitement Informatique des Données.

. Nous terminons par une conclusion.

CHAPITRE 1

# Préliminaires

Dans ce chapitre nous présentons dans une première quelques définitions et outils de base utilisé en analyse des données.

## 1.1 Introduction

L'analyse statistique multivariée consiste à analyser et comprendre des données de grandes dimension. L'analyse factorielle est une famille de méthodes géométriques dont les objectifs sont :

- résumer l'information contenues dans un tableau de données. En remplaçons ce dernier par un tableau de plus faible dimension.

- visualiser cette information. En éliminant les redondances d'information contenues dans le tableau de départ.

De l'analyse factorielle, plusieurs variantes se sont développées, tel que l'analyse en composantes principale (ACP) et ACP normée, l'analyse factorielle des correspondances AFC, l'analyse factorielle des correspondances multiple AFCM et l'analyse discriminante.

## 1.2 Notion des population et variables

**Définition 1.2.1** *On appelle population, un ensemble fini de personnes ou d'objets, noté par $I$ sur lequel se fera l'étude. Un élément de la population est dit individu et le cardinal de $I$ est dit taille de la population, notée $n$.*

**Définition 1.2.2** *Une variable ou caractère est une application qui à chaque individu associe un élément d'un ensemble appelé ensemble des modalités du caractère.*

On note $X : \quad I \to E$

$$I_i \longmapsto X(I_i) = X_i$$

Si $E \subset \mathbb{R}$ la variable $X$ est dite variable quantitative.

Si $E \subset \mathbb{R}^p$ la variable $X$ est dite vectorielle.

Si $E$ échappe à la mesure, $X$ est dite variable qualitative

## 1.3 Tableau des données

Soient $I$ une population de taille $n$, $X^1, \cdots, X^p$ $p$ variables quantitatives observées sur les individus de $I$. On construit un tableau de données quantitatives à partir de la population et les $p$ variables, on pose

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^p \\ \vdots & & \vdots \\ x_n^1 & \cdots & x_n^p \end{pmatrix}$$

où $x_i^j$ est la valeur du $i^{ème}$ individu par la variable $X^j$.

### 1.3.1 Notations

On note par $X^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix} \in \mathbb{R}^n$, la $j^{ème}$ variable mesurée sur les $n$ individus. Et par

${}^tX_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \end{pmatrix} \in \mathbb{R}^p$, le $i^{ème}$ individu observée sur les $p$ variables.

On suppose que chaque individu est muni d'un poids $p_i$.

$N(I) = \{({}^tX_i, p_i) / {}^tX_i \in \mathbb{R}^p \text{ pour } i = 1, \cdots, n\}$ le nuage des individus.

$N(J) = \{X^j / X^j \in \mathbb{R}^n \text{ pour } j = 1, \cdots, p\}$ le nuage des variables.

$z = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$.

$D_p = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_n \end{pmatrix}$ la matrice diagonale des poids

### 1.3.2 Eléments descriptifs du nuage des variables

#### Moyenne empirique

Pour $j = 1, \cdots, p$

$$\overline{X}^j = \sum_{i=1}^{n} p_i x_i^j$$

On peut facilement vérifier que $\overline{X}^j = \langle X^j, z \rangle_{D_p} = {}^tX^j D_p z$.

La moyenne empirique est donc la projection $D_p$-orthogonale du point $X^j$ de $\mathbb{R}^n$ sur la droite portée par le vecteur $z$.

**La variance**

$$Var\left(X^j\right) = \sum_{i=1}^{n} p_i \left(x_i^j - \overline{X}^j\right)^2$$

On a,

$$\begin{aligned} Var\left(X^j\right) &= \left\langle X^j - \overline{X}^j z, X^j - \overline{X}^j z \right\rangle_{D_p} \\ &= {}^t\left(X^j - \overline{X}^j z\right) D_p \left(X^j - \overline{X}^j z\right) \\ &= \left\| X^j - \overline{X}^j z \right\|_{D_p}^2 \\ &= \sigma_j^2 \end{aligned}$$

La variance est le carré de la norme relativement à la métrique $D_p$ du vecteur centré $\widetilde{X}^j = X^j - \overline{X}^j z$. L'ecart-type $\sigma_j$ est sa norme.

**Covariance**

$$cov\left(X^j, X^{j'}\right) = \sum_{i=1}^{n} p_i \left(x_i^j - \overline{X}^j\right)\left(x_i^{j'} - \overline{X}^{j'}\right)$$

On a,

$$\begin{aligned} cov\left(X^j, X^{j'}\right) &= \left\langle X^j - \overline{X}^j z, X^{j'} - \overline{X}^{j'} z \right\rangle_{D_p} \\ &= {}^t\left(X^j - \overline{X}^j z\right) D_p \left(X^{j'} - \overline{X}^{j'} z\right) \end{aligned}$$

La covariance est la projection du vecteur centré $\widetilde{X}^j = X^j - \overline{X}^j z$ sur la droite portée par le vecteur centré $\widetilde{X}^{j'} = X^{j'} - \overline{X}^{j'} z$.

**Coefficient de corrélation linéaire**

$$\begin{aligned}
\rho_{jj'} &= \frac{cov\left(X^j, X^{'}\right)}{\sqrt{Var\left(X^j\right) Var\left(X^{j'}\right)}} \\
&= \frac{\left\langle X^j - \overline{X}^j z, X^{j'} - \overline{X}^{j'} z \right\rangle_{D_p}}{\left\| X^j - \overline{X}^j z \right\|_{D_p} \left\| X^{j'} - \overline{X}^{j'} z \right\|_{D_p}} \\
&= \frac{\left\| X^j - \overline{X}^j z \right\|_{D_p} \left\| X^{j'} - \overline{X}^{j'} z \right\|_{D_p} \cos\left(\theta\right)}{\left\| X^j - \overline{X}^j z \right\|_{D_p} \left\| X^{j'} - \overline{X}^{j'} z \right\|_{D_p}} \\
&= \cos\left(\theta\right)
\end{aligned}$$

$\theta$ est l'angle entre les vecteurs centrés $\widetilde{X}^j = X^j - \overline{X}^j z$ et $\widetilde{X}^{j'} = X^{j'} - \overline{X}^{j'} z$ .

**Matrice des variances-covariances**

La matrice des variances-covariances est la matrice symétrique suivante

$$V = {}^t\widetilde{X} D_p \widetilde{X}$$

$\widetilde{X}$ est la matrice centrée donnée par $\widetilde{X} = \left(\widetilde{x}_i^j\right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} = \left(x_i^j - \overline{X}^j\right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$

**Matrice des corrélations**

La matrice des corrélations est la matrice symétrique suivante

$$R = {}^t\widetilde{\widetilde{X}} D_p \widetilde{\widetilde{X}}$$

$\widetilde{\widetilde{X}} = \left(\frac{x_i^j - \overline{X}^j}{\sigma_j}\right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ est la matrice centrée-réduite. On pose

$$D_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_p} \end{pmatrix}$$

Alors

$$R = D_{\frac{1}{\sigma}} \, {}^t\widetilde{X} D_p \widetilde{X} D_{\frac{1}{\sigma}}$$

### 1.3.3 Eléments descriptifs du nuage des individus

On muni $\mathbb{R}^p$ de la métrique identité $I_p$

**Centre de gravité**

$$g = \sum_{i=1}^{n} p_i \ {}^t X_i = \begin{pmatrix} \overline{X}^1 \\ \vdots \\ \overline{X}^j \end{pmatrix}$$

**Inertie au point** $a \in \mathbb{R}^p$

**Définition 1.3.1** *On appelle inertie au point $a \in \mathbb{R}^p$ la quantité*

$$\mathcal{I}_a = \sum_{i=1}^{n} p_i \left\| {}^t X_i - a \right\|_{I_p}^2$$

$\mathcal{I}_a$ mesure la proximité du nuage $N(I)$ au point $a$.

**Théorème 1.3.1** *Huyghens : $\forall a \in \mathbb{R}^p$*

$$\mathcal{I}_a = \mathcal{I}_g + \|g - a\|_{I_p}^2$$

**Remarque 1.3.1** *L'inertie est minimale au centre de gravité.*

Inertie au centre de gravité est donnée par,

$$\mathcal{I}_g = \sum_{i=1}^{n} p_i \left\| {}^t X_i - g \right\|_{I_p}^2$$

On pose ${}^t\widetilde{X}_i = {}^t X_i - g$, on aura,

$$\begin{aligned} \mathcal{I}_g &= \sum_{i=1}^{n} p_i \left\| {}^t\widetilde{X}_i \right\|_{I_p}^2 \\ &= \sum_{i=1}^{n} p_i \ \widetilde{X}_i \ {}^t\widetilde{X}_i \end{aligned}$$

Comme $\widetilde{X}_i\ {}^t\widetilde{X}_i$ est un scalaire, alors $\widetilde{X}_i\ {}^t\widetilde{X}_i = tr\left(\widetilde{X}_i\ {}^t\widetilde{X}_i\right) = tr\left({}^t\widetilde{X}_i\widetilde{X}_i\right)$, et

$$\begin{aligned}\mathcal{I}_g &= \sum_{i=1}^{n} p_i tr\left({}^t\widetilde{X}_i\widetilde{X}\right)\\ &= tr\left(\sum_{i=1}^{n} p_i\ {}^t\widetilde{X}_i\widetilde{X}\right)\\ &= tr\left(V\right)\end{aligned}$$

**Moment d'inertie par rapport à un sous espace vectoriel**

Soit $E$ un sous espace vectoriel de $\mathbb{R}^p$ et $E^\perp$ son supplémentaire ie. $\mathbb{R}^p = E \oplus E^\perp$ (somme directe).

Pour $i = 1, \cdots, n$

$${}^t\widetilde{X}_i = \alpha_i + \beta_i$$

avec $\alpha_i \in E$ et $\beta_i \in E^\perp$.

La quantité suivante est utilisée pour calculer la proximité du nuage au sous espace vectoriel $E$

$$\mathcal{I}_E = \sum_{i=1}^{n} p_i \left\|\beta_i\right\|_{I_p}^2$$

$\mathcal{I}_E$ est le moment d'inertie du nuage au sous espace vectoriel $E$. Et

$$\mathcal{I}_{E^\perp} = \sum_{i=1}^{n} p_i \left\|\alpha_i\right\|_{I_p}^2$$

est l'inertie du nuage au sous espace supplémentaire $E^\perp$.

**Proposition 1.3.1** *On a,*

$$\mathcal{I}_g = \mathcal{I}_E + \mathcal{I}_{E^\perp}$$

**Démonstration.** D'après le théorème de Pythagore

$$\left\|\widetilde{X}_i\right\|_{I_p}^2 = \left\|\alpha_i\right\|_{I_p}^2 + \left\|\beta_i\right\|_{I_p}^2$$

Donc,

$$\begin{aligned}\sum_{i=1}^{n} p_i \left\|X_i\right\|_{I_p}^2 &= \sum_{i=1}^{n} p_i \left\|\alpha_i\right\|_{I_p}^2 + \sum_{i=1}^{n} p_i \left\|\beta_i\right\|_{I_p}^2 \\ &= \mathcal{I}_E + \mathcal{I}_{E^\perp}\end{aligned}$$

■

Moment d'inertie par rapport à une droite et moment d'inertie par rapport à un hyperplan.

Soit $\Delta u$ la droite passant par l'origine et engendrée par le vecteur directeur normé $u$, $\Delta u^\perp$ l'hyperplan orthogonal à $\Delta u$ : $\mathbb{R}^p = \Delta u \oplus \Delta u^\perp$.

On peut écrire ${}^t\widetilde{X}_i = \alpha_i + \beta_i$,

$$\begin{aligned}\mathcal{I}_{\Delta u} &= \sum_{i=1}^{n} p_i \left\|{}^t\widetilde{X}_i - \alpha_i\right\|_{I_p}^2 \\ &= \sum_{i=1}^{n} p_i \left\|\beta_i\right\|_{I_p}^2\end{aligned}$$

et

$$\begin{aligned}\mathcal{I}_{\Delta u^\perp} &= \sum_{i=1}^{n} p_i \left\|{}^t\widetilde{X}_i - \beta_i\right\|_{I_p}^2 \\ &= \sum_{i=1}^{n} p_i \left\|\alpha_i\right\|_{I_p}^2\end{aligned}$$

Il vient,

$$\mathcal{I}_g = tr\,(V) = \mathcal{I}_{\Delta u} + \mathcal{I}_{\Delta u^\perp}$$

On a $\alpha_i \in \Delta u$, $\alpha_i = \omega_i u$. Où $\omega_i = \left\langle {}^t\widetilde{X}_i, u \right\rangle_{I_p} = \widetilde{X}_i u$ est la projection orthogonale de ${}^t\widetilde{X}_i$ sur $\Delta u$. Ainsi

$$\begin{aligned}
\mathcal{I}_{\Delta u^\perp} &= \sum_{i=1}^{n} p_i \left\|\alpha_i\right\|_{I_p}^2 \\
&= \sum_{i=1}^{n} p_i \left\|\omega_i u\right\|_{I_p}^2 \\
&= \sum_{i=1}^{n} p_i \omega_i^2 \left\|u\right\|_{I_p}^2 \\
&= \sum_{i=1}^{n} p_i \omega_i^2
\end{aligned}$$

On pose $\omega = \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_n \end{pmatrix} = \begin{pmatrix} \widetilde{X}_1 u \\ \vdots \\ \widetilde{X}_n u \end{pmatrix} = \widetilde{X} u$,

On obtient

$$\begin{aligned}
\mathcal{I}_{\Delta u^\perp} &= \left\|\omega\right\|_{D_p}^2 \\
&= {}^t\omega D_p \omega \\
&= {}^t u \, {}^t\widetilde{X} D_p \widetilde{X} u
\end{aligned}$$

Il suit

$$\mathcal{I}_g = tr\left(V\right) \Leftrightarrow tr\left(V\right) = \mathcal{I}_{\Delta u} + {}^t u \, {}^t\widetilde{X} D_p \widetilde{X} u = \mathcal{I}_{\Delta u} + {}^t u \, V u$$

D'ou

$$\mathcal{I}_{\Delta u} = tr\left(V\right) - {}^t u \, V u$$

CHAPITRE 2

# Analyse en composantes principales normée (ACP Normée)

## 2.1 Introduction

l'objectif de ce chapitre est d'étudier la méthode classiquement utilisée pour d'écrire et visualiser des données multivariées issues des variables continues : Analyse en composantes principales normée.

Les techniques d'analyse descriptive sont utilisées, notamment pour visualiser des données dans un sous espace représentatif, pour détecter des groupes d'individus et /ou variables, des valeurs aberrantes ou pour aider au choix de variables.

Ces méthodes permettent aussi de répondre aux questions de type : Quels individus se ressemblent du point de vue des différentes variables considérées. Ou bien, quelles sont les variables qui sont semblables du point de vue des individus de la population étudiée.

L'analyse en composantes principales est un outil de réduction de dimension tout en retirant la redondance d'informations apportées par plusieurs variables corrélées. Les

variables initiales seront alors remplacées par un nombre réduit de nouvelles variables construites comme combinaisons linéaires des variables initiales. Ces dernières sont alors indépendantes les unes des autres et peuvent être classées par ordre d'importance.

L'ACP normée, contrairement à l'ACP, permet d'éliminer la contrainte d'échelles de mesure des variables observées, dans le cas où ces dernières ne sont pas exprimées dans la même échelle de mesure.

Soit un tableau de données quantitatives $X$ de dimension $(n, p)$ ($n$ individus et $p$ variables). $\underset{(n,p)}{X} = (x_{ij})_{\substack{1\leq i\leq n\\ 1\leq j\leq p}}$

On suppose que chaque individu est muni d'un poids $p_i$ et on note par $\widetilde{X}$ le tableau centré-réduit. $\underset{(n,p)}{\widetilde{X}} = (\widetilde{x}_{ij})_{\substack{1\leq i\leq n\\ 1\leq j\leq p}}$

où $\widetilde{x}_{ij} = \frac{x_{ij}-\overline{X^j}}{\sigma_j}$.

Le tableau $\widetilde{X}$ est dit à deux entrées :

$$\widetilde{X} = \begin{cases} \left[\widetilde{X}^1, \cdots, \widetilde{X}^P\right] \\ \begin{bmatrix} \widetilde{X}_1 \\ \vdots \\ \widetilde{X}_n \end{bmatrix} \end{cases}$$

Où $\widetilde{X}^j \in \mathbb{R}^n$ représente la $j^{éme}$ colonne de $\widetilde{X}$. Et ${}^t\widetilde{X}_i \in \mathbb{R}^p$, $\widetilde{X}_i$ représente la i$^{éme}$ ligne de $\widetilde{X}$.

**Remarque 2.1.1** *la matrice $\widetilde{X}$ peut être assimilée à l'application linéaire définie comme suit : On note par $(e_1, \cdots e_n)$ la base canonique de $\mathbb{R}^n$ et par $\left(e'_1, \cdots e'_p\right)$ la base canonique de $\mathbb{R}^p$*

$$\widetilde{X} : \mathbb{R}^p \to \mathbb{R}^n$$

*Pour tout $j = 1, \cdots, p$,*

$$\widetilde{X}\left(e'_j\right) = \widetilde{X}^j = \sum_{i=1}^{n} \widetilde{x}_{ij} e_i$$

$$^t\widetilde{X} : \mathbb{R}^n \to \mathbb{R}^p$$

*Pour tout* $i = 1, \cdots, n$

$$^t\widetilde{X}(e_i) = \widetilde{X}_i = \sum_{j=1}^{p} \widetilde{x}_{ij} e'_j$$

**Remarque 2.1.2** *On a deux espaces vectoriels* $\mathbb{R}^n$ *et* $\mathbb{R}^p$*. L'approche consiste à se placer sur l'un des deux espaces vectoriels et projeter les points (de* $\mathbb{R}^n$ *ou* $\mathbb{R}^p$*) sur un sous espaces et établir des formules de transitions permettant de passer à l'analyse sur l'autre espace.*

## 2.2 Position du problème

### 2.2.1 On se place sur $\mathbb{R}^p$ muni de la métrique identité $(\mathbb{R}^p, \mathbb{I}_p)$

On considère le nuage des $n$ points de $\mathbb{R}^p$ noté $N(I) = \left\{ \left( {}^t\widetilde{X}_i, p_i \right) / {}^t\widetilde{X}_i \in \mathbb{R}^p, i = 1, \cdots, n \right\}$.

Le problème est de déterminer un sous espace vectoriel de $\mathbb{R}^p$ de dimension $1, 2, \cdots$ ,ou $r$ $(r < p)$ qui ajuste aux mieux les $n$ points du nuage $N(I)$ , au sens du critère des moindres carrés. La méthode consiste à chercher le sous espace vectoriel $E_1$ de $\mathbb{R}^p$ de dimension 1 qui ajuste au mieux le nuage. Si l'information apportée par ce sous espace vectoriel est considérable, on remplace les $p$ variables initiales par une seule nouvelle variable. Sinon, on cherche un sous espace $E_2$ de dimension 2. Et, ainsi de suite jusqu'à ce que la perte d'information soit négligeable.

**Recherche du sous espace vectoriel de $\mathbb{R}^p$ de dimension 1 qui ajuste aux mieux les $n$ points du nuage $N(I)$**

**Remarque 2.2.1** *Un sous espace vectoriel de* $\mathbb{R}^p$ *de dimension* 1 *est engendré par un seul vecteur, qu'on peut choisir normé.*

On note par $u_1$ le vecteur normé qui engendre $E_1$. Alors, $E_1 = \Delta u_1$ est une droite passant par l'origine de $\mathbb{R}^p$ et portée par le vecteur $u_1$. Schématiquement, cela se présente comme suit

$\omega_i^{(1)}$ est la projection orthogonale de ${}^t\widetilde{X}_i$ sur $\Delta u_1$. On a alors,

$$\omega_i^{(1)} = \left\langle {}^t\widetilde{X}_i, u_1 \right\rangle_{\mathbb{I}_p} = \widetilde{X}_i u_1 \ \in \mathbb{R}^p$$

Il est facile de voir, que le vecteur des $n$ projections

$$\omega^{(1)} = \begin{pmatrix} \omega_1^1 \\ \vdots \\ \omega_n^1 \end{pmatrix} = \widetilde{X} u_1$$

**Proposition 2.2.1** *Pour tout* $i = 1, \cdots, n$

$$\widetilde{X}_i = \underline{X}_i D_{\frac{1}{\sigma}}$$

*avec* $\widetilde{X}_i = (\widetilde{x}_i^1, \cdots, \widetilde{x}_i^p)$, $\widetilde{x}_i^j = \frac{x_i^j - \overline{X}^j}{\sigma_j}$ *est la donnée centrée réduite et* $\overline{X}_i = \left(x_i^1 - \overline{X}^1, \cdots, x_i^p - \overline{X}^p\right)$

.

**Démonstration.**

$$\begin{aligned} \underline{X}_i D_{\frac{1}{\sigma}} &= \left(x_i^1 - \overline{X}^1, \cdots, x_i^p - \overline{X}^p\right) \begin{pmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma p} \end{pmatrix} \\ &= \left(\frac{x_i^1 - \overline{X}^1}{\sigma_1}, \cdots, \frac{x_i^p - \overline{X}^p}{\sigma_p}\right) \\ &= \widetilde{X}_i \end{aligned}$$

■

**Proposition 2.2.2** *On note par $\underline{X}$ le tableau centré. On alors,*

$$\widetilde{X} = \underline{X} D_{\frac{1}{\sigma}}$$

**Démonstration.** On a

$$\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_n \end{pmatrix}$$

donc

$$\begin{aligned} \underline{X} D_{\frac{1}{\sigma}} &= \begin{pmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_n \end{pmatrix} D_{\frac{1}{\sigma}} = \begin{pmatrix} \underline{X}_1 D_{\frac{1}{\sigma}} \\ \vdots \\ \underline{X}_n D_{\frac{1}{\sigma}} \end{pmatrix} \\ &= \begin{pmatrix} \widetilde{X}_1 \\ \vdots \\ \widetilde{X}_n \end{pmatrix} = \widetilde{X} \end{aligned}$$

■

**Corollaire 2.2.1** *Le vecteur des n projections $\omega^{(1)} = \widetilde{X} u_1 = \underline{X} D_{\frac{1}{\sigma}} u_1$*

**Remarque 2.2.2** *Déterminer l'axe $\Delta u_1$ qui ajuste au mieux les points du nuage revient à déterminer le vecteur normé $u_1$.*

On utilise le critère des moindres carrés, qui consiste à minimiser les erreurs quadratiques moyennes sous la contrainte $\|u_1\|^2_{\mathbb{I}_p} = 1$.

**Problème d'optimisation s'écrit**

$$\min_{\|u_1\|^2_{\mathbb{I}_p}=1} \sum_{i=1}^{n} p_i \varepsilon_i^2$$

D'après le théorème de Pythagore on a,

$$\left\| {}^t\widetilde{X}_i \right\|_{\mathbb{I}_p}^2 = \omega_i^2 + \varepsilon_i^2$$

Le problème est alors équivalent à maximiser l'inertie

$$\begin{aligned}
\min_{\|u_i\|_{\mathbb{I}_p}^2=1} \sum_{i=1}^{n} p_i \left( \left\| {}^t\widetilde{X}_i \right\|_{\mathbb{I}_p}^2 - \omega_i^2 \right) \quad &\Leftrightarrow \quad \max_{\|u_i\|_{\mathbb{I}_p}^2=1} \sum_{i=1}^{n} p_i \omega_i^2 \\
&\Leftrightarrow \quad \max_{\|u_i\|_{\mathbb{I}_p}^2=1} \|\omega\|_{D_p}^2 \\
&\Leftrightarrow \quad \max_{\|u_i\|_{\mathbb{I}_p}^2=1} {}^t\omega D_p \omega \\
&\Leftrightarrow \quad \max_{\|u_i\|_{\mathbb{I}_p}^2=1} {}^t\left(\widetilde{X}u_1\right) D_p \left(\widetilde{X}u_1\right) \\
&\Leftrightarrow \quad \max_{\|u_i\|_{\mathbb{I}_p}^2=1} {}^tu_1 \, {}^t\widetilde{X} D_p \widetilde{X} u_1 \\
&\Leftrightarrow \quad \max_{\|u_i\|_{\mathbb{I}_p}^2=1} I_{\Delta u^\perp} \qquad (2.2.1)
\end{aligned}$$

**Remarque 2.2.3** *La matrice* $R = {}^t\widetilde{X} D_p \widetilde{X} = D_{\frac{1}{\sigma}}^t \underline{X} D_p \underline{X}$ *représente la matrice des corrélations.*

**Résolution du problème d'optimisation** On pose $H(u_1) = {}^tu_1 R u_1$ on cherche un extremum de $H$ sous la contrainte $\|u_1\|_{\mathbb{I}_p}^2 = {}^tu_1 u_1 = 1$. Le Lagrangien s'écrit,

$$\mathcal{L}(u_1) = {}^tu_1 R u_1 - \lambda \left( {}^tu_1 u_1 - 1 \right)$$

La condition nécessaire

$$\begin{aligned}
\frac{\partial \mathcal{L}(u_1)}{\partial u_1} &= 0 \Leftrightarrow 2{}^tu_1 R - 2\lambda {}^tu_1 u_1 = 0 \\
&\Leftrightarrow \quad {}^tu_1 R = \lambda {}^tu_1 u_1 \\
&\Leftrightarrow \quad R u_1 = \lambda u_1
\end{aligned}$$

$u_1$ est alors un vecteur propre de la matrice des corrélations $R$ associé à la valeur propre $\lambda$.

On alors $H(u_1) = {}^t u_1 R u_1 = \lambda\, {}^t u_1 u_1 = \lambda$ et

$$\max_{\|u_i\|^2_{\mathbb{I}_p}=1} H(u_1) = \max_{\|u_i\|^2_{\mathbb{I}_p}=1} \lambda$$

Le maximum est donc atteint au point $u_1 \in \mathbb{R}^p$, qui est vecteur propre de $R$ associé à la plus grande valeur propre $\lambda$.

Conclusion : Le sous espace vectoriel $E_1 = \Delta u_1$ de $\mathbb{R}^p$ de dimension 1 qui ajuste au mieux les points du nuage $N(I)$, est la droite engendrée par le vecteur propre normé $u_1$ de la matrice des corrélations $R$ associé à la plus grande valeur propre $\lambda_1$. $\left(\lambda_1 = \max_{1\leq j\leq p} \lambda j\right)$

**Proposition 2.2.3** *1. Les valeurs propres de la matrices $R$ sont non négatives.*

*2. Si $\lambda_m$ est la première valeur propre nulle $(\lambda_1 \geq \cdots \geq \lambda_{m-1} > 0)$, alors $rg(R) = m-1$*

**Démonstration.**

1. Soit $\lambda$ une valeur propre de $R$ associée à un vecteur propre $u$,

$$\begin{aligned} Ru &= \lambda u \Leftrightarrow {}^t uRu = \lambda {}^t u\, u \\ &\Leftrightarrow {}^t u {}^t\widetilde{X} D_p \widetilde{X} u = \lambda \\ &\Leftrightarrow \left\langle u\widetilde{X}, \widetilde{X}u \right\rangle_{D_p} = \left\| \widetilde{X}u \right\|^2_{D_p} = \lambda \geq 0 \end{aligned}$$

2. $R$ est une matrice carrée définie positive car

   . ${}^t R = R$

   . $\forall u \in \mathbb{R}^p$, $\langle u, u \rangle_R = {}^t uRu = \left\langle u\widetilde{X}, \widetilde{X}u \right\rangle_{D_p} = \left\| \widetilde{X}u \right\|^2_{D_p} \geq 0$

   $R$ est donc diagonalisable ie. il existe une matrice inversible. $P$ et une matrice diagonale des valeurs propres $D$ telle que

$$R = P^{-1} D P$$

$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{m-1} \end{pmatrix}$ et $P = (u_1, \cdots, u_{m-1})$, $u_i$ est un vecteur propre de $R$ associé à $\lambda_j$

Comme $P$ est inversible. alors $rg(P) = m-1$, et comme $\lambda_j \neq 0$ alors $rg(D) = m-1$

$$rg(R) = rg\left(P^{-1}DP\right) = rg\left(P^{-1}PD\right) = rg(D) = m-1$$

■

**Projection des individus sur $\Delta u_1$**

A chaque individu $I_i$ on associe un vecteur de $\mathbb{R}^p$ : $I_i \rightarrow {}^t\widetilde{X}_i = \begin{pmatrix} \widetilde{x}_i^p \\ \vdots \\ \widetilde{x}_i^p \end{pmatrix} \in \mathbb{R}^p$

A près réduction, on associe à $I_i$ un vecteur de $\mathbb{R}$ : $I_i \rightarrow \omega_i^{(1)} = \widetilde{X}_i u_1$ sa projection sur $\Delta u_1$.

Le vecteur des projections des $n$ individus est noté $\omega^{(1)} = \begin{pmatrix} \omega_n^{(1)} \\ \vdots \\ \omega_n^{(1)} \end{pmatrix} = \widetilde{X} u_1$

**Définition 2.2.1** *La nouvelle variable $\omega^{(1)}$ est dite composante principale.*

**Recherche du sous espace vectoriel de $\mathbb{R}^p$ de dimension 2 qui ajuste aux mieux les $n$ points du nuage $N(I)$**

**Proposition 2.2.4** *Le sous espace vectoriel $E_2$ de $\mathbb{R}^p$ de dimension 2 qui ajuste au mieux les points du nuage $N(I)$ contient $E_1$.*

**Démonstration.** Le sous espace vectoriel $E_2$ est engendré par deux axes passant par l'origine de $\mathbb{R}^p$. $E_2 = (\Delta x, \Delta y)$ où $\{x, y\}$ est une base de $E_2$. On suppose que $E_1 \nsubseteq E_2$. On alors, $E_1 \neq \Delta x$ et $E_1 \pitchfork \Delta x$

$$E_1 \neq \Delta y \text{ et } E_1 \pitchfork \Delta y$$

Considérons le sous espace $F_2$ de $\mathbb{R}^p$ de dimension 2 engendré par $E_1$ et $\Delta x$, comme $E_1 = \Delta u_1$ alors $\{u_1, x\}$ est une base de $F_2$ et la somme des projections des $n$ points sur $E_1$ est meilleur que celle sur $\Delta y$ et la projection des $n$ points sur $F_2$ est donc meilleur que sur $E_2$.

Contradiction avec le fait que $E_2$ est le meilleur sous espace qui ajuste les points du nuage $N(I)$. D'où $E_1 \subset E_2$. ■

**Remarque 2.2.4** *Le résultats de la proposition précédente se généralise facilement au cas* $E_{r-1} \subset E_r$.

*Avec* $E_k$ *le sous espace vectoriel de* $\mathbb{R}^p$ *de dimension* $k$ *qui ajuste au mieux les points du nuage* $N(I)$.

**Remarque 2.2.5** *La recherche de* $E_2$ *revient à la recherche d'un sous espace vectoriel de dimension* 1 *orthogonal à* $E_1$

On note par $\Delta u_2$ ce sous espace de dimension 1 orthogonale à $E_1 = \Delta u_1$, on a alors $\langle u_1, u_2 \rangle_{\mathbb{I}_p} = 0$. On suppose que $\|u_2\|^2_{\mathbb{I}_p} = 1$.

**Problème d'optimisation s'écrit**

$$\min_{\|u_2\|^2_{\mathbb{I}_p}=1} \sum_{i=1}^{n} p_i \varepsilon_i^2 \tag{2.2.2}$$

En procédant de la même manière que précédemment (2.2.1), on en déduit que $u_2$ est un vecteur propre normé de la matrice des corrélations $R$ associé à la deuxième plus grande valeur propre $\lambda_2$.

**Généralisation** Le sous espace vectoriel de dimension $r$ qui ajuste au mieux les $n$ points du nuage $N(I)$ est l'espace engendré par les $r$ vecteurs propres normés, $u_1, \cdots, u_p$ de la matrice des corrélations $R$ associés aux $r$ plus grandes valeurs propres $\lambda_1 \geq \cdots \geq \lambda_r$ respectivement.

**Définition 2.2.2** *On appelle $r$ $^{ième}$ axe factoriel, l'axe passant par l'origine et engendré par le vecteur propre normé $u_r$ de la matrice $R$ associé à la $r$ $^{ième}$ plus grande valeur propre $\lambda_r$.*

**Définition 2.2.3** *Les vecteurs projections des $n$ points du nuage $N(I)$ sur les axes factoriels sont dits facteurs.*

**Définition 2.2.4** *Le plan $(\Delta u_1, \Delta u_2)$ est dit premier plan factoriel, le plan $(\Delta u_1, \Delta u_3)$ est dit deuxième plan factoriel et $(\Delta u_2, \Delta u_3)$ est dit troisième plan factoriel.*

**Remarque 2.2.6** *L'oeil humain ne sachant pas regarder dans un espace de dimension supérieur à trois, on se restreint généralement en pratique, à deux ou trois axes factoriels. D'une part la représentation est plus lisible, d'autre part, on verra ultérieurement que l'information recueillie sur les deux ou trois premiers axes factoriels est plus importante, car cette information est directement liée aux plus grandes valeurs propres.*

**Moyennes et variances des composantes principales**

**Proposition 2.2.5** *les composantes principales sont :*

1. *Centrées*

2. *De variances égales aux valeurs propres.*

**Démonstration.** Soit $\omega^{(r)} = \begin{pmatrix} \omega_1^{(r)} \\ \vdots \\ \omega_n^{(r)} \end{pmatrix}$ la $r$ $^{ième}$ composante principale

1.

$$\begin{aligned}
\overline{\omega}^{(r)} &= \sum_{i=1}^{n} p_i \omega_i^{(r)} \\
&= \sum_{i=1}^{n} p_i \left( \widetilde{X}_i u_r \right) \\
&= \left( \sum_{i=1}^{n} p_i \widetilde{X}_i \right) u_r \\
&= \left( \sum_{i=1}^{n} p_i \widetilde{x}_i^1, \cdots, \sum_{i=1}^{n} p_i \widetilde{x}_i^p \right) u_r \\
&= (0, \cdots, 0)\, u_r = 0
\end{aligned}$$

2.

$$\begin{aligned}
Var\left(\omega^{(r)}\right) &= \left\langle \omega^{(r)}, \omega^{(r)} \right\rangle_{D_p} \\
&= {}^t\omega^{(r)} D_p \omega^{(r)} \\
&= {}^t\left( \widetilde{X} u_r \right) D_p \left( \widetilde{X} u_r \right) \\
&= {}^t u_r \, {}^t\widetilde{X} D_p \widetilde{X} u_r \\
&= {}^t u_r \lambda_r u_r \\
&= \lambda_r
\end{aligned}$$

■

### 2.2.2 On se place sur $\mathbb{R}^n$ muni de la métrique des poids $(\mathbb{R}^n, D_p)$

Considérons le tableau initial centré-réduit

$$\widetilde{X} = \left( \widetilde{X}^1, \cdots, \widetilde{X}^p \right)$$

**Remarque 2.2.7** *1.* $\overline{\widetilde{X}}^j = 0,\ j = 1, \cdots, p$

*2.* $Var\left(\widetilde{X}^j\right) = Var\left(\frac{X^j - \overline{X}\mathbf{1}}{\sigma_j}\right) = \frac{1}{\sigma_j^2} Var\left(X^j\right) = 1$

On note par $N(J) = \left\{ \widetilde{X}^j / \widetilde{X}^j \in \mathbb{R}^n \text{ pour } j = 1, \cdots p \right\}$ le nuage des $p$ points de $\mathbb{R}^n$ .

On procède de manière analogue au cas où on se place sur $\mathbb{R}^p$. Le problème d'optimisation lié à la recherche d'un sous espace vectoriel de $\mathbb{R}^n$ de dimension 1, noté $\Delta s_1$, qui ajuste au mieux les p points du nuage $N(J)$, consiste à minimiser la somme des erreurs quadratiques sous la contrainte que le vecteur directeur $s_1$ est normé. $\left( \|s_1\|^2_{D_p} = 1 \right)$.

**Problème d'optimisation s'écrit**

$$\min_{\|s_1\|^2_{\mathbb{D}_p}=1} \sum_{j=1}^{p} \varepsilon_j^2$$

D'après le théorème de Pythagore on a,

$$\left\| \widetilde{X}^j \right\|^2_{D_p} = \left( \psi_j^{(1)} \right)^2 + \varepsilon_j^2$$

$\psi_j^{(1)}$ est la projection $D_p$-orthogonale de la $j^{ième}$ variable centrée-réduite $\widetilde{X}^j$ sur $\Delta s_1$.

$$\psi_j^{(1)} = \left\langle \widetilde{X}^j, s_1 \right\rangle_{D_p} = {}^t\widetilde{X}^j D_p s_1$$

On note par $\psi^{(1)} = \begin{pmatrix} \psi_1^{(1)} \\ \vdots \\ \psi_p^{(1)} \end{pmatrix} = {}^t\widetilde{X} D_p s_1$ le vecteur des projections des $p$ variables.

Le problème est alors équivalent à

$$\begin{aligned}
\min_{\|s_1\|^2_{D_p}=1} \sum_{j=1}^{p} \left( \left\| \widetilde{X}^j \right\|^2_{D_p} - \psi_i^2 \right) \quad &\Leftrightarrow \quad \max_{\|s_1\|^2_{D_p}=1} \sum_{j=1}^{p} \left( \psi_i^{(1)} \right)^2 \\
&\Leftrightarrow \quad \max_{\|s_1\|^2_{D_p}=1} \left\| \psi^{(1)} \right\|^2_{\mathbb{I}_p} \\
&\Leftrightarrow \quad \max_{\|s_1\|^2_{D_p}=1} {}^t\psi^{(1)} \psi^{(1)} \\
&\Leftrightarrow \quad \max_{\|s_1\|^2_{D_p}=1} {}^t \left( {}^t\widetilde{X} D_p s_1 \right) \left( {}^t\widetilde{X} D_p s_1 \right) \\
&\Leftrightarrow \quad \max_{\|s_1\|^2_{D_p}=1} {}^t s_1 \; D_p \; \widetilde{X} \; {}^t\widetilde{X} D_p s_1
\end{aligned}$$

**Résolution du problème d'optimisation** On pose $H(u_1) = {}^t s_1 \; D_p \; \widetilde{X} \; {}^t\widetilde{X} D_p s_1$ on cherche un extremum de $H$ sous la contrainte $\|s_1\|^2_{D_p} = {}^t s_1 D_p s_1 = 1$. Le Lagrangien s'écrit,

$$\mathcal{L}(s_1) = {}^t s_1 \; D_p \widetilde{X} \; {}^t\widetilde{X} D_p s_1 - \lambda \left( {}^t s_1 D_p s_1 - 1 \right)$$

La condition nécessaire

$$\begin{aligned} \frac{\partial \mathcal{L}(s_1)}{\partial s_1} &= 0 \Leftrightarrow 2 {}^t s_1 \; D_p \widetilde{X} \; {}^t\widetilde{X} D_p - 2\lambda {}^t s_1 D_p = 0 \\ &\Leftrightarrow \quad {}^t s_1 \; D_p \widetilde{X} \; {}^t\widetilde{X} = \lambda {}^t s_1 \\ &\Leftrightarrow \quad \widetilde{X} \; {}^t\widetilde{X} D_p s_1 = \lambda s_1 \end{aligned}$$

$s_1$ est alors un vecteur propre de la matrice $\widetilde{X} \; {}^t\widetilde{X} D_p$ associé à la valeur propre $\lambda$.

On alors $H(s_1) = {}^t s_1 D_p \widetilde{X} \; {}^t\widetilde{X} D_p s_1 = \lambda \; {}^t s_1 D_p s_1 = \lambda$ et

$$\max_{\|s_1\|^2_{D_p}=1} H(s_1) = \max_{\|s_1\|^2_{D_p}=1} \lambda$$

Le maximum est donc atteint au point $s_1 \in \mathbb{R}^n$, qui est vecteur propre de $\widetilde{X} \; {}^t\widetilde{X} D_p$ associé à la plus grande valeur propre $\lambda$.

Conclusion : Le sous espace vectoriel $\Delta s_1$ de $\mathbb{R}^n$ de dimension 1 qui ajuste au mieux les points du nuage $N(J)$, est la droite engendrée par le vecteur propre normé $s_1$ de la matrice $\widetilde{X} \; {}^t\widetilde{X} D_p$ associé à la plus grande valeur propre $\lambda_1$. $\left( \lambda_1 = \max\limits_{1 \leq j \leq n} \lambda j \right)$

**Remarque 2.2.8** *1. Les propositions (1) et (2) et les remarque (1) et (2) précédentes sont vérifiées lorsqu'on remplace $R$ par $\widetilde{X} \; {}^t\widetilde{X} D_p$.*

*2. Les axes factoriels sont les droites de $\mathbb{R}^n$ passant par l'origine et engendrées par les $r$ vecteurs propres normées $s_1, \cdots, s_r$ de la matrice $\widetilde{X} \; {}^t\widetilde{X} D_p$ associés aux $r$ plus grandes valeurs propres respectives $\lambda_1 \geq \cdots \geq \lambda_r$.*

## 2.3 Relation entre les sous espaces propres de $\mathbb{R}^p$ et de $\mathbb{R}^n$

**Proposition 2.3.1** *Les matrices $R$ et $\widetilde{X}\ {}^t\widetilde{X}D_p$ ont les mêmes valeurs propres.*

**Démonstration.** $\Rightarrow$ / Soit $\lambda$ une valeur propre de $R = {}^t\widetilde{X}D_p\widetilde{X}$ associée au vecteur propre $u$. Montrons que $\lambda$ est encore valeur propre de $\widetilde{X}\ {}^t\widetilde{X}D_p$.

On a,

$$ {}^t\widetilde{X}D_p\widetilde{X}u = \lambda u \Longleftrightarrow \widetilde{X}{}^t\widetilde{X}D_p\left(\widetilde{X}u\right) = \lambda\left(\widetilde{X}u\right)$$

$\lambda$ est donc valeur propre de $\widetilde{X}\ {}^t\widetilde{X}D_p$ associée au vecteur propre $\widetilde{X}u$.

$\Leftarrow$ / Soit $\lambda$ une valeur propre de $\widetilde{X}\ {}^t\widetilde{X}D_p$ associée au vecteur propre $s$. Montrons que $\lambda$ est encore valeur propre de ${}^t\widetilde{X}D_p\widetilde{X}$.

On a,

$$\widetilde{X}\ {}^t\widetilde{X}D_ps = \lambda s \Longleftrightarrow\ {}^t\widetilde{X}D_p\widetilde{X}\left({}^t\widetilde{X}D_ps\right) = \lambda\left({}^t\widetilde{X}D_ps\right)$$

$\lambda$ est donc valeur propre de ${}^t\widetilde{X}D_p\widetilde{X}$ associée au vecteur propre ${}^t\widetilde{X}D_pu$. ■

**Proposition 2.3.2** *Formule de transition*

$\forall r \geq 1$, *on note par $u_r$ le vecteur propre de ${}^t\widetilde{X}D_p\widetilde{X}$ et $s_r$ le vecteur propre de $\widetilde{X}\ {}^t\widetilde{X}D_p$ associés à la même valeur propre $\lambda_r$. On a alors les deux formules de transition suivantes :*

1. $u_r = \frac{1}{\sqrt{\lambda_r}}\ {}^t\widetilde{X}D_ps_r$
2. $s_r = \frac{1}{\sqrt{\lambda_r}}\widetilde{X}u_r$

**Démonstration.**

1. La démonstration de 1 revient à vérifier les deux points suivants :

**1 a/** $\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r$ est un vecteur propre de $R$.

**1 b/** $\left\|\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r\right\|^2_{\mathbb{I}_p} = 1$.

**1 a/**

$$\begin{aligned} R\left(\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r\right) &= \frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p\left(\widetilde{X}{}^t\widetilde{X}D_p s_r\right) \\ &= \frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p\left(\lambda_r s_r\right) \\ &= \lambda_r\left(\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r\right) \end{aligned}$$

**1 b/**

$$\begin{aligned} \left\|\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r\right\|^2_{\mathbb{I}_p} &= {}^t\left(\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r\right)\left(\frac{1}{\sqrt{\lambda_r}}{}^t\widetilde{X}D_p s_r\right) \\ &= \frac{1}{\lambda_r}\left({}^t s D_p\widetilde{X}\right)\left({}^t\widetilde{X}D_p s_r\right) \\ &= \frac{1}{\lambda_r}{}^t s D_p\left(\widetilde{X}{}^t\widetilde{X}D_p s_r\right) \\ &= \frac{1}{\lambda_r}{}^t s D_p\left(\lambda_r s_r\right) \\ &= \|s_r\|^2_{D_p} = 1 \end{aligned}$$

La démonstration de la formule 2 s'obtient exactement de la même manière. ■

## 2.4 Représentation simultanée des individus et variables sur les axes factoriels

Comme il existe des formules de transition entre les deux espaces $\mathbb{R}^p$ et $\mathbb{R}^n$, l'ACP normée permet une interprétation simultanée des nuages $N(I)$ et $N(J)$ et de les représenter simultanément dans les plans factoriels. Notons que les nuages $N(I)$ et $N(J)$ ne sont en

réalités pas dans les mêmes espaces qui ont des dimensions différentes. Cette représentation simultanée est essentiellement graphique. L'inertie totale des nuages $N(I)$ et $N(J)$ est la même et est égale à

$$\lambda = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\widetilde{x}_i^j\right)^2$$

qui représente la variance ou dispersion totale.

### 2.4.1 Cas où on se place sur $\mathbb{R}^p$

Si $n \geq p$, on se place sur $(\mathbb{R}^p, \mathbb{I}_p)$, car la matrice à diagonaliser $R$ est de dimension $(p, p)$.

Soient $u_1, \cdots, u_r$ les r vecteurs propres normés de $R$ associés aux $r$ plus grandes valeurs propres $\lambda_1 \geq \cdots \geq \lambda_r$ respectivement. $\Delta u_1, \cdots, \Delta u_r$ les $r$ premiers axes factoriels.

#### Projection des individus

La projection du $i^{ème}$ individu sur le $k^{ème}$ axe factoriel $\Delta u_k$ est donnée par

$$\omega_i^{(k)} = \widetilde{X}_i u_k$$

Le vecteur des projections des $n$ individus est

$$\omega^{(k)} = \begin{pmatrix} \omega_1^{(k)} \\ \vdots \\ \omega_n^{(k)} \end{pmatrix} = \widetilde{X} u_k$$

#### Projection des variables

La projection de la $j^{ème}$ variable sur le $k^{ème}$ axe factoriel $\Delta s_k$ est donnée par

$$\psi_j^{(k)} = {}^t\widetilde{X}^j D_p s_k$$

Le vecteur des projections des $p$ variables est

$$\begin{aligned}
\psi^{(k)} &= \begin{pmatrix} \psi_1^{(k)} \\ \vdots \\ \psi_p^{(k)} \end{pmatrix} = {}^t\widetilde{X} D_p s_k \\
&= {}^t\widetilde{X} D_p \left( \frac{1}{\sqrt{\lambda_k}} \widetilde{X} u_k \right) \\
&= \frac{1}{\sqrt{\lambda_k}} (\lambda_k u_k) \\
&= \sqrt{\lambda_k} u_k
\end{aligned}$$

Ainsi la projection de la $j^{ème}$ variable sur le $k^{ème}$ axe factoriel $\Delta u_k$ est donnée par le $j^{ème}$ composante de $\sqrt{\lambda_k} u_k \in \mathbb{R}^p$.

### 2.4.2 Cas où on se place sur $\mathbb{R}^n$

Si $n \leq p$, on se place sur $(\mathbb{R}^n, D_p)$, car la matrice à diagonaliser $\widetilde{X}\ {}^t\widetilde{X} D_p$ est de dimension $(n, n)$.

Soient $s_1, \cdots, s_r$ les $r$ vecteurs propres normés de $\widetilde{X}\ {}^t\widetilde{X} D_p$ associés aux $r$ plus grandes valeurs propres $\lambda_1 \geq \cdots \geq \lambda_r$ respectivement. $\Delta s_1, \cdots, \Delta s_r$ les $r$ premiers axes factoriels.

#### Projection des individus

La projection du $i^{ème}$ individu sur le $k^{ème}$ axe factoriel $\Delta u_k$ est donnée par

$$\omega_i^{(k)} = \widetilde{X}_i u_k$$

Le vecteur des projections des $n$ individus est

$$\begin{aligned}\omega^{(k)} &= \begin{pmatrix}\omega_1^{(k)}\\ \vdots \\ \omega_n^{(k)}\end{pmatrix} = \widetilde{X}u_k \\ &= \widetilde{X}\left(\frac{1}{\sqrt{\lambda_k}}\,{}^t\widetilde{X}D_p s_k\right) \\ &= \frac{1}{\sqrt{\lambda_k}}(\lambda_k s_k) \\ &= \sqrt{\lambda_k}s_k\end{aligned}$$

Ainsi la projection du $i^{ème}$ individu sur le $k^{ème}$ axe factoriel $\Delta s_k$ est donnée par la $i^{ème}$ composante de $\sqrt{\lambda_k}s_k \in \mathbb{R}^n$.

**Projection des variables**

La projection de la $j^{ème}$ variable sur le $k^{ème}$ axe factoriel $\Delta s_k$ est donnée par

$$\psi_j^{(k)} = {}^t\widetilde{X}^j D_p s_k$$

Le vecteur des projections des $p$ variables est

$$\psi^{(k)} = \begin{pmatrix}\psi_1^{(k)}\\ \vdots \\ \psi_p^{(k)}\end{pmatrix} = {}^t\widetilde{X}D_p s_k$$

## 2.5 Aide à l'interprétation

A partir des relations données précédemment, on définit quelques règles d'interprétation

. Un individu sera du côté des variables pour les quelles il a de fortes valeurs, inversement, il sera du côté opposé des variables pour lesquelles il a de faibles valeurs.

. Plus les valeurs d'un individu sont fortes pour une variable, plus il sera éloigné de l'origine suivant l'axe factoriel qui décrit au mieux cette variable.

. Deux individus qui sont à une même extrémité d'un axe et sont éloignés de l'origine sont semblables (proches).

. Deux variables très corrélées positivement sont de même côté sur un axe.

Les axes factoriels donnent les images approchées des deux nuages $N(I)$ des individus et $N(J)$ des variables. Il est donc nécessaire de définir des indicateurs (ou contributions) pour mesurer la qualité de l'approximation ( ou représentation).

**Définition 2.5.1** *On appelle inertie expliquée par le $r^{ième}$ axe factoriel, la quantité*

$$I_r = \frac{\lambda_r}{\sum\limits_{k\geq 1}\lambda_k}$$

**Remarque 2.5.1** *On sait que $tr(R) = \sum\limits_{k\geq 1}\lambda_k$ où $R$ est la matrice des corrélations alors $tr(R) = p$ et*

$$I_r = \frac{\lambda_r}{p}$$

**Remarque 2.5.2** *En pratique l'inertie est exprimée en pourcentage $100I_r\%$, et représente la quantité d'information recueilli sur l'axe factoriel.*

**Définition 2.5.2** *L'inertie expliquée par le plan factoriel $(\Delta_r, \Delta_{r'})$ est donnée par*

$$I_{r,r'} = I_r + I_{r'} = \frac{\lambda_r + \lambda_{r'}}{p}$$

*L'inertie expliquée par le sous espace factoriel de dimension $r$ est donnée par*

$$I_{1,\cdots,r} = \frac{\lambda_1 + \cdots + \lambda_r}{p}$$

**Remarque 2.5.3** *En pratique, en générale, dés que le pourcentage d'inertie expliquée par le sous espace factoriel de dimension $r$ dépasse $75\%$, on dit qu'on a une bonne visualisation. On remplace les $p$ variables initiales par les $r$ composantes principales.*

### 2.5.1 Interprétation des nouvelles variables

**Corrélation entre les nouvelles et les variables initiales;**

Soit la $k^{ème}$ nouvelle variable (ou composante principale) $\omega^{(k)}$ et $\widetilde{X}^j$ la $j^{ème}$ variable initiale centrée-réduite.

**Proposition 2.5.1** *La corrélation entre* $\omega^{(k)}$ *et* $\widetilde{X}^j$ *est égale à* $\sqrt{\lambda_k}u_k^j = \psi_j^{(k)}$. *Où* $u_k^j$ *représente la* $j^{ème}$ *composante du vecteur propre* $u_k$ .

**Démonstration.**

$$Corr\left(\omega^{(k)}, \widetilde{X}^j\right) = \frac{Covar\left(\widetilde{X}^j, \omega^{(k)}\right)}{\sqrt{Var\left(\omega^{(k)}\right)}\sqrt{Var\left(\widetilde{X}^j\right)}}$$

$Var\left(\omega^{(k)}\right) = \lambda_k$ et $Var\left(\widetilde{X}^j\right) = 1$

$$\begin{aligned} Covar\left(\widetilde{X}^j, \omega^{(k)}\right) &= \left\langle \widetilde{X}^j, \omega^{(k)} \right\rangle_{D_p} \\ &= {}^t\widetilde{X}^j D_p \omega^{(k)} \\ &= {}^t\widetilde{X}^j D_p \widetilde{X} u_k \end{aligned}$$

or ${}^t\widetilde{X}^j D_p \widetilde{X} u_k$ est la $j^{ème}$ composante de la matrice $Ru_k$ et $Ru_k = \lambda_k u_k$.

Alors,

$$Corr\left(\omega^{(k)}, \widetilde{X}^j\right) = \sqrt{\lambda_k}u_k^j = \psi_j^{(k)}$$

■

**Cercle des corrélations**

Pour tout couple de composantes principales $\left(\omega^{(k)}, \omega^{(k')}\right)$, dans un cercle de centre $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$,

on représente les points $M_j = \begin{pmatrix} Corr\left(\omega^{(k)}, \widetilde{X}^j\right) \\ Corr\left(\omega^{(k')}, \widetilde{X}^j\right) \end{pmatrix}$.

Plus les points $M_j, j = 1 \cdots, p$ sont éloignés du centre, plus les variables $\widetilde{X}^j$ sont fortement corrélées avec les composantes principales. Donc bien représentées dans le plan $\left(\Delta\omega^{(k)}, \Delta\omega^{(k')}\right)$.

### 2.5.2 Les contributions

#### Qualité de représentation d'un élément sur l'axe factoriel

Un élément représente soit un individu ou une variable.

**Définition 2.5.3** *On appelle contribution relative du $i^{ème}$ individu sur le $k^{ème}$ axe factoriel, la quantité donnée par*

$$C_r^k(i) = \frac{\left|\omega_i^{(k)}\right|}{\left\|{}^t\widetilde{X}_i\right\|_{\mathbb{I}_p}^2} = \cos^2(\theta)$$

*où $\theta$ est l'angle formé par ${}^t\widetilde{X}_i$ et $u_k$.*

*On appelle contribution relative de la $j^{ème}$ variable sur le $k^{ème}$ axe factoriel, la quantité donnée par*

$$C_r^k(j) = \frac{\left|\psi_j^{(k)}\right|}{\left\|\widetilde{X}^j\right\|_{D_p}^2} = \cos^2(\theta)$$

*où $\theta$ est l'angle formé par $\widetilde{X}^j$ et $s_k$.*

La contribution relative d'un élément représente la qualité de visualisation de l'élément sur l'axe factoriel. Elle est donnée par le rapport de l'inertie de la projection sur l'axe et l'inertie totale. Ainsi, si la contribution relative est proche de 1, l'élément est proche de l'axe et donc, proche du plan de projection contenant l'axe. La projection de l'élément est donc proche de la valeur réelle. On dit que l'élément est très bien représenté sur l'axe.

**Définition 2.5.4** *On appelle contribution relative du $i^{ème}$ individu sur le sous espace factoriel de dimension r $(\Delta u_1, \cdots, \Delta u_r)$, la quantité donnée par*

$$C_r^{1,\cdots,r}(i) = \sum_{1\leq k\leq r} C_r^k(i) = \cos^2(\theta)$$

*où $\theta$ est l'angle formé par ${}^t\widetilde{X}_i$ et $(\Delta u_1, \cdots \Delta u_r)$.*

*On appelle contribution relative de la $j^{ème}$ variable sur le sous espace factoriel de dimension r $(\Delta s_1, \cdots, \Delta s_r)$, la quantité donnée par*

$$C_r^{1,\cdots,r}(j) = \sum_{1\leq k\leq r} C_r^k(j) = \cos^2(\theta)$$

*où $\theta$ est l'angle formé par $\widetilde{X}^j$ et $(\Delta s_1, \cdots \Delta s_r)$.*

### Qualité de représentation du nuage des individus

La qualité de représentation du nuage est donnée par le pourcentage d'inertie associée à l'axe, ie. le rapport d'inertie de la projection du nuage sur l'axe et l'inertie totale du nuage. Donnée par

$$QLTn = \sum_{1\leq j\leq n} C_r^k(i)$$

Cette quantité mesure l'importance d'un axe factoriel. Bien sûr, les premiers axes auront plus d'importance que les suivant. Nous jugons ces pourcentages en fonction de la taille du tableau. Par exemple, 10% est une valeur faible pour un tableau qui comporte 10 variables, mais c'est une valeur forte pour dans le cas de 100 variables.

### Contribution d'un élément à l'inertie d'un axe ou contribution absolue

**Définition 2.5.5** *On appelle contribution absolue du $i^{ème}$ individu à l'inertie du $k^{ème}$ axe factoriel, la quantité donnée par*

$$C_a^k(i) = p_i \frac{\left|\omega_i^{(k)}\right|}{\lambda_k}$$

*On appelle contribution absolue de la $j^{ème}$ variable à l'inertie du $k^{ème}$ axe factoriel, la quantité donnée par*

$$C_a^k(j) = \frac{\left|\psi_j^{(k)}\right|}{\lambda_k}$$

La contribution absolue est le rapport de l'inertie de la projection de l'élément sur l'axe et l'inertie de l'ensemble des points du nuage. Elle représente la part de l'élément pour construire l'axe factoriel et permet de mettre en évidence le sous ensemble des éléments ayant participés essentiellement à la construction de l'axe.

**Définition 2.5.6** *On appelle contribution absolue de deux individus $i_1$ et $i_2$ à l'inertie du $k^{ème}$ axe factoriel, la quantité donnée par*

$$C_a^{k,}(i_1, i_2) = C_a^k(i_1) + C_a^k(i_2)$$

*On appelle contribution absolue de deux variables $j_1$ et $j_2$ à l'inertie du $k^{ème}$ axe factoriel, la quantité donnée par*

$$C_a^k(j_1, j_2) = C_a^k(j_1) + C_a^k(j_2)$$

**Remarque 2.5.4** *1.*

*2.* $\forall i = 1, \cdots, n$ *et pour tout* $k \geq 1$, *et* $0 \leq C_a^k(i) \leq 1$.

*3.* $\forall j = 1, \cdots, p$ *et pour tout* $k \geq 1$, *et* $0 \leq C_a^k(j) \leq 1$.

*4.* $\sum_{1 \leq i \leq n} C_a^k(i) = 1$ *et* $\sum_{1 \leq j \leq n} C_a^k(j) = 1$

En effet,

$$\begin{aligned}
\sum_{1 \leq i \leq n} C_a^k(i) &= \sum_{1 \leq i \leq n} p_i \frac{\left|\omega_i^{(k)}\right|}{\lambda_k} \\
&= \frac{1}{\lambda_k} \left\|\omega^{(k)}\right\|_{D_p}^2 \\
&= \frac{1}{\lambda_k} Var\left(\omega^{(k)}\right) \\
&= 1
\end{aligned}$$

et

$$\begin{aligned}
\sum_{1\leq j\leq p} C_a^k(i) &= \sum_{1\leq j\leq p} \frac{\left|\psi_i^{(k)}\right|}{\lambda_k} \\
&= \frac{1}{\lambda_k}\left\|\psi^{(k)}\right\|_{\mathbb{I}_p}^2 \\
&= \frac{1}{\lambda_k}\ {}^t\left(\psi^{(k)}\right)\left(\psi^{(k)}\right) \\
&= \frac{1}{\lambda_k}\ {}^t\left({}^t\widetilde{X}D_p s_k\right)\left({}^t\widetilde{X}D_p s_k\right) \\
&= \frac{1}{\lambda_k}\ {}^t s_k D_p \widetilde{X}\,{}^t\widetilde{X}D_p s_k \\
&= \|s_k\|_{D_p}^2 \\
&= 1
\end{aligned}$$

**Définition 2.5.7** *On appelle contribution absolue du $i^{ème}$ individu à l'inertie du sous espace factoriel de dimension r $(\Delta u_1, \cdots, \Delta u_r)$, la quantité donnée par*

$$C_a^{1,\cdots,r}(i) = \sum_{1\leq k\leq r} C_r^k(i)$$

*On appelle contribution absolue de la $j^{ème}$ variable à l'inertie du sous espace factoriel de dimension r $(\Delta s_1, \cdots, \Delta s_r)$, la quantité donnée par*

$$C_a^{1,\cdots,r}(j) = \sum_{1\leq k\leq r} C_a^k(j)$$

Pour aider l'interprétation, nous devons s'appuyer sur les points suivants :

- Choisir le nombre d'axes. Notons que le choix du nombre d'axes reste un problème car il n'y a pas de solution rigoureuse. Les valeurs propres permettent de choisir le nombre d'axes de sorte que le pourcentage d'inertie cumulée soit supérieur ou égale à 75% environ, ou de sorte que les valeurs propres soient supérieurs où égales à 1, ou encore si on note un saut important dans l'histogramme des valeurs propres. Le nombre d'axes ne doit pas être trop grand.

- Etudier les indicateurs de la qualité des approximations.

- Interpréter les facteurs simultanément : En calculant

  **a/** les contributions des individus

  **b/** les coordonnés des individus et variables

CHAPITRE 3

# Problèmes

Ce chapitre a pour objet de présenter quelques problèmes concret rencontrés dans la vie courante. Ces problèmes sont des mémoire de licence et mini-projets réalisés par les étudiants de l'université de Bejaia.

## 3.1 Problème 1: Mémoire de licence STID (2014)

L'étude est réalisée sur les données récoltées au niveau de l'établissement hospitalier de Sidi Aich. Pour différentes communes limitrophes de Sidi Aich, nous nous sommes intéressé au nombre de patients de ces communes admis dans les différents services de l'hopital durant les années 2011 et 2012.

Le tableau de données est une matrice de dimension (,7) pour les deux années. Les lignes représentes les communes limitrophes de Sidi Aich et les collones sont les différents de l'hopital de Sidi Aich. On se place alors, sur l'espace des services.

### 3.1.1 Analyse des données de l'année (2011)

Tab. (1.1) : Matrice des corrélations

$$\begin{pmatrix} 1 & 0.9715 & 0.9073 & 0.8915 & 0.6469 & 0.9874 & 0.9390 \\ & 1 & 0.9111 & 0.8993 & 0.6384 & 0.9297 & 0.9595 \\ & & 1 & 0.9100 & 0.8827 & 0.8953 & 0.8801 \\ & & & 1 & 0.6951 & 0.8617 & 0.8911 \\ & & & & 1 & 0.6471 & 0.5821 \\ & & & & & 1 & 0.9331 \\ & & & & & & 1 \end{pmatrix}$$

Tab. (1.2) : Les valeurs propres, Inerties et Inerties cummulées

| valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 6.1053 | 87.2179 | 87.2179 |
| 0.5595 | 7.9922 | 95.2101 |
| 0.1413 | 2.0180 | 97.2281 |
| 0.0714 | 1.0568 | 98.2849 |
| 0.0545 | 0.7788 | 99.0637 |
| 0.0439 | 0.6269 | 99.6906 |
| 0.0217 | 0.3094 | 100 |

Tab. (1.3) : Projection des communes, contributions absolues et relatives

| Communes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Bejaia | −0.2159 | 0.02 | 0.06 |
| Sidi Aich | 8.5594 | 24.49 | 96.23 |
| El fleye | 2.7714 | 2.57 | 10.09 |
| Tnebdar | 1.8765 | 1.18 | 4.63 |
| Tibane | 0.3399 | 0.04 | 0.15 |
| Tifra | 2.6945 | 2.43 | 9.54 |
| Akfadou | 1.4125 | 0.67 | 2.62 |
| Chemini | 6.4233 | 13.79 | 5.42 |
| S. oufela | 2.0054 | 1.34 | 5.28 |
| Sidi Ayad | 2.0892 | 1.46 | 5.73 |
| Timezrit | 8.5812 | 24.61 | 92.72 |
| Fenaia | 1.0449 | 0.36 | 1.43 |
| Adkar | 0.5877 | 0.12 | 0.45 |
| T. Ighil | −0.744 | 0.01 | 0.04 |
| B. K'sila | −1.3188 | 0.58 | 2.28 |
| El Kseur | 0.3451 | 0.04 | 0.16 |
| Amizour | −0.7415 | 0.18 | 0.72 |
| Smaoun | −1.0083 | 0.34 | 1.34 |
| B. Djellil | −1.0629 | 0.38 | 1.48 |

Suite Tab. (1.3) : Projection des communes, contributions absolues et relatives

| Communes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Feraoun | −1.6827 | 0.95 | 3.72 |
| Barbacha | −1.5254 | 0.78 | 1.06 |
| Kendira | −1.6827 | 0.75 | 3.72 |
| Seddouk | 2.6912 | 2.42 | 9.51 |
| M'Cisna | 0.7187 | 0.17 | 0.68 |
| Amalou | 0.7196 | 0.65 | 2.56 |
| Ouzelaguen | −1.3959 | 0.04 | 0.14 |
| Akbou | −0.3258 | 0.51 | 2.00 |
| Chelata | 1.2328 | 0.81 | 3.19 |
| Tamokra | −1.5580 | 0.85 | 3.32 |
| Tazmalt | −1.5906 | 0.06 | 0.22 |
| B. Melokeche | −0.4114 | 0.67 | 2.63 |
| Ighil Ali | −1.4137 | 0.82 | 3.23 |
| Ighram | −1.4727 | 0.88 | 3.46 |
| B. Maouche | −1.6231 | 0.01 | 0.04 |
| Bouhamza | −0.1795 | 0.69 | 2.70 |
| Ait Rezine | −1.4341 | 0.60 | 2.37 |
| Toudja | −1.3423 | 0.95 | 3.72 |
| Oued ghir | −1.6827 | 0.70 | 2.95 |
| Tichy | −1.4467 | 0.93 | 3.67 |

Suite Tab. (1.3) : Projection des communes, contributions absolues et relatives

| Communes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Tala Hamza | −1.6712 | 0.90 | 3.55 |
| Boukhelifa | −1.6438 | 0.95 | 3.73 |
| Melbou | −1.6841 | 0.95 | 3.73 |
| Oukas | −1.6700 | 0.93 | 3.66 |
| Derguina | −1.6684 | 0.93 | 3.66 |
| Taskeriout | −1.6769 | 0.94 | 3.69 |
| Tamridjet | −1.6755 | 0.94 | 3.69 |
| S. El tnini | −1.6394 | 0.90 | 3.53 |
| Kherrata | −1.5151 | 0.77 | 3.02 |
| Boudjellil | −1.5680 | 0.73 | 2.85 |

Tab. (1.4) : Projection des variables, contributions absolues et relatives

| Services | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Medecine | 96.85 | 15.36 | 93.79 |
| Pédiatrie | 97.12 | 15.45 | 94.33 |
| Chirurgie | 96.84 | 15.36 | 93.78 |
| Orthopédie | 94.36 | 14.59 | 89.05 |
| Ophtalmologie | 75.92 | 09.44 | 57.64 |
| Maternité | 95.42 | 14.91 | 91.06 |
| Gynécologie | 95.33 | 14.89 | 90.88 |

### 3.1.2 Analyse des données de l'année (2012)

Tab. (1.5) : Matrice des corrélations (2012)

$$\begin{pmatrix} 1 & 0.9534 & 0.9502 & 0.8984 & 0.5844 & 0.9519 & 0.8879 \\ & 1 & 0.9541 & 0.9012 & 0.6029 & 0.9658 & 0.9022 \\ & & 1 & 0.9404 & 0.6938 & 0.9719 & 0.9195 \\ & & & 1 & 0.7121 & 0.9043 & 0.8327 \\ & & & & 1 & 0.6359 & 0.6271 \\ & & & & & 1 & 0.9354 \\ & & & & & & 1 \end{pmatrix}$$

Tab. (1.6) : Les valeurs propres

| valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 6.1100 | 87.2852 | 87.2852 |
| 0.5476 | 7.8224 | 95.1077 |
| 0.1654 | 2.3628 | 97.4704 |
| 0.0762 | 1.0880 | 98.5585 |
| 0.0473 | 0.6763 | 99.2348 |
| 0.0336 | 0.4796 | 99.7144 |
| 0.0200 | 0.2856 | 100 |

Tab. (1.7) : Projection des communes, contributions absolues et relatives

| Communes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Ighil Ali | −1.4436 | 0.73 | 2.86 |
| Sidi Aich | 8.4298 | 24.75 | 97.50 |
| El fleye | 2.3094 | 1.86 | 7.32 |
| Tinebdar | 2.2765 | 1.80 | 7.11 |
| Tibane | 0.8011 | 0.22 | 0.88 |
| Tifra | 2.6657 | 2.47 | 9.75 |
| Ighram | −1.6223 | 0.92 | 3.61 |
| Imoula | −1.7350 | 1.05 | 4.13 |
| S. oufela | 2.3429 | 1.91 | 7.53 |
| Sidi Ayad | 1.7720 | 1.09 | 4.31 |
| Timzrit | 8.4897 | 25.10 | 98.89 |
| Fenaia | 1.4150 | 0.70 | 2.75 |
| Kherrata | −1.5178 | 0.80 | 3.16 |
| T. Ighil | −0.2944 | 0.03 | 0.12 |
| Melbou | −1.7183 | 1.03 | 4.05 |
| El Kseur | 0.4961 | 0.09 | 0.34 |
| Oued ghir | −1.6910 | 1.00 | 3.92 |
| Smaoun | −0.8821 | 0.27 | 1.07 |
| Ouzelaguen | −0.0095 | 0.00 | 0.00 |
| Feraoun | −1.7306 | 1.04 | 4.11 |

Suite du tab (1.7) : Projection des communes, contributions absolues et relatives

| Communes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| S. el thinine | −1.7163 | 1.03 | 4.04 |
| Tazmalt | −0.5995 | 0.13 | 0.49 |
| Seddouk | 2.4090 | 2.02 | 7.96 |
| M'Cisna | 0.5255 | 0.10 | 0.38 |
| Tala hamza | −1.7450 | 1.06 | 4.18 |
| Toudja | −1.5835 | 0.87 | 3.44 |
| Tichy | −1.7135 | 1.02 | 4.03 |
| Tamokra | −1.6381 | 0.93 | 3.68 |
| Akfadou | 1.5447 | 0.83 | 3.27 |
| Adekar | 0.3566 | 0.04 | 017 |
| Akbou | 0.3540 | 0.04 | 017 |
| Amizour | −0.9591 | 0.32 | 1.26 |
| Amalou | −1.4607 | 0.74 | 2.93 |
| Aghbalou | −1.6866 | 0.99 | 3.90 |
| Aokas | −1.6515 | 0.95 | 3.74 |
| Ait rezine | −1.4653 | 0.75 | 2.95 |
| Bejaia | −0.9770 | 0.33 | 1.31 |
| Beni Maouch | −0.2264 | 0.02 | 0.07 |
| Beni Djellil | −1.2947 | 0.58 | 2.30 |
| Barbacha | −1.5584 | 0.85 | 3.33 |

Suite du tab (1.7) : Projection des communes, contributions absolues et relatives

| Communes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Bouhamza | −1.4578 | 0.74 | 2.92 |
| Beni Melikeche | −1.7214 | 1.03 | 4.07 |
| Chemini | 5.9690 | 12.41 | 48.89 |
| Chelata | −1.5902 | 0.88 | 3.47 |
| Derguina | −1.7500 | 1.07 | 4.20 |
| Beni ksila | −2.3462 | 0.63 | 2.49 |
| Boudjellil | −1.3713 | 0.65 | 2.58 |

Tab. (1.8) : Projection des variables, contributions absolues et relatives

| Services | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Medecine | 96.03 | 15.09 | 92.23 |
| Pédiatrie | 96.81 | 15.34 | 93.72 |
| Chirurgie | 98.83 | 15.98 | 97.66 |
| Orthopédie | 94.94 | 14.75 | 90.13 |
| Ophtalmologie | 72.78 | 08.67 | 52.96 |
| Maternité | 98.03 | 15.73 | 96.10 |
| Gynécologie | 93.91 | 14.43 | 88.19 |

**Conclusion 3.1.1** *Les patients de Sidi Aich et Timezrit sont généralement hospitalisés au sein de l'hopital de Sidi Aich contrairement aux patients des autres communes considérées. Touts les services de cet hopital, à part l'ophtalmologie, sont d'egals importances ( en terme du nombre d'admission).*

## 3.2 Problème 2: Mémoire de licence STID (2014)

L'étude est réalisée sur les données récoltées au niveau de l'entreprise Tchin-lait : Condia, et conserne les caractéristiques des différentes gammes de laits selon leurs composition.

Le tableau de données est une matrice de dimension (5,15), dont les lignes représentes 5 gammes de lait : et les colonnes sont les 15 élèments de la composition de chaque gamme. On se place alors sur l'espace des élèments de la composition.

Tab.(2.1) : Matrice à diagonaliser

$$\begin{pmatrix} 0.4150 & 0.1586 & 1.0821 & 0.2291 & 1.0548 \\ 0.1586 & 0.2357 & 0.3234 & 1.1542 & 0.5547 \\ 1.0821 & 0.3234 & 0.5586 & 0.1081 & 0.9080 \\ 0.2291 & 1.1542 & 0.1081 & 0.1951 & 0.2284 \\ 1.0548 & 0.5547 & 0.9080 & 0.2284 & 1.6363 \end{pmatrix}$$

Tab. (2.2) : Valeurs propres, Inerties et inerties cummulées

| valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 4.7767 | 95.5331 | 95.5331 |
| 0.1794 | 2.5881 | 99.1212 |
| 0.029 | 0.8588 | 99.98 |
| 0.001 | 0.0198 | 99.9998 |
| 0.0000 | 0.0002 | 100 |

Tab. (2.3) : Projections des gammes, contributions absolues et relatives

| gammes | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| viva | 0.9845 | 20.29 | 96.93 |
| Demi-écrimé | 0.9922 | 20.61 | 98.45 |
| Silouhaite | 0.9800 | 20.11 | 96.05 |
| Entier | 0.9943 | 20.70 | 98.86 |
| Candy-choco | 0.9346 | 18.29 | 87.35 |

Tab. (2.4) : Projections des élèments de la composition, contributions absolues et relatives

| Composition | Proj | $C_a$ (%) | $C_r$ (%) |
|---|---|---|---|
| Valeur inergitique | −1.1498 | 27.68 | 85.97 |
| Protéine | 0.1678 | 00.59 | 94.37 |
| Glucide | −0.0797 | 00.13 | 89.47 |
| Lipide | 0.1181 | 00.29 | 96.69 |
| Calcium | 0.6501 | 47.50 | 99.12 |
| Vitamine D | 0.1553 | 00.55 | 96.91 |
| Vitamine E | 0.1073 | 01.24 | 98.52 |
| Vitamine B1 | 0.1567 | 01.51 | 96.99 |
| Vitamine B2 | 0.1555 | 00.51 | 97.03 |
| Vitamine B3 | −0.0892 | 00.17 | 98.94 |
| Vitamine B5 | 0.1346 | 00.38 | 97.74 |
| Vitamine B6 | 0.1555 | 00.51 | 97.03 |
| Vitamine B8 | −0.0647 | 00.09 | 96.55 |
| Vitamine B9 | −0.7464 | 00.11 | 00.15 |
| Vitamine B12 | 0.1503 | 00.47 | 97.21 |

**Conclusion 3.2.1** *Touts les éléments de la composition des gammes considérées à part la vitamine B9, sont essentiels dans la compositions de lait. Les gammes sont de même importances.*

## 3.3 Problème 3: $4^{ème}$ année ingénieur en informatique (2008/2009).

Etude des critères de performances des micros portables décris par six variables suivantes :

CPU : puissance de traitement de données par le processeur (GHz)

DD : capacité du disque dure (Géga)

RAM : capacité de la mémoire vivre (Géga)

CG : capacité de la carte graphique (Méga)

DVB : la durée de vie de la batterie (minutes)

ECR : la dimension de l'écran (pousse)

Tab. (3.1) : Matrice de données

| | CPU | DD | RAM | CG | DVB | ECR |
|---|---|---|---|---|---|---|
| HP | 1.86 | 160 | 2 | 256 | 180 | 15.4 |
| Toshiba | 1.6 | 120 | 2 | 256 | 180 | 15.4 |
| Acer | 1.5 | 80 | 1 | 64 | 120 | 15.4 |
| Samsung | 1.8 | 250 | 2 | 512 | 180 | 17 |
| Sony | 2 | 250 | 3 | 1024 | 240 | 17 |
| Ibm | 1.73 | 140 | 1 | 128 | 180 | 15 |
| Siemens | 1.63 | 120 | 1 | 128 | 120 | 15 |
| Zala | 1.4 | 80 | 1 | 64 | 90 | 14 |

Tab. (3.2) : Matrice des corrélations

$$\begin{pmatrix} 1 & 0.8775 & 0.79102 & 0.81598 & 0.95583 & 0.77402 \\ & 1 & 0.79764 & 0.88341 & 0.85792 & 0.90232 \\ & & 1 & 0.87067 & 0.81969 & 0.78633 \\ & & & 1 & 0.83684 & 0.83553 \\ & & & & 1 & 0.79321 \\ & & & & & 1 \end{pmatrix}$$

Tab. (3.3) : Les valeurs propres, inerties et inerties cumulées

| valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 5.201 | 86.68 | 86.68 |
| 0.31044 | 05.17 | 91.85 |
| 0.25509 | 04.25 | 96.10 |
| 0.1262 | 02.10 | 98.21 |
| 0.073959 | 01.23 | 99.44 |
| 0.033343 | 00.56 | 100 |

Tab. (3.6) : Projections des Micro-portables, contributions absolues et relatives

| Micro-port | Proj | $C_a\,(1,2)$ (%) | $C_r\,(1,2)$ (%) |
|---|---|---|---|
| HP | $(0.7238, 0.7129)$ | 21.72 | 81.08 |
| Toshiba | $(-0.4862, -0.2045)$ | 02.25 | 29.77 |
| Acer | $(-1.8172, -0.4092)$ | 14.68 | 90.96 |
| Samsung | $(2.1012, -0.6361)$ | 26.90 | 86.96 |
| Sony | $(4.2101, -0.1100)$ | 43.08 | 97.75 |
| Ibm | $(-0.5204, 0.9556)$ | 37.42 | 79.37 |
| Siemens | $(-1.3757, 0.0430)$ | 04.62 | 89.37 |
| Zala | $(-2.8356, -0.3516)$ | 24.30 | 94.07 |

Tab. (3.7) : Projections des caractéristiques, contributions absolues et relatives

| caractéristiques | Proj | $C_a\ (1,2)(\%)$ | $C_r\,(1,2)$ (%) |
|---|---|---|---|
| CPU | $(0.9342, 0.3272)$ | 51.27 | 97.99 |
| DD | $(0.9527, -0.0720)$ | 19.12 | 91.21 |
| RAM | $(0.9058, -0.1198)$ | 20.40 | 83.49 |
| CG | $(0.9385, -0.1568)$ | 24.86 | 90.55 |
| DVB | $(0.9428, 0.2873)$ | 43.69 | 97.15 |
| ECR | $(0.9110, -0.2769)$ | 40.65 | 90.66 |

**Conclusion 3.3.1** *Toutes caractéristiques considérées sont fortement corrélées avec le premier axe, ce qui explique leurs importances pour la caractérisation des micro-portables. Les micro-portables Sumsug et Sony sont proches dans le sens où ils ont les mêmes caractéristiques.*

## 3.4 Problème 4: $4^{ème}$ année ingénieur en informatique (2008/2009).

Etude des différents critères de performance des voitures.

12 voitures de dernières génération sont décrites par 8 variables quantitatives, les données sont résumés dans le tableau suivant

Tab (4.1) : Tableau des données

| | Cyl | Pui | Long | Larg | Poids | Vmax | Prix | EmiCO2 |
|---|---|---|---|---|---|---|---|---|
| BMW x6 | 2993 | 286 | 445 | 174 | 2185 | 260 | 7900 | 220 |
| Toyota Aris | 1775 | 158 | 454 | 176 | 1345 | 220 | 2010 | 195 |
| Volvo c30 | 2360 | 170 | 425 | 176 | 1937 | 205 | 2240 | 156 |
| Chevrolet captiva | 2370 | 170 | 423 | 178 | 2640 | 210 | 2300 | 264 |
| Opel Astra | 1690 | 110 | 418 | 171 | 1930 | 215 | 1990 | 149 |
| Nissan Navara | 2598 | 110 | 410 | 169 | 1107 | 180 | 2190 | 149 |
| Hundai I30 | 2100 | 190 | 420 | 175 | 2300 | 210 | 1970 | 145 |
| Passat TDI | 1875 | 105 | 430 | 178 | 2200 | 240 | 2650 | 170 |
| Ford monda | 1990 | 155 | 415 | 174 | 1900 | 235 | 2479 | 185 |
| Renault Laguna | 1990 | 140 | 434 | 170 | 1840 | 205 | 2038 | 180 |
| Peugeot 207 | 1580 | 115 | 418 | 172 | 1810 | 220 | 1819 | 175 |
| Citroen C3HDI | 2000 | 138 | 415 | 173 | 2100 | 210 | 2570 | 143 |

Cyl : Cylindrée

Pui : Puissance

Long : Longueur

Larg : Largeur

Poids : Poids

Vmax : Vitesse maximale.

Prix : Prix x$10^3$

EmiCO2 : Emission CO2

Tab. (4.2) : Matrice des corrélations

$$\begin{pmatrix} 1 & 0.69647 & 0.11684 & 0.02962 & 0.10812 & 0.088888 & 0.70757 & 0.33104 \\ 0.69647 & 1 & 0.49384 & 0.27489 & 0.36496 & 0.52918 & 0.81426 & 0.46028 \\ 0.11684 & 0.49384 & 1 & 0.35269 & -0.048789 & 0.49091 & 0.43566 & 0.43523 \\ 0.02962 & 0.27489 & 0.35269 & 1 & 0.579 & 0.42557 & 0.080567 & 0.45983 \\ 0.10812 & 0.36496 & -0.048789 & 0.579 & 1 & 0.40611 & 0.23064 & 0.40471 \\ 0.088888 & 0.52918 & 0.49091 & 0.42557 & 0.40611 & 1 & 0.69075 & 0.38091 \\ 0.70757 & 0.81426 & 0.43566 & 0.080567 & 0.23064 & 0.69075 & 1 & 0.37672 \\ 0.33104 & 0.46028 & 0.43523 & 0.45983 & 0.40471 & 0.38091 & 0.37672 & 1 \end{pmatrix}$$

Tab. (4.3) : Les valeurs propres, inerties et inerties cumulées

| valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 3.7858 | 47.3225 | 47.3225 |
| 1.5549 | 19.43625 | 66.7587 |
| 1.0614 | 13.2675 | 80.0262 |
| 0.7541 | 9.42625 | 89.4525 |
| 0 .4464 | 5.58 | 95.0325 |
| 0.2660 | 3.325 | 98.3575 |
| 0 .1157 | 1.444625 | 99.8021 |
| 0.0156 | 0.1975 | 100 |

Tab. (4.4) : Projections des voitures sur l'espace de dimension 3, contributions absolues et relatives

| Voiture | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| BMW x6 | $(5.0729, 1.9117, 0.1519)$ | 83.34 | 98.38 |
| Toyota Aris | $(0.3262, -0.4481, 2.5882)$ | 58.79 | 82?96 |
| Volvo c30 | $(0.0507, 0.0871, 0.1676)$ | 1.92 | 12.43 |
| Chevrolet captiva | $(1.731, -1.6458, -1.5424)$ | 44.1 | 70.36 |
| Opel Astra | $(-1.6703, -0.1222, 0.1169)$ | 6.88 | 71.78 |
| Nissan Navara | $(-2.5166, 2.5798, -0.5228)$ | 56.45 | 90.86 |
| Hundai I30 | $(-0.1249, -0.3666, -0.9098)$ | 7.89 | 35.23 |
| Passat TDI | $(0.3872, -1.7421, 0.4498)$ | 19.83 | 66.71 |
| Ford monda | $(0.0109, -0.3321, -0.173)$ | 0.89 | 9.03 |
| Renault Laguna | $(-0.8148, 0.5699, 0.6011)$ | 6.57 | 43.21 |
| Peugeot 207 | $(-1.3839, -0.5065, 0.4039)$ | 6.13 | 68.26 |
| Citroen C3HDI | $(-0.9671, 0.015, -0.6961)$ | 6.39 | 66.59 |

Tab. (4.5) : Projections des caractéristiques sur le sous espace de dimension 3, contributions absolues et relatives

| Caractéristiques | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| Cyl | $(0.5648, 0.5833, -0.3829)$ | 48.13 | 87.92 |
| Pui | $(0.8405, 0.2804, -0.0876)$ | 26.66 | 86.47 |
| Long | $(0.5841, -0.0231, 0.684)$ | 57.95 | 88.31 |
| Larg | $(0.5079, -0.6734, -0.037)$ | 39.39 | 77.76 |
| Poids | $(0.4948, -0.532, -0.5304)$ | 55.82 | 88.26 |
| Vmax | $(0.7199, -0.1921, 0.2594)$ | 24.42 | 67.91 |
| Prix | $(0.8039, 0.4251, 0.0018)$ | 31.29 | 90.22 |
| EmiCO2 | $(0.6601, -0.2269, -0.0251)$ | 16.22 | 53.21 |

**Conclusion 3.4.1** *Emission en CO2 est moins important que les autres caractéristiques d'une voiture. La meilleure voiture est la BMWx6, puis la Toyota Aris et Nissan Navara. Volvo c30 et la ford Monda sont les dernières clasées.*

## 3.5 Problème 5 : $4^{ème}$ année ingénieur en informatique (2008/2009).

On considères les moyennes annuelles $X^1, X^2, X^3$ et $X^4$ obtenues par 50 étudiants de la quatrième année ingénieur informatique, depuis le BAC.

$X^1$ est la moyenne obtenue au BAC.

$X^2$ est la moyenne obtenue à la première année universitaire.

$X^3$ est la moyenne obtenue à la deuxième année universitaire.

$X^1$ est la moyenne obtenue à la troisième année universitaire.

Influence de la moyenne d'une année sur les années avenirs.

Tab. (5.1) : Tableau des données

| Etudiants | $X^1$ | $X^2$ | $X^3$ | $X^4$ |
|---|---|---|---|---|
| $I_1$ | 10.80 | 12.34 | 12.38 | 11.08 |
| $I_2$ | 10.60 | 11.32 | 12.59 | 11.69 |
| $I_3$ | 11.00 | 10.00 | 10.93 | 10.94 |
| $I_4$ | 11.83 | 10.22 | 12.09 | 11.11 |
| $I_5$ | 12.50 | 10.11 | 1.33 | 10.47 |
| $I_6$ | 12.82 | 10.39 | 10.97 | 11.11 |
| $I_7$ | 10.05 | 11.83 | 10.02 | 10.24 |
| $I_8$ | 10.63 | 10.00 | 10.00 | 10.98 |
| $I_9$ | 11.34 | 11.11 | 12.26 | 12.25 |
| $I_{10}$ | 10.69 | 10.48 | 10.01 | 10.98 |
| $I_{11}$ | 10.75 | 10.23 | 12.03 | 10.65 |
| $I_{12}$ | 10.19 | 10.00 | 12.27 | 10.34 |
| $I_{13}$ | 10.34 | 10.47 | 10.53 | 10.24 |
| $I_{14}$ | 10.15 | 10.46 | 10.50 | 10.60 |
| $I_{15}$ | 10.96 | 12.91 | 13.52 | 13.94 |
| $I_{16}$ | 11.00 | 10.00 | 10.73 | 10.74 |
| $I_{17}$ | 12.11 | 10.03 | 10.00 | 10.44 |
| $I_{18}$ | 11.83 | 10.13 | 10.53 | 10.87 |
| $I_{19}$ | 11.11 | 13.05 | 10.53 | 10.89 |
| $I_{20}$ | 11.91 | 10.10 | 10.56 | 10.54 |

Suite du tableau (5.1)

| Inds | $X^1$ | $X^2$ | $X^3$ | $X^4$ |
|---|---|---|---|---|
| $I_{21}$ | 10.97 | 10.65 | 10.25 | 10.00 |
| $I_{22}$ | 11.85 | 10.03 | 10.40 | 10.08 |
| $I_{23}$ | 10.19 | 12.75 | 10.49 | 11.11 |
| $I_{24}$ | 13.00 | 10.00 | 10.50 | 10.00 |
| $I_{25}$ | 11.21 | 10.95 | 11.64 | 11.71 |
| $I_{26}$ | 12.01 | 11.97 | 11.30 | 11.99 |
| $I_{27}$ | 10.73 | 11.11 | 11.23 | 10.93 |
| $I_{28}$ | 12.27 | 10.21 | 10.21 | 12.32 |
| $I_{29}$ | 11.77 | 10.96 | 10.99 | 11.83 |
| $I_{30}$ | 10.15 | 10.46 | 10.50 | 10.60 |
| $I_{31}$ | 10.27 | 10.17 | 10.05 | 11.02 |
| $I_{32}$ | 12.13 | 13.96 | 11.27 | 11.63 |
| $I_{33}$ | 10.56 | 10.03 | 12.07 | 11.36 |
| $I_{34}$ | 12.25 | 11.17 | 11.07 | 11.30 |
| $I_{35}$ | 12.56 | 13.08 | 11.50 | 11.98 |
| $I_{36}$ | 12.38 | 11.53 | 10.86 | 10.57 |
| $I_{37}$ | 10.75 | 10.25 | 11.28 | 11.17 |
| $I_{38}$ | 10.86 | 10.04 | 11.24 | 10.95 |
| $I_{39}$ | 10.34 | 10.80 | 10.46 | 10.69 |

Suite du Tab. (5.1)

| Inds | $X^1$ | $X^2$ | $X^3$ | $X^4$ |
|---|---|---|---|---|
| $I_{40}$ | 11.11 | 10.24 | 10.45 | 10.75 |
| $I_{41}$ | 10.25 | 11.22 | 11.11 | 10.73 |
| $I_{42}$ | 11.00 | 12.30 | 11.25 | 11.75 |
| $I_{43}$ | 10.93 | 11.80 | 12.25 | 10.64 |
| $I_{44}$ | 11.30 | 11.19 | 10.37 | 10.19 |
| $I_{45}$ | 11.34 | 11.17 | 12.43 | 10.02 |
| $I_{46}$ | 12.16 | 12.71 | 11.10 | 10.98 |
| $I_{47}$ | 11.81 | 12.64 | 12.27 | 11.89 |
| $I_{48}$ | 10.00 | 10.00 | 11.79 | 10.75 |
| $I_{49}$ | 12.16 | 10.46 | 11.87 | 12.87 |
| $I_{50}$ | 12.15 | 11.37 | 10.47 | 10.96 |

Tab. (5.2) : Matrice des corrélations.

$$\begin{pmatrix} 1 & 0,0904 & -0,0577 & 0,1787 \\ & 1 & 0,2797 & 0,3921 \\ & & 1 & 0,4887 \\ & & & 1 \end{pmatrix}$$

Tab. (5.3) : Valeurs propres, Inerties et Inerties cumulées

| Valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 1.8002 | 45.0040 | 45.0040 |
| 1.0308 | 25.9502 | 70.9542 |
| 0.7140 | 17.8508 | 88.8050 |
| 0.4478 | 11.2050 | 100 |

Tab. (5.4) : Projections des étudiantss sur l'espace de dimension 3, contributions absolues et relatives

| Etudiants | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| $I_1$ | $(-1.4333, -1.0580, -0.4980)$ | 5.1336 | 79.3408 |
| $I_2$ | $(-1.5276, -1.3831, 0.5742)$ | 7.2021 | 98.5826 |
| $I_3$ | $(0.7926, -0.2872; 0.5909)$ | 1.8349 | 92.1948 |
| $I_4$ | $(-0.4037, 0.1708, 1.3218)$ | 5.1313 | 80.5514 |
| $I_5$ | $(0.5593, 1.2341, 0.9667)$ | 5.8997 | 74.7169 |
| $I_6$ | $(0.0682, 1.8231, 0.8241)$ | 8.3111 | 98.5097 |
| $I_7$ | $(1.2542, -0.9455, -1.8443)$ | 12.9983 | 98.7518 |
| $I_8$ | $(1.4587, -0.2933, -0.0057)$ | 2.5300 | 63.4660 |
| $I_9$ | $(-1.8102, -0.3501, 0.9260)$ | 6.2785 | 97.7885 |
| $I_{10}$ | $(1.1993, -0.2079, -0.3703)$ | 2.0653 | 60.9815 |
| $I_{11}$ | $(0.2264, -0.2079, 0.8584)$ | 4.3708 | 78.8087 |
| $I_{12}$ | $(0.5478, -1.8665, 0.9346)$ | 9.4927 | 83.2290 |
| $I_{13}$ | $(1.5358, -0.8989, -0.3950)$ | 4.6146 | 99.9335 |
| $I_{14}$ | $(1.3010, -1.0790, -0.3415)$ | 4.4505 | 92.7041 |
| $I_{15}$ | $(64.8835, -1.1547, 0.5989)$ | 30.0698 | 97.1712 |
| $I_{16}$ | $(1.0926, -0.2118, 0.4187)$ | 1.9038 | 95.8120 |
| $I_{17}$ | $(1.5966, 1.3760, 0.1869)$ | 6.5783 | 99.7759 |
| $I_{18}$ | $(0.8897, 0.8504, 0.4593)$ | 2.8637 | 97.3859 |
| $I_{19}$ | $(-0.4490, 0.1475, -2.0675)$ | 12.2385 | 99.9764 |
| $I_{20}$ | $(1.1416, 0.9041, 0.4187)$ | 3.5138 | 99.0514 |

Suite du tableau (5.4) :Projections des étudiantss sur l'espace de dimension 3, contributions absolues et relatives

| Etudiants | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| $I_{21}$ | $(1.7074, -0.0560, -0.6050)$ | 4.2702 | 98.4225 |
| $I_{22}$ | $(1.6768, 0.8706, 0.2288)$ | 4.7306 | 94.7553 |
| $I_{23}$ | $(-0.2709, -0.8927, -2.0190)$ | 13.0345 | 93.7891 |
| $I_{24}$ | $(1.4625, 2.1450, 0.5815)$ | 12.1883 | 88.0589 |
| $I_{25}$ | $(-0.8403, -0.2674, 0.5118)$ | 1.6560 | 91.9438 |
| $I_{26}$ | $(-1.5136, 0.8723, -0?2013)$ | 4.1251 | 95.7856 |
| $I_{27}$ | $(0.0956, -0.6861, -0.2086)$ | 1.0391 | 99.8531 |
| $I_{28}$ | $(-0.2216, 1.6095, 0.7810)$ | 6.7547 | 53.2470 |
| $I_{29}$ | $(-0.6183, 0.6796, 0?3275)$ | 1.6150 | 66.4607 |
| $I_{30}$ | $(1.3010, -1.0790, 60.3415)$ | 4.4505 | 92.7041 |
| $I_{31}$ | $(1.1797, 0.4326, 0.0636)$ | 1.9182 | 66.8313 |
| $I_{32}$ | $(-2.2179, 1.0842, -1.8942)$ | 17.8115 | 98.7718 |
| $I_{33}$ | $(-0.2496, -1.2752, 1.2115)$ | 7.3134 | 99.9627 |
| $I_{34}$ | $(-0.4344, 1.1681, 0.1634)$ | 2.9137 | 99.6225 |
| $I_{35}$ | $(-2.3060, 1.4641, -0.8434)$ | 13.0305 | 99.6741 |
| $I_{36}$ | $(0.1042, 1.3753, -0.4337)$ | 4.1837 | 79.1407 |
| $I_{37}$ | $(0.2910, -0.7054, 0.5894)$ | 2.0259 | 90.4725 |
| $I_{38}$ | $(0.5834, -0.5855, 0.6966)$ | 2.3977 | 98.6261 |
| $I_{39}$ | $(1.0452, -0.8208, -0.5596)$ | 3.3889 | 90.1647 |

Suite du tab. (5.4) : Projections des étudiantss sur l'espace de dimension 3, contributions absolues et relatives

| | | | |
|---|---|---|---|
| $I_{40}$ | $(1.1305, 0.0520, 0.1030)$ | 1.4548 | 89.9562 |
| $I_{41}$ | $(0.3819, -1.1955, -0.5495)$ | 3.7615 | 99.9976 |
| $I_{42}$ | $(-1.2466, -0.2732, -0.8302)$ | 3.8008 | 91.8070 |
| $I_{43}$ | $(-0.7368, -0.9047, -0.2382)$ | 2.3389 | 48.4868 |
| $I_{44}$ | $(1.1331, 0.3077, -0.8284)$ | 3.5311 | 94.7983 |
| $I_{45}$ | $(-0.1104, -0.5845, 0.2819)$ | 0.8945 | 9.8971 |
| $I_{46}$ | $(-0.9448, 1.0943, -1.1798)$ | 7.1976 | 89.2351 |
| $I_{47}$ | $(-2.3801, 0.2277, -0.2874)$ | 6.6250 | 96.2985 |
| $I_{48}$ | $(0.5695, -1.8405, 0.7473)$ | 8.4516 | 99.3807 |
| $I_{49}$ | $(-1.8972, 0.7875, 1.6375)$ | 12.7040 | 86.7166 |
| $I_{50}$ | $(0.1700, 1.3064, -0.4593)$ | 3.9117 | 99.9775 |

Tab. (1.5) : Projections des moyennes sur le sous espace de dimension 3, contributions absolues et relatives

| moyennes | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| $X^1$ | $(-0.2138, 0.9436, 0.1759)$ | 92.6606 | 96.7134 |
| $X^2$ | $(-0.6981, 0.0457, -0.7042)$ | 96.7174 | 98.5300 |
| $X^3$ | $(-0.7443, -0.3773, 0.3843)$ | 65.17199 | 84.4040 |
| $X^4$ | $(-0.8444, 0.0559, 0.1989)$ | 45.4502 | 75.5727 |

**Conclusion 3.5.1** *La moyenne du BAC et la moyenne de la première année ont une très grande importance sur les trois années d'étude utiversitaire. L'étudiant n° 15 est le meilleuilleur étudiant.*

## 3.6 Problème 6 : $2^{ème}$ année STID (2013/2014).

Classification des eaux minirales de différentes sources d'eau selon leurs composition et prix.

Différentes eux minérales sont considérées et pour chacune on donne sa composition son prix. Le but est de déterminer

1/ la caractéristiques de chacune des eaux .

2/ les composants les plus importants.

3/ les eaux qui sont proches au sens de leurs contenances en matières minérales.

Tab. (6.1) : Tableau des données

| | Na | Mg | Ca | Potassium | Sulfates | Cl | Nitrates | Résiuds | Ph |
|---|---|---|---|---|---|---|---|---|---|
| Ifri | 15.8 | 24 | 81 | 2.1 | 73 | 72 | 15 | 380 | 7.2 |
| Toudja | 32.9 | 14.4 | 55.4 | 0.6 | 16.2 | 60.2 | 1.01 | 334 | 7.43 |
| Lala Kha | 5.5 | 7 | 53 | 0.54 | 7 | 11 | 0.42 | 187 | 7.22 |
| Ayris | 28.5 | 6.8 | 65.6 | 1.9 | 75 | 37 | 2.7 | 276 | 7.78 |
| Texana | 11 | 9.1 | 30 | 1 | 11 | 28.4 | 0 | 152 | 7 |
| Nestlé | 12 | 16.3 | 57.9 | 0.5 | 31 | 15 | 8 | 300 | 7.6 |
| Djurdj. | 52 | 28 | 74 | 2 | 50 | 82 | 19.8 | 551 | 7.2 |
| Cordial | 23.33 | 15.98 | 94.47 | 1.98 | 84 | 70 | 10 | 410 | 7.25 |
| Togi | 36 | 19.25 | 73.41 | 1.8 | 28.9 | 43.76 | 5.93 | 366 | 7.46 |
| Boura. | 22.8 | 13.47 | 62.88 | 2.1 | 37.6 | 23.1 | 0.02 | 285 | 7.44 |
| Ifren | 32 | 10.69 | 68.8 | 2.4 | 62.5 | 17.04 | 3.22 | 300 | 7.48 |
| N'Gaous | 63.4 | 65.4 | 143 | 3.76 | 44.4 | 75 | 2.07 | 962 | 7.66 |
| Ovitale | 30 | 14 | 91 | 1 | 86 | 50 | 15 | 420 | 6.92 |

Tab. (6.2) Matrice des corrélations

$$\begin{pmatrix}
1 & 0.7525 & 0.7045 & 0.6978 & 0.2162 & 0.6333 & 0.2073 & 0.8547 & 0.3081 \\
 & 1 & 0.8307 & 0.7166 & 0.0646 & 0.6068 & 0.1399 & 0.9575 & 0.2573 \\
 & & 1 & 0.7486 & 0.5018 & 0.6117 & 0.2702 & 0.9143 & 0.2244 \\
 & & & 1 & 0.4072 & 0.4589 & 0.0459 & 0.7307 & 0.1720 \\
 & & & & 1 & 0.4095 & 0.5761 & 0.2573 & -0.0364 \\
 & & & & & 1 & 0.5839 & 0.6870 & -0.1340 \\
 & & & & & & 1 & 0.2594 & -0.4133 \\
 & & & & & & & 1 & 0.2449 \\
 & & & & & & & & 1
\end{pmatrix}$$

Tab. (6.3) Valeurs propres, Inerties, Inerties cumulées

| Valeurs propres | Inerties (%) | Inerties cumulées (%) |
|---|---|---|
| 4.9869 | 55.4104 | 55.4104 |
| 1.8878 | 20.9756 | 76.3860 |
| 0.9276 | 10.3061 | 86.6921 |
| 0.4447 | 4.9411 | 91.6332 |
| 0.3114 | 3.4599 | 95.0931 |
| 0.2231 | 2.4787 | 97.5718 |
| 0.1872 | 2.0803 | 99.6521 |
| 0.0286 | 0.3180 | 99.9701 |
| 0.0027 | 0.0299 | 100.00 |

Tab. (6.4) : Projection des eaux sur le sous espace de dimension 3, contributions absolues et relatives.

| Caractéristiques | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| Ifri | $(0.8805, -1.8449, 0.2822)$ | 15.7256 | 76.7714 |
| Toudja | $(-0.9462, 0.6941, -1.1237)$ | 13.8170 | 56.4696 |
| Lala Khadidja | $(-3.0954, 0.6564, -0.7817)$ | 21.6028 | 93.7330 |
| Ayris | $(-0.3648, 0.8192, 2.0263)$ | 36.9903 | 85.7910 |
| Texana | $(-2.9884, 0.0610, -1.2765)$ | 27.3036 | 89.7790 |
| Nestlé | $(-1.7659, 0.5684, 0.1854)$ | 6.4120 | 55.2658 |
| Djurdjura | $(2.1980, -1.7072, -1.0354)$ | 28.2192 | 81.4064 |
| Cordial | $(1.0336, -1.4393, 0.7925)$ | 15.2974 | 83.0090 |
| Togi | $(0.1211, 0.6435, -0.2885)$ | 2.4001 | 55.6283 |
| Bourached | $(-0.9727, 1.1370, 0.4706)$ | 8.5634 | 84.1998 |
| Ifren | $(-0.3636, 0.7221, 1.3554)$ | 17.5637 | 72.3199 |
| N'Gaous | $(5.8509, 2.2907, -0.6964)$ | 78.2068 | 98.6804 |
| Ovitale | $(0.4128, -2.6011, 0.0898)$ | 27.8982 | 81.6200 |

Tab. (6.5) : Projection des caractéristiques sur le sous espace de dimension 3, contributions absolues et relatives.

| Caractéristiques | Proj | $C_a\,(1,2,3)$ (%) | $C_r\,(1,2,3)$ (%) |
|---|---|---|---|
| Sodium | $(0.8660, 0.1616, -0.1223)$ | 18.0333 | 79.1002 |
| Magnésium | $(0.8882, 0.2426, -0.2989)$ | 28.5689 | 93.7116 |
| Calcium | $(0.9265, 0.0201, 0.1019)$ | 18.3546 | 86.9190 |
| Potassium | $(0.8232, 0.2414, 0.2569)$ | 23.7898 | 80.1860 |
| Sulfate | $(0.4532, -0.5514, 0.6823)$ | 70.4131 | 97.4967 |
| Clorure | $(0.7671, -0.4081, -0.2330)$ | 26.4747 | 80.9237 |
| Nitrate | $(0.3635, -0.8421, -0.0035)$ | 40.2154 | 84.1303 |
| Résidus secs | $(0.9624, 0.1194, -0.1785)$ | 22.7655 | 97.2371 |
| ph | $(0.2447, 0.7418, 0.4417)$ | 51.3847 | 80.5243 |

**Conclusion 3.6.1** *Le sulfate est l'élément le plus représentatif des différentes eaux étudier, puis ph ensuite le Nitrate, les autes éléments sont presque d'égals importance. L'eau minérale la plus riche est N'Gaous et la plus pauvre est Togi.*

# Conclusion

L'analyse en composantes principales est une méthode puissante pour synthétiser et résumer les populations de grandes tailles qui sont décrites par plusieurs variables quantitatives.

L'ACP permet, entre autre de dégager des catégories d'individus et de réaliser un bilan de liaison entre variables. Ainsi, on peut mettre en évidence, les grandes tendances des données tel que le regroupement d'individus qui se ressembles ou opposition entre individus, ce qui traduit un comportement radicalement différent des individus. Ou encore, opposition entre variable, qui traduit le fait que les variables sont inversement corrélées.

Les représentations graphiques obtenues sont simples et riches en informations.

Schématiquement, l'analyse en composante principale est un changement de repère qui vise à préviligier les axes de variance maximale par rapport à des ensembles de données. La transformation est par essence linéaire. Dans certains cas, en pratique, les données ont une structure compliquée qui ne peuvent pas être simplifier dans des espaces linéaires ( comme exemple, application en vidéo), il est souhaitable de pouvoir atteindre des relations non linéaires. La méthode "Kernel CPA" ACP à noyau est une première extention qui l'envisage. Cette dernière est une généralisation de la méthode ACP traditionnelle qui consiste à projeter les individus en utilisant une fonction noyau qui n'est pas linéaire.

L'ACP peut être une première étape pour réaliser une analyse discriminante ou une classification automatique des données.

# Bibliographie


[1] J. P. Benzekri. Analyse des données (Tome 1). La taxinomie. Dunod. (1980).

[2] J. P. Benzekri. Analyse des données (Tome 2). L'analyse des correspondances. Dunod. (1980).

[3] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier et H. Ralambondrainy. Classification automatique des données. Dunod (1989).

[4] B. Escoffier et J. Pagées. Analyse factorielle simple et multiples objectifs, Méthodes et interprétations. dunod. (1990)

[5] G. Hebrail et Y. Lechevallier. Analyse des données, chapitre Data Mining et analyse des données. Hermes sciences publications. (2003).

[6] L. Lebart, A. Morineau et M. Piron. Statistique exploratoire multidimensionnelle. Dunod. (1995).

[7] A. Martin. L'analyse de données. Polycopier de cours ESNIETA- ref 1463. (2003).