utfcode
utf8.sty 3.10 UTF-8 input encoding 13.06.2000
scanner for code UTF-8 installed.
utf8
4.137figure.caption.40

$N^o d'ordre$: ....

République Algerienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la Recherche scientifique
Université Mouloud Mammeri de Tizi Ouzou
Faculté des Sciences
Département de Mathématiques
Laboratoire (L2CSP)

Mémoire de Master

Filière : Mathématiques
Spécialité : Mathématiques Appliqué à la Gestion

Par

Mansouri Cylia
Sellami Hacene

Analyse en Composantes Principales

Soutenue le Sept 2022 devant le jury :

Dr.	BOUALAM KARIMA	UMMTO	Président du jury
Dr.	ATIL LYNDA	UMMTO	Examinateur
Dr.	MEHIRI MOHAMED	UMMTO	Examinateur
Dr.	AMIROU AHMED	UMMTO	Encadreur

Année Universitaire : 2021/2022

DÉDICACES

A mon cher pére AREZKI
Pour l'amour et l'éducation qu'il m'a donnée .
A ma chere mère KHADOUDJA
Pour son grand amour, ses sacrifices et toute l'affection qu'elle m'a toujours offerte
A mes fréres ESSAID,SAID,RACHID,MOHAMED et ma soeur LYNDA
A mes amis (amies)
A mon binôme CYLIA
A tous les étudiants de mathématique

HACENE

DÉDICACES

A mon cher pére AREZKI
Pour l'amour et l'éducation qu'il m'a donnée .
A ma chere mère KHADOUDJA
Pour son grand amour, ses sacrifices et toute l'affection qu'elle m'a toujours offerte
A mes fréres ESSAID,SAID,RACHID,MOHAMED et ma soeur LYNDA
A mes amis (amies)
A mon binôme CYLIA
A tous les étudiants de mathématique

HACENE

Remerciements

Nous remercions chaleureusement et spécialement notre promoteur **M.Amirou** pour son aide, ses conseils, son entière disponibilité tout au long de notre projet et pour la bonne ambiance de travail qu'il a su créer.

Nos remerciements vont également :

Aux membres du jury de la soutenance, qui nous ont fait l'honneur de juger nos travail.

A nos familles et amis respectifs et spécialement nos très chères mamans, pour le soutien inconditionnel qu'elles nous ont apporté tout au long de nos études.

Enfin, à toute personne qui, d'une façon ou d'une autre, a contribué à la réalisation de cette mémoire.

Merci à Tous

Tizi-Ouzou, le 27.09.2022

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

L'analyse en composantes principales est une méthode statistique exploratoire permettant une description essentiellement graphique de l'information contenue dans de grands tableaux de données. Dans la plupart des applications, il s'agit d'étudier p variables mesurées sur un ensemble de n individus. Lorsque n et p sont grands, on cherche à synthétiser la masse d'information sous une forme exploitable et compréhensible. Grâce aux outils de la statistique descriptive, il est possible d'étudier une à une ou deux à deux les variables à travers notamment des résumés graphique ou numérique (moyenne,la variance,corrélation). Cependant, ces études préalables simples, si elles sont indispensables dans toute étude statistique, sont insuffisantes ici par elles laissent de coté les liaisons éventuelles entre le variables qui sont souvent l'aspect le plus important.

L'analyse en composantes principales, notée ACP par la suite, est souvent considérée comme la méthode de base de l'analyse factorielle des données dont l'objectif est de déterminer des fonctions des p variables ou facteurs qui serviront à visualiser les observations de façon simplifiée. En ramenant un grand nombre de variables, souvent corrélées entre elles, à un petit nombre de composantes principales (les premières) non corrélées, l'ACP est une méthode de réduction de la dimension. Nous verrons cependant que l'ACP peut aussi être utilisée comme une méthode de détection de valeurs aberrantes multidimensionnelles, notamment par l'exploitation des dernières composantes. Cette propriété se révèle utile en contrôle de qualité multidimensionnel.

Dans ce mémoire, on a élaboré un plan de travail qui se compose de deux parties, une partie théorique et une partie pratique. Dans un premier lieu on a cité les différents rappels algébriques et principes d'analyse en composantes principales (ACP), en seconde partie l'application de l'ACP sur des données réelles "La température mensuelle de 15 villes de France sur 30 ans" sous logiciel R.

Rappels d'algèbre linéaire et de géométrie euclidienne 1

1.1 Les matrices

1.1.1 Notation

On appelle matrice de dimension $n \times p$ un tableau de nombres comportant n lignes et p colonnes, (voir Gergaid [octobre 2006]).

$$\begin{array}{cl} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{\mathrm{np}} \end{pmatrix} & \begin{array}{l} \\ \longleftarrow \textit{ligne } 2 \\ \\ \longleftarrow \textit{ligne } n \end{array} \end{array}$$

• Toute matrice s'écrit sous la forme ci-dessus. Les nombres a_{ij} sont appelés coefficients de la matrice.

•Le coefficient de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne est noté a_{ij}.

1.1.2 Type de matrice

Une matrice est dite :

•Vecteur ligne (colonne) si $n = 1$ $(p = 1)$,

•Vecteur-unité d'ordre p si elle vaut $1_p = [1 \ldots..1]^{'}$,

•scalaire c'est une matrice diagonale telle que :a_{ii}=λ; $\lambda \in \mathbb{R}$,

•carrée si $n = p$,

Une matrice carrée est dite :

•Identité (I_n) est une matrice scalaire telle que, $\lambda = 1$,

•Diagonale si : a_i^j =0 et a_i^i=λ_{i} ; $\forall$: i, λ_{i} $\in \mathbb{R}$,

•Triangulaire supérieure (inférieure) si : a_i^j=0 lorsque $\frac{i>j}{i<j}$.

•Symétrique : une matrice carrée A est symétrique si et seulement si elle est égale à sa transposée $(A^{'} = A)$.

•Matrice orthogonale : c'est une matrice carée telle que ; $A^t.A = A.A^t = I_n$,

1.1.3 Opération sur les matrices

-Somme de deux matrice

On appelle somme de deux matrices de même dimension la matrice obtenue en additionnant les coefficients qui ont la même positions .

$$\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{\mathrm{np}} \end{pmatrix} + \begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{\mathrm{np}} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & \cdots & a_{1p}+b_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1}+b_{1n} & \cdots & a_{\mathrm{np}}+b_{\mathrm{np}} \end{pmatrix}$$

Remarque 1.1 *On ne peut pas additionner deux matrices de dimensions différentes.*

-Produit d'un scalaire par une matrice

Soit une matrice A telle que $A = (a_{ij})$ et un scalaire α, on peut multiplier une matrice A par un scalaire α

$$\alpha A = \left(\alpha a_{ij}\right) = \begin{pmatrix} \alpha a_{11} & \cdots & \alpha a_{1n} \\ \alpha a_{12} & \ddots & \vdots \\ \ddots & \vdots & \vdots \\ \alpha a_{n1} & \cdots & \alpha a_{nn} \end{pmatrix}$$

Remarque 1.2 *On peut pas écrire le point de la multiplication, ainsi on écrit αA mais plutôt que $\alpha.A$.*

Le scalaire s'écrit toujours à la gauche, ainsi on écrit αA mais surtout pas $A\alpha$.

De même on écrit $\frac{1}{\alpha}A$ mais surtout $\frac{A}{\alpha}$.

-Transposition de matrice

Soit $A = (a_{ij})$ une matrice de type (n,p). On appelle transposée de A la matrice $B = (b_{ij})$ de type (p,n) définie par :

$$b_{ij} = a_{ji}$$

.

On notra B=$A^{'}$.

Soient A et B deux matrices quelconqueq.On a les propriétés suivantes :

1.$(A^{'}) = A$

2.$(A+B)^{'} = A^{'} + B'$

3.$(AB)^{'} = B^{'}A'$

-Produit de deux matrices

Soit une matrice de dimension $n \times p$ et soit B une matrice de dimension $p \times m$.

On appelle le produit A×B la matrice X de dimension $n \times m$ où chaque coefficient x_{ij} est le produit de la i^{me} ligne de A par la j^{me} colonne de B.

Autrement dit : si on pose $A = (a_{\text{ij}})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ et $B = (b_{\text{ij}})_{\substack{1 \leq i \leq p \\ 1 \leq j \leq m}}$

$A \times B = (x_{\text{ij}})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$

Avec : $x_{ij=} \sum_{k=1}^{p} a_{\text{ik}} b_{\text{kj}}$

Remarque 1.3 *La matrice $A \times B$ n'est définie que si le nombre de colonnes de A est égale au nombre de lignes de B.*

Exemple : $\begin{pmatrix} 1 & 2 & -3 \\ 4 & -5 & 6 \end{pmatrix} \times \begin{pmatrix} 7 & 8 \\ 9 & -1 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} 16 & -9 \\ 1 & 67 \end{pmatrix}$

Propriété 1.1 • *La multiplication de matrices n'est pas commutative :* $A \times B \neq B \times A$

•*La multiplication de matrices est associative :* $(A \times B) \times C = A \times (B \times C)$

•*La multiplication de matrices est distributive par rapport à l'addition :*
$A \times (B + C) = A \times B + A \times C$

1.1.4 Propriétés des matrices carrées

-Trace

Par définition, si A est une matrice carrée : trace $(A) = \sum_{j=1}^{p} a_{\text{jj}}$

Propriété 1.2 *trace* $(\alpha) = \alpha$ *;*

trace $(A + B) =$ *trace* $(A) +$ *trace* (B) *;*

Soit A une matrice de type $(n \times p)$ et B une matrice de type $(p \times n)$ alors :
trace $(AB) =$ *trace* (BA)

-Comatrice(Matrice,adjointe)

Soit $A = (a_{\text{ij}})$ une matrice carrée d'ordre n et Δ_{ij} le cofacteur de l'élément a_{ij}

Définition 1.1 *On appelle comatrice (ou matrice adjointe) de A, la matrice carrée d'ordre n, notée com (A) (ou $Adj(A)$) définie par :*

$$\text{com}\,(A) = \begin{pmatrix} \Delta_{11} & \Delta_{12} & \cdots & \cdots & \Delta_{1n} \\ \Delta_{21} & \Delta_{22} & \cdots & \cdots & \Delta_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Delta_{n1} & \Delta_{n2} & \cdots & \cdots & \Delta_{\text{nn}} \end{pmatrix}$$

Où Δ_{ij} est le cofacteur de l'élément a_{ij} de A défini à partir du mineur $\left|M_{\text{ij}}\right|$ par la relation : $\Delta_{\text{ij}} = (-1)^{i+j} \left|M_{\text{ij}}\right|$.

On appelle mineur $\left|M_{\text{ij}}\right|$ de l'élément a_{ij} du déterminant d'ordre n , le déterminant d'ordre $(n-1)$ obtenue en supprimant la $i^{\text{ème}}$ ligne et la $j^{\text{ème}}$ colonne de $|A|$.

-Déterminant et l'inverse d'une matrice

Théorème 1.1 *Soit une matrice* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$,

On appelle déterminant de la matrice A, noté $(\det A)$, *le réel* $ad - bc$.

Si $\det(A) \neq 0$ *,alors*A *admet une matrice inverse unique noté* A^{-1} *définie par :*

$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Où : $A^{-1} = \frac{com(A)'}{det|A|}$

Remarque 1.4 $A.A^{-1} = A^{-1}.A = I_n$

Exemple 1.1 *Calculer de la matrice inverse de* $A= \begin{pmatrix} 2 & -4 \\ -3 & 1 \end{pmatrix}$

$A= \begin{pmatrix} 2 & -4 \\ -3 & 1 \end{pmatrix}; |A| = \begin{vmatrix} 2 & -4 \\ -3 & 1 \end{vmatrix} = 10;\ com\ (A) = \begin{pmatrix} 1 & 3 \\ 4 & 2 \end{pmatrix};$

$(t_{com}(A) = \begin{pmatrix} 1 & 4 \\ 3 & 2 \end{pmatrix} \Rightarrow A^{-1} = -\frac{1}{10} \begin{pmatrix} 1 & 4 \\ 3 & 2 \end{pmatrix}$.

1.2 Valeurs et Vecteurs propres

Définition 1.2 *(Valeur propre)*
Soit A *une matrice carrée , alors* $\lambda \in R$ *est une valeur propre de* A *si et seulement si'l existe un vecteur non nul* x *telle que* $Ax=\lambda x$

Théorème 1.2 *Soit* A *une matrice carrée alors* λ *est une valeur propre de* A *si et seulement si* $Det(A - \lambda\ I_n) = 0$

Où $Det(A - \lambda\ I_n)$ *est le déterminant de la matrice* $(A - \lambda\ I_n)$*, et*I_n *représente la matrice identité d'ordre* n.

Définition 1.3 *(Vecteur propre)*
Soit A *une matrice carrée et* λ *une valeur propre de* A.

Alors tout vecteur x *non nul tel que* $Ax = \lambda x$ *est un vecteur propre associé à la valeur propre* λ.

Définition 1.4 *(Espace propre)*
Soit A *une matrice carrée et* λ *une valeur propre .On appelle espace propre* V_λ *le sous-espace vectoriel* $\ker(A - \lambda I_n)$.

Définition 1.5 *(Ordre de multiplicité d'une valeur propre)*

Soit A*une matrice carrée d'ordre* n *et* λ *une valeur propre .On appelle ordre de multiplicité de la valeur propre* λ *la dimension de l'espace propre* V_λ .

1.3 ESPACES EUCLIDIENS

1.3.1 Produit scalaire

Un espace vectoriel de dimension p est dit euclidien, s'il est muni d'un produit scalaire qui est définit par :

$$f(x,y) = < x,y >, \forall x,y \in E$$

Où f est une forme bilinéaire , symétrique et définie positive c'est-à-dire :

$$\forall \alpha, \beta \in \mathbb{R},\ \forall x,y,z \in E :< \alpha x + \beta y, z > \geq \alpha < x,z > +\beta < y,z >$$

$< x,y >=< y,x >, \forall x,y \in E$.

$< x,x >= \| x \|^2 \geq 0, \forall x \in E - \{0\}$

-Représentation matricielle du produit scalaire

Soit $e_1, \ldots, e_p$ une base de E , les vecteurs $x = (\alpha_1, \ldots, \alpha_p)$ et $y = (\beta_1, \ldots, \beta_p)$ s'écrivent alors :

$x = \sum_{i=1}^{p} \alpha_i\ e_i$, $y = \sum_{i=1}^{p} \beta_i\ e_i$.

Grâce à la bilinéarité de f on écrira :

$< x,y >= \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \beta_j e_i e_j$.

La matrice M de terme général $< e_i, e_j >$ est appelée métrique. Le produit scalaire s'écrira :
$< x,y >_M = x'My$

M est symétrique définie positive.

On dit que x et y sont M-orthogonaux si $< x,y >_M =0$

Dans ce cas :

$$\| x+y \|_M^2 = \| x \|_M^2 + \| y \|_M^2$$

1.3.2 Métrique euclidienne

Soit M une matrice carrée $(p \times p)$, symétrique, définie –positive, M définit sur l'espace E :

-Un produit scalaire : $< x.y >_M = x' My$,

-Une norme : $\| x \|_M = < x,x >_M^{\frac{1}{2}}$

-Des angles : $\cos \theta_M (x,y) = \frac{< x.y >_M}{\|x\|_M \|y\|_M}$.

-Une distance : $d_M(x,y) = \| x - y \|_M$.

Étant donnée une matrice M, on dit que :
-Une matrice A est M-symétrique si $(MA)' = MA$,

-Un vecteur x et M-normé si $\parallel x \parallel_M = 1$,

-Une base $\epsilon_q = e_1, \quad . \ . \ . \ . \ , e_q$ est $M-$orthonormée si :

$\forall\,(i,j)\,, < e_i, e_j >_M = \delta_i^j$.

1.3.3 Projection

Soit W un sous-espace de E et $\beta = b^1, \ . \ . \ . \ . \ , b^q$ une base de W , $P\,(p \times p)$ est une matrice de projection $M-$orthogonale sur W si et seulement si :

$$\begin{cases} \forall y \in E, \mathrm{Py} \in W \\ \text{et} \\ < \mathrm{Py}, y - \mathrm{Py} >_M = 0 \end{cases}$$

Toute matrice idempotente ($P^2 = P$) et $M-$symétrique ($P^{'}M = \mathrm{MP}$) est une matrice de projection $M-$orthogonale et réciproquement.

Propriété 1.3 *-Les valeurs propres de P sont 0 et1 :*

$u \in W,\ Pu = u\ , \lambda = 1,$ *de multiplicité Dim* (W) ,

$v \perp W,\ \ Pv = 0\ \ ,\ \lambda = 0,$ *de multiplicité Dim* $(W^{\perp}),$ (*on note* $(v \in W^{\perp})$),

trace $(P) = $ *Dim* (W) ,

$P = B(B^{'}MB)^{-1}B^{'}M,$ *où* $B = [b^1, . \ . \ . \ . \ ., b^q]$.

-Dans le cas particulier où les b^j sont M-orthonormés :

$P = BB'M = \sum_{i=1}^{i=q} b^j b^{j'} M.$

-Dans le cas particulier où $q = 1$ alors :

$P = \frac{bb'}{b'Mb} M = \frac{1}{\|b\|_M} bb'M.$

-Si $P_1, . \ . \ . \ ., P_q$ sont des matrices de projection M-orthogonales alors la somme $P_1 + . \ . \ . \ . \ + \ P_q$ est une matrice de projection M-orthogonale si et seulement si : $P_k P_j = \delta_k^j P_j$.
-La matrice $I - P$ est la matrice de projection $M-$orthogonale sur $W^{\perp}$.

2 Analyse En composantes Principales

2.1 Introduction

L'analyse en composantes principales notée ACP est une méthode d'analyse statistique multivariée, qui a pour but d'étudier simultanément un nombre important de variables quantitatives. L'ACP permet d'obtenir des représentations graphique des distances entre les individus et des corrélations entre les variables, (voir Escofier [1979]).

2.2 L'objectif de l'ACP

- Réduction de la dimension des vecteurs, donner leur approximation avec p variables $(p < P)$. Autrement dit, on cherche à définir p nouvelles variables, combinaisons linéaires des P variables initiales qui feront perdre le moins d'informations possible. Ces nouvelles variables seront appelées **composantes principales**. Les axes qu'elles déterminent sont des axes principaux et les formes linéaires associées sont des facteurs principaux.
- Représentation graphique "optimale" des observations minimisant la déformation du nuage de points dans un sous-espace de dimension $p \ll P$.

2.3 principe de L'ACP, Tableaux de données et Espace associés

2.3.1 Principe de l'ACP

Le principe est simple : Il s'agit en fait de résumer l'information qui est contenue dans une large base de données en un certain nombre de variables synthétiques appelées : Composantes principales.

L'idée est ensuite de pouvoir projeter ces données sur l'hyperplan le plus proche afin d'avoir une représentation simple de nos données.

Évidemment, qui dit réduction de dimension dit perte d'informations. C'est la tout l'enjeu que représente une Analyse en Composantes principales. Il faut pouvoir réduire la dimension de nos données tout en conservant un maximum d'informations, (voir Boumaza [2007]).

2.3.2 Tableau des données

Les observations de p variables sur n individus sont rassemblées en un tableau rectangulaire X à n lignes et p colonnes : (voir Saporta [1980]).

$$X=\begin{pmatrix} x_1^1 & x_1^2 & . & . & x_1^j & . & x_1^p \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ x_i^1 & x_i^2 & . & . & x_i^j & . & x_i^p \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ x_n^1 & x_n^2 & . & . & x_n^j & . & x_n^p \end{pmatrix}$$

On identifie la $j^{\text{ième}}$ variable à la $j^{\text{ième}}$ colonne de x par :

$$x^j=\begin{pmatrix} x_1^j \\ x_2^j \\ . \\ . \\ . \\ x_n^j \end{pmatrix}$$

Ce sont les valeurs prises par x^j $(j = 1, ..., p)$ sur les n individus.
On identifie le $i^{\text{ème}}$ individu à la $i^{\text{ème}}$ ligne de x noté : x_i $(i = 1, 2, . \ . \ . \ ., n)$

avec :

$$x_i=\left(x_i^1, x_i^2, \ldots, x_i^p\right)'$$

2.3.3 Matrice des poids

On suppose que chaque individu est muni d'un poids p_i tel que $p_i \geq 0$ et $\sum_{i=1}^{n} p_i = 1$

Dans certain cas, il est utile de travailler avec des poids p_i différents (données regroupées...).

Ces poids sont regroupés dans la matrice diagonale D de taille n :

$$D=\begin{pmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix}$$

Si les poids sont égaux à $\frac{1}{n}$ on a D= $\frac{1}{n} I_n$ où I_n est la matrice identité.

Preuve

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

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

Par conséquent

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

Et

$$D = \begin{bmatrix} \frac{1}{n} & \cdots & 0 \\ & & \\ 0 & \cdots & \frac{1}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & \cdots & 0 \\ & & \\ 0 & \cdots & 1 \end{bmatrix} = \frac{1}{n} I_n$$

2.3.4 Centre de gravité(le point moyen)

On appellera centre de gravité associé à la matrice des poids D le vecteur g définir par :

$$g = \begin{pmatrix} \overline{x^1} \\ \vdots \\ \overline{x^p} \end{pmatrix}$$

Avec : $\overline{x^j} = \sum_{i=1}^{n} p_i x_i^j, j = 1, \ldots, p.$

Et : $\overline{x^j} = \left(x^j\right)' D1_n,$

$\overline{x^j}$: représente la moyenne arithmétique de la variable j.

La forme matricielle :

$g = X' D1_n$.

Où $1_n = (1, \ . \ . \ ., 1)$ désigne un vecteur de $\mathbb{R}^n$.

preuve

$$X^{'}D1_n = \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ . & . & . \\ . & . & . \\ x_{1p} & \cdots & x_{\text{np}} \end{bmatrix} \begin{bmatrix} p_1 & \cdots & 0 \\ . & . & . \\ . & . & . \\ 0 & \cdots & p_n \end{bmatrix} \begin{bmatrix} 1 \\ . \\ . \\ 1 \end{bmatrix}$$

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

2.4 STANDARDISATION DU TABLEAU

Dans l'analyse en composantes principales les variables sont souvent normalisées. Ceci est particulièrement recommandé lorsque les variables sont mesurées dans différentes unités par exemple : (kilogrammes, kilomètres, centimètres, ...ect) ; sinon, le résultat de lanalyse obtenue sera fortement affecté. L'objectif est de rendre les variables comparables. Généralement, les variables sont normalisées de manière à ce qu'elles aient au final, (voirBouroche [1992])

1. Un écart type égale à un.

2. Une moyenne égale à zéro.

2.4.1 Tableau centré associé à X

Prendre g comme origine de nuage des points, revient alors à travailler sur le tableau de données centrées Y associé à X :

$$Y= \begin{pmatrix} x_1^1 - \overline{x^1} & \cdots & x_1^j - \overline{x^j} & \cdots & x_1^p - \overline{x^p} \\ x_2^1 - \overline{x^1} & \vdots & x_2^j - \overline{x^j} & \cdots & x_2^p - \overline{x^p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^1 - \overline{x^1} & \cdots & x_i^j - \overline{x^j} & \cdots & x_i^p - \overline{x^p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 - \overline{x^1} & \cdots & x_n^j - \overline{x^j} & \cdots & x_n^p - \overline{x^p} \end{pmatrix} = \left(Y^1, \ . \ . \ ., Y^j, . . ., Y^p\right)$$

La forme matricielle :

$$Y = X - 1_n \, g^{'}$$

Preuve :

$$X - 1_n g' = \begin{bmatrix} x_1^1 & \cdots & x_1^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^p \end{bmatrix} - \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \left(\overline{x^1}, . \ . \ . \ ., \overline{x^p} \right).$$

$$= \begin{bmatrix} x_1^1 - \overline{x^1} & \cdots & x_1^p - \overline{x^p} \\ \vdots & \vdots & \vdots \\ x_n^1 - \overline{x^1} & \cdots & x_n^p - \overline{x^p} \end{bmatrix} = \begin{bmatrix} y_1^1 & \cdots & y_1^p \\ \vdots & \vdots & \vdots \\ y_n^1 & \cdots & y_n^p \end{bmatrix} = Y.$$

2.4.2 Tableau reduit associé à X

La réduction des données nous permet de ramener toutes les variables à un même écart-type 1, d'aprés le tableau Y on construit un tableau standard noté par Z de terme général :

$$Z = \frac{y_{\text{ij}}}{s_j} = \frac{x_i^j - \overline{x^j}}{s_j}$$

Où :

$$Z = \begin{bmatrix} \frac{X_1^1 - \overline{X}^1}{S_1} & \cdots & \frac{X_1^p - \overline{X}^p}{S_p} \\ \vdots & \vdots & \vdots \\ \frac{X_i^1 - \overline{X}^1}{S_1} & \cdots & \frac{X_i^p - \overline{X}^p}{S_p} \\ \vdots & \vdots & \vdots \\ \frac{X_n^1 - \overline{X}^1}{S_1} & \cdots & \frac{X_n^p - \overline{X}^p}{S_p} \end{bmatrix}$$

Avec

$$s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$$

La forme matricielle :

$$Z = YD_{\frac{1}{s}}.$$

Avec : $D_{\frac{1}{s}}$ est la matrice diagonale des inverses des écarts-types :

$$D_{\frac{1}{s}} = \begin{bmatrix} \frac{1}{s_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_p} \end{bmatrix}$$

Preuve :

$$\text{YD}_{\frac{1}{s}} = \begin{bmatrix} y_1^1 & \cdots & y_1^p \\ \vdots & & \vdots \\ y_n^1 & \cdots & y_n^p \end{bmatrix} \begin{bmatrix} \frac{1}{s_1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \frac{1}{s_p} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{y_1^1}{s_1} & \cdots & 0 \\ \vdots & & \vdots \\ \frac{y_n^1}{s_1} & \cdots & \frac{y_n^p}{s_p} \end{bmatrix} = \begin{bmatrix} z_1^1 & \cdots & z_1^p \\ \vdots & & \vdots \\ z_n^1 & \cdots & z_n^p \end{bmatrix} = Z$$

2.5 Matrice de variance-covariance et matrice de corrélation

2.5.1 Matrice de variance-covariance

On appelle matrice de variance-covariance V associée à p variable aléatoires X^1,X^2,. . . . .,X^j,. . . , X^p mesurées sur un ensemble de n individus, la matrice à p lignes et p colonnes contenant sur sa diagonale principale les variances empiriques des p variables,et ailleurs, les covariances empiriques de ces variables deux à deux on a :(voir Caillez [1982]).

$$V = \begin{pmatrix} var(X^1) & cov(X^1, X^2) & . \; . \; . \; . \; . & cov(X^1, X^p) \\ cov(X^2, X^1) & var(X^2) & . \; . \; . \; . \; . & . \\ . & . & . \; . \; . \; . \; . & . \\ . & . & . \; . \; . \; . \; . & . \\ . & . & . \; . \; . \; . \; . & . \\ cov(X^p, X^1) & . & . \; . \; . \; . \; . & var(X^p) \end{pmatrix}$$

Avec :

$\text{var}\left(X^j\right) = \sum_{i=1}^{n} p_{\text{i}} \left(x_i^j - \overline{x}^j\right)^2$

Et :

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

$$\left(j, j' = 1, . \; . \; . \; ., p\right)$$

La forme matriciel :

$$V = X' DX - gg' = Y' \text{DY}$$

Preuve

On a

$$Y = X - 1_n g', \text{ alors}$$

$$V = \left(X - 1_n g'\right)' D \left(X - 1_n g'\right)$$

$$= X' DX - X' D 1_n g' - g 1_n' DX + g 1_n' D 1_n g'$$

$$= X' DX - gg' - gg' + gg'.$$

car :

$$1_n' D 1_n = \sum_{i=1}^{n} p_i = 1$$

$$= X' DX - gg'$$

Remarque 2.1 *Dans le cas ou les poids sont égaux,la forme matricielle devient :*

$$V = \frac{1}{n}Y^{'}Y = \frac{1}{n}X^{'}DX - gg'$$

2.5.2 Matrice de corrélation

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux est notée R :

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & . & . & r_{1p} \\ r_{21} & 1 & . & . & . & . \\ r_{31} & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ r_{p1} & . & . & . & . & 1 \end{pmatrix}$$

avec :

$$r_{jj'} = \frac{V_{jj'}}{S_j S_{j'}}$$

La forme matricielle :

$R = D_{\frac{1}{s}} V D_{\frac{1}{S}} = Z^{'}\text{DZ}.$

Avec :

Z est le tableau centré réduit .

D est la matrice des poids.

Preuve

On montre d'abord que $R = D_{\frac{1}{s}} V D_{\frac{1}{s}}$ On a :

$$D_{\frac{1}{s}} V D_{\frac{1}{s}} = \begin{bmatrix} \frac{1}{s_1} & \cdots & 0 \\ . & & . \\ . & & . \\ 0 & \cdots & \frac{1}{s_p} \end{bmatrix} \begin{bmatrix} V_1^2 & \cdots & V_{1p} \\ . & & . \\ . & & . \\ V_{p1} & \cdots & V_p^2 \end{bmatrix} \begin{bmatrix} \frac{1}{s_1} & \cdots & 0 \\ . & & . \\ . & & . \\ 0 & \cdots & \frac{1}{s_p} \end{bmatrix}$$

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

Ensuite, on montre que $Z^{'}DZ = R$. On a

$$\begin{aligned} Z^{'}DZ &= \left(YD_{\frac{1}{s}}\right)^{'} D \left(YD_{\frac{1}{s}}\right) \\ &= D_{\frac{1}{s}} Y^{'} \mathrm{DY} D_{\frac{1}{s}} \\ &= D_{\frac{1}{s}} V D_{\frac{1}{s}} \\ &= R \end{aligned}$$

2.6 Espace des individus (Nuage de points)

Chaque individu étant un point défini par p coordonées est considéré comme un vecteur d'un espace vectoriel défini dans R^p appelé l'espace des individus . L'ensemble des n individus est un nuage de points appelé nuage des individus, (voir Rousselet [2005]).

2.6.1 Le rôle de la métrique « Choix de la distance »

Pour faire une représentation géométrique il faut choisir une distance entre deux points de l'espace. La distance utilisé par l'ACP dans l'espace où sont représentées les individus, est la distance euclidienne classique. La distance entre deux individus e_i et e_j est définie par la forme quadratique suivante :

$$d_M^2\left(e_i, e_j\right) = \left(e_i - e_j\right)^{'} M \left(e_i - e_j\right) = \parallel e_i - e_j \parallel_M^2$$

Où : M est une matrice symétrique ($M^{'} = M$) de taille p définie positive,

Et : $\left(e_i - e_j\right)^{'}$ est la transposer de vecteur $\left(e_i - e_j\right)$

Ceci revient à munir l'espace des individus du produit scalaire :

$$\langle e_i, \quad e_j \rangle_M = e_i^{'} M e_j$$

Les métriques les plus utilisées sont les métriques $M = I_p$,correspondant au produit scalaire usuel, et la métrique $M = D_{\frac{1}{s^2}}$ (est métrique diagonal des inverses des variances) :

$$\mathrm{M} = \mathrm{D}_{\frac{1}{s^2}} = \begin{bmatrix} \frac{1}{s_1^2} & & & & 0 \\ & \frac{1}{s_2^2} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & & \frac{1}{s_p^2} \end{bmatrix}$$

Ce qui revient à diviser chaque caractère par son écart-type. Les éventuelles unités de mesure disparaissent alors car les nombres $\frac{x_i^j}{s_j}$ sont sans dimension et l'importance d'un caractère devient indépendant de sa dispersion.

Remarque 2.2 *On utise la métrique $D_{\frac{1}{s^2}}$ pour le tableau Y et la métrique I_p pour le tableau Z*

Preuve :

On a :

•Le $i^{\text{ème}}$ individu du tableau Y est ; $e_i^y = (y_i^1, \ldots, y_i^p)' \in \mathbb{R}^p$.

• Le $i^{\text{ème}}$ individu du tableau Z est ; $e_i^z = (z_i^1, \ldots, z_i^p)' \in \mathbb{R}^p$.

$$\left\langle e_i^y,\ e_i^y \right\rangle_{D_{\frac{1}{s^2}}} = \left(e_i^y\right)' D_{\frac{1}{s^2}} e_i^y$$

$$= \left(\frac{y_i^1}{s_1}\right)^2 + \ldots\ldots + \left(\frac{y_i^p}{s_p}\right)^2 = \sum_{j=1}^{p} \left(\frac{y_i^j}{s_j}\right)^2$$

$$= \sum_{j=1}^{p} \left(Z_i^j\right)^2 = \sum_{j=1}^{p} \left(\frac{Z_i^j}{1}\right)^2 = (e_i^z)' I_p e_i^z = \langle e_i^z, e_i^z \rangle_{I_p}.$$

2.6.2 L'inertie

On appelle inertie totale du nuage de points la moyenne des carrés des distances des n points au centre de gravité g. Elle est exprimée comme ceci :

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

On peut aussi l'ecrire comme :

$$I_g = \sum_{i=1}^{n} p_i \left(e_i - g\right)' M \left(e_i - g\right) = \sum_{i=1}^{n} p_i \| e_i - g \|_M^2 = \sum_{i=1}^{n} p_i \left\langle e_i - g,\ e_i - g \right\rangle_M$$

Remarque 2.3 *L'inertie en un point a quelconque définie par :*

$$I_a = \sum_{i=1}^{n} p_i \left(e_i - a\right)' M \left(e_i - a\right) = \sum_{i=1}^{n} p_i d_M^2 \left(e_i, a\right) = \sum_{i=1}^{n} p_i \| e_i - a \|_M^2$$

Si $g = 0$, on a : $I_g = \sum_{i=1}^{n} p_i e_i{}' M e_i$

Formule de Huyghens :

$I_a = I_g + \left(g - a\right)' M \left(g - a\right) = I_g + \| g - a \|^2.$

Par ailleurs, on peut montrer que l'inertie totale est égale à la moitié de la moyenne des carrés de toutes les distances entre les n individus. Mais l'égalité la plus utilisée est la suivante :

$$I_{g=trace(MV)=trace\ (VM)}.$$

Proposition 2.1 *-Si $M = I_p$, l'inertie est égale à la somme des variances des p variables :*

$$I_g = \sum_{j=1}^{p} S_j^2.$$

- *Si :M =* $D_{\frac{1}{S^2}}$, l'inertie est égale au nombre de variable :
$\mathrm{I}_g = P$

2.7 Espaces des variables

Chaque variable X^j est considérée comme vecteur d'un espace E à n dimensions, appelé espace des variables, (voir Volle [1980]).

2.7.1 Liaison entre deux variables

Le coefficient r_{jk} de corrélation mesure la liaison entre deux variables X_j et X_k,qui prend ses valeurs dans $[-1, 1]$

$r\left(X_j, X_k\right) = \frac{cov(X_j,X_k)}{\sqrt{var(X_j)var(X_k)}} \ =\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{X_{ij}}-\overline{X_j}}{s_j}\right)\left(\frac{\mathrm{X_{ik}}-\overline{X_k}}{s_k}\right)$, pour $j, k = \overline{1, p}.$

Avec

$\mathrm{cov}\left(X_j, X_k\right) = \left(\mathrm{X_{ij}} - \overline{X_j}\right)\ \left(\mathrm{X_{ik}} - \overline{X_k}\right), \overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X_{ij}}$
et

$s_j{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{X_{ij}} - \overline{X_j}\right)^2.$

2.7.2 La métrique des poids

La métrique utilisée dans cet espace sera la métrique des poids D(matrice carrée de taille n) définie précédemment.
En effet, si les variables sont centrées, on a les propriétés suivantes :

•$\langle X^j, X^k\rangle_D = \left(X^j\right)' DX^k = \sum_{i=1}^{n} p_i X_i^j X_i^k = s_{\mathrm{jk}}$: c'est la covariance des deux variables.

•$\| \ X^j \ \|_D^2 = s_j^2$:la « longueur »d'une variable est son écart-type.

•Si θ_{jk} est l'angle entre deux variables centrées, on a :

$$\cos\left(\theta_{\mathrm{jk}}\right) = \frac{\langle X^j, X^k\rangle_D}{\left\|X^j\right\|_D \left\|X^k\right\|_D} = \frac{s_{\mathrm{jk}}}{s_j s_k}$$

C'est le coefficient de corrélation linéaire.

•Si dans l'espace des individus on s'intéresse aux distance entre points, dans l'espace des variables on s'interessera plutôt aux angles en raison de la propriété précédente.

2.8 L'analyse

2.8.1 Projection des individus sur un sous-espace

Le principe de la méthode est d'obtenir une représentation approchée du nuage des n individus sur le sous-espace de dimension faible. Ceci s'effectue par projection ainsi que l'illustre la figure 2.1,(voirSaporta [1980])

-Nuage projeté

Le critére du choix de l'espace de projection s'effectue tel que la moyenne des carrées des distances entre les projections et leur centre de gravité soit la plus grande possible. Ce qui implique qu'il faut que l'inertie du nuage projeté sur ce sous espace soit maximale on note F_K le sous-espace de projection.

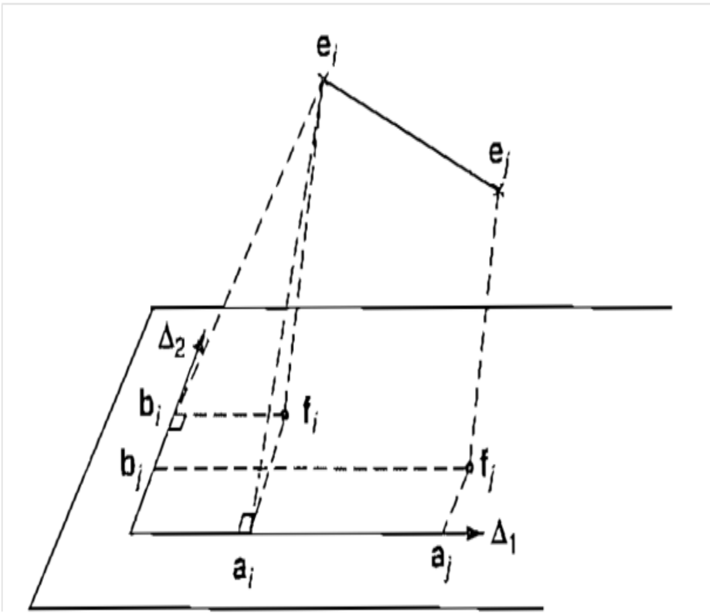

FIGURE 2.1 – *Projection du nuage des individus dans un sous -espace de dimension* 2

f_i et f_j sont les projections respectives de e_i et e_j sur F_2 ; a_i et a_j sont les coordonnées respectives de f_i et f_j sur le premier axe ;$\triangle_1$, b_i et b_j sont les coordonées respéctives de f_i et f_j sur le deuxième axe $\triangle_2$.

Soit P une matrice (opérateur) de projection M-orthogonale sur F_K, elle vérifie les deux conditions suivantes :

- $\mathrm{P}^2 = \mathrm{P}$.
- MP=P′M.

Définition 2.1 *Soit f_i la projection d'un individu e_i telle que $f_i = Pe_i$ d'où $f_i^{'} = e_i{}'P^{'}$ c'est la $i^{ème}$ ligne du tableau $XP^{'}$.*

On écrit : $X_{proj} = XP^{'}$.

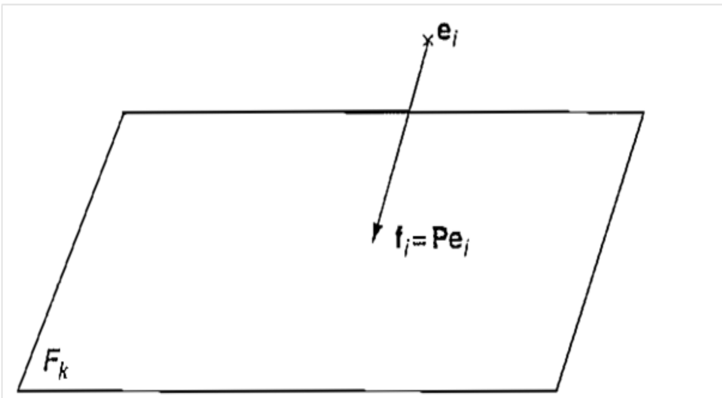

FIGURE 2.2 – *Projection d'un individu sur un espace f_K*

Proposition 2.2 1.*La matrice de variance du tableau $XP^{'}$est pour des variables centrées :*

$(XP^{'})^{'} D(XP^{'})$=$PX^{'}DXP^{'} = PVP^{'}$

2.*L'inertie du nuage projeté vaut :*

trace $\left(PVP^{'}M\right) = trace\,(VMP)$ *ou P est l'operateur de projection M-orthogonale sur F_k.*

3. $I_{G \bigoplus G'} = I_G + I_{G'}$ *ou G et $G^{'}$ sont deux sous-espaces orthogonaux de F.*

Théorème 2.1 *Soit F_K un sous-espace portant l'inertie maximale; alors, le sous-espace de dimension $K+1$ portant l'inertie maximale est la somme directe de F_k et du sous-espace de dimension 1 M-orthogonal à F_k portant l'inertie maximale.Les solutions sont donc «emboîtées».*

Il suffit alors de déterminer le sous espace -vectoriel de dimension 1 de R^p qui maximise l'inertie du nuage projeté sur cette droite.

En notant a un vecteur directeur de cette droite, l'operateur de projection M-orthogonale sur cette droite est :

$P_a = a(a' \ Ma)^{-1}a^{'}M = \frac{aa^{'}M}{a^{'}\mathrm{Ma}}$,car $(a^{'}Ma) \in R$

Il est relativement aisé d'aboutir à la relation :$VMa = \lambda a$

Or l'inertie du nuage projeté sur cette droite vaut :

$\text{trace}\,(\mathrm{VM}P_a) = \frac{a^{'}\mathrm{MVMa}}{a^{'}\mathrm{Ma}} = \lambda$

On vertu du résultat définit en propriétée 2. Et de l'écriture de P_a. Donc en choisissant pour a le vecteur propre associé à la plus grande valeur propre de VM, on obtient le sous-espace vectoriel de dimension 1 sur lequel l'inértie du nuage projeté est maximale.

Cette propriété se généralise à l'ordre K :

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

Remarque 2.4 *La matrice VM étant M-symétrique, elle possède des vecteurs propres M-orthogonaux deux à deux.*

2.9 Éléments principaux

L'ACP repose essentiellement sur trois éléments qui sont : (voir Lebart [1995]).

2.9.1 Axes principaux

Ce sont les p vecteurs propres $(a_i. \ . \ . \ .a_p)$de la matrice VM associée à la valeur propre λ_j , M-normé à 1 :

$$\begin{cases} \mathrm{VM}a_j = \lambda_j a_j \\ \\ \| a_j \|_M^2 = 1 \end{cases}$$

-Propriétés des axes principaux

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

.Les axes principaux a_j sont M orthonormé.

Preuve

Soit (a_j, a_k) deux axes principaux tel que

$$\langle a_j, a_k \rangle_{V^{-1}} = a_j^{'} V^{-1} a_k$$

$$= \frac{1}{\lambda_j} \left(\mathrm{VM}a_j\right)^{'} V^{-1} a_k$$

$$= \frac{1}{\lambda_j} a_j^{'} \mathrm{MV} V^{-1} a_k$$

$$= \frac{1}{\lambda_j} a_j^{'} M a_k$$

$$= \frac{1}{\lambda_j} \langle a_j, a_k \rangle_M$$

$$= \begin{cases} \frac{1}{\lambda_j} & si \quad j = k \\ 0 & si \quad non \end{cases}$$

2.9.2 Facteur principaux

Soit a_j un axe principal, le facteur principal noté u_j est un vecteur propre de la matrice MV associé à la valeurs propre λ_j ; M^{-1}-normé à 1 :

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

Où $u_j = Ma_j \in \mathbb{R}^p$.

-Propriétés des facteurs principaux

.u_j sont V-orthogonaux.

. u_j sont M^{-1}-orthonormé.

.u_j sont aussi les vecteurs propres de la matrice MV.

Preuve

$$1\,. \quad \langle u_j, u_k \rangle_V = u_j^{'} V u_k$$

$$= a_j^{'} \mathrm{MVM} a_k$$

$$= a_j^{'} M \lambda_k a_k$$

$$= \lambda_k a_j^{'} M a_k$$

$$= \lambda_k \langle a_j, a_k \rangle_M$$

$$= \begin{cases} \lambda_k & si \quad j = k \\ 0 & si \quad non \end{cases}$$

$$2. \quad \langle u_j, u_k \rangle_{M^{-1}} = u_j^{'} M^{-1} u_k$$

$$= a_j^{'} M M^{-1} M a_k$$

$$= a_j^{'} M a_k$$

$$= \langle a_k, a_k \rangle_M$$

$$= \begin{cases} 1 & si \quad j = k \\ 0 & si \quad non \end{cases}$$

3. Comme a_{j} est un vecteur propre de la matrice VM on a :

$$\mathrm{VM} a_j = \lambda_j a_j$$

$$\mathrm{MVM} a_j = \lambda_j \mathrm{Ma}_j$$

$$\mathrm{MV} u_j = \lambda_j u_j$$

2.9.3 Composantes principales

Chaque axe a_j est associé à une variable appelée composante principale. Ce sont de nouvelles variables $c_j = (c_{1j}, c_{2j}, \ldots, c_{nj}) \in \mathbb{R}^n$ définies en fonction des facteurs principaux :

$$c_j = XMa_j = Xu_j.$$

Si on travaille avec le tableau centré réduit devient :

$$c_j = Zu_j$$

Chaque c_j contient des coordonnés $(c_1, c_2, \ldots . c_n)$ qui sont les mesures algébriques des projections des individus e_i sur ces axes.

-Propriétés des composantes principales

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

$$\text{cov}\left(c_j, c_k\right) = 0$$

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

$$\text{var}\left(c_j\right) = \lambda_j$$

Les composantes principales sont les vecteurs propres de la matrice $XMX'D$:

$$\text{XM}X'Dc_j = \lambda_j c_j$$

Preuve

1 . $\text{cov}\left(c_j, c_k\right) = c_j' Dc_k - g_{\text{cj}} g_{\text{ck}}'$

$= u_j' X' \text{DX} u_k - c_j' D1_n 1_n' Dc_k$

$= u_j' X' \text{DX} u_k - u_j' X' D1_n 1_n' \text{DX} u_k$

$= u_j' X' \text{DX} u_k - u_j' X' D1_n 1_n' \text{DX} u_k$

$= u_j' (X' DX - gg') u_k$

$=\langle u_j, u_k\rangle_V$

$= 0$

2 .Même démonstration que la précédente

$$\text{var}\left(c_j\right) = c_j^{'}Dc_j - g_{\text{cj}}g_{\text{cj}}^{'}$$

$$= \langle u_j, u_j\rangle_V$$

$$= \parallel u_j \parallel_V^2$$

$$= \lambda_j.$$

3 .Dans le cas où g est centré.

$$\text{XMX}^{'}Dc_j = XMX^{'}\text{DX}u_j$$

$$= XMVu_j$$

$$= X\lambda_j u_j$$

$$= \lambda_j X u_j$$

$$= \lambda_j c_j$$

Remarque 2.5 1 *.Les composantes principales c_j sont des combinaisons linéaires des variables centrées et réduites. On a :*

$$c_j = \sum_{k=1}^{p} u_{kj} x_k.$$

3 *.La première composante principale doit être de variance maximale.*

2.9.4 Résumé des propriétés des éléments principaux

Éléments principaux	Définition	Propriété	Relation
a_i : *axes principaux* $\in R^p$	$VMa_i=\lambda_i a_i$	M-orthonormés	
u_i : facteurs principaux $\in (R^p)^*$	$MVu_i=\lambda_i u_i$	M^{-1}-orthonormés	$u_i = Ma_i$
c^i : composantes principales $\in R^n$	$XMX'Dc^i = \lambda_i c^i$	D-orthogonales	$c^i = Xu_i$

TABLE 2.1 – *propriétés des éléments principaux issus d'une ACP*

2.10 L'analyse en Composantes Principales Normée

Travailler avec la métrique $M = D_{\frac{1}{s^2}}$ sur le tableau centré X revient à travailler avec la métrique $M = I_p$ sur le tableau centré réduit Z : c'est l'ACP sur données centrées réduites, communément appelée ACPN,(voir Duby [2006])

Dans ce cas, on a $MV = R$: c'est la matrice des corrélations entre variables. La diagonalisation de R nous fournira facteurs principaux et composantes principales. Dans le cas d'une ACPN, on a une propriété supplémentaire :

Propriété 2.1 *La composante principale c^1 associée à la plus grande valeur propre est la variable la plus liée aux variables $x^1, \ldots\ldots., x^p$ au sens de la somme des carrés des corrélations :*

$\sum_{i=1}^{p} r^2(c, x^j)$ *est maximal.*

2.11 Interprétation et qualité des résultats d'une ACP

2.11.1 Interprétation « interne »

-Corrélations entre composantes principales et variables initiales

L'interprétation des axes se fait par l'intermédiaire de l'étude des corrélations entre la composante principale définissant cet axe et les variables du tableau de données initial. Lorsque l'on effectue une ACPN, les coefficients de corrélation linéaire $r(c^k, x^j)$ s'expriment simplement en fonction des facteurs principaux , (voir Chessel [2008]).

$$r\left(c^k, x^j\right) = \sqrt{\lambda_k} u_j^k$$

Ou :

c^k :$k^{\text{ème}}$ composante principale ;

x^j : $j^{\text{ème}}$ variable initiale ;

λ_k : valeur propre associée à c^k ;

u_j^k : $j^{\text{ème}}$ composante du facteur principal u_k. ;

Ainsi, pour un couple de composantes principales c^1 *et* c^2, on représente ces corrélations linéaires sur une figure appelée cercle de corrélations où chaque variable x^j est repérée par un point d'abscisse $r\left(c^1, x^j\right)$ et d'ordonnée r $\left(c^2, x^j\right)$ (voir figure 2.3)

Remarque 2.6 *Lorsque l'on ne travaille pas sur des données centrées réduites, il vaut mieux éviter d'interpréter des proximités entre points variables si ceux-ci ne sont pas proches de la circonférence. Par contre, dans le cas de l'ACPN, le cercle des corrélations est la projection exacte de l'ensemble des variables centrées réduites sur le sous-espace engendré par* c^1 *et* c^2.

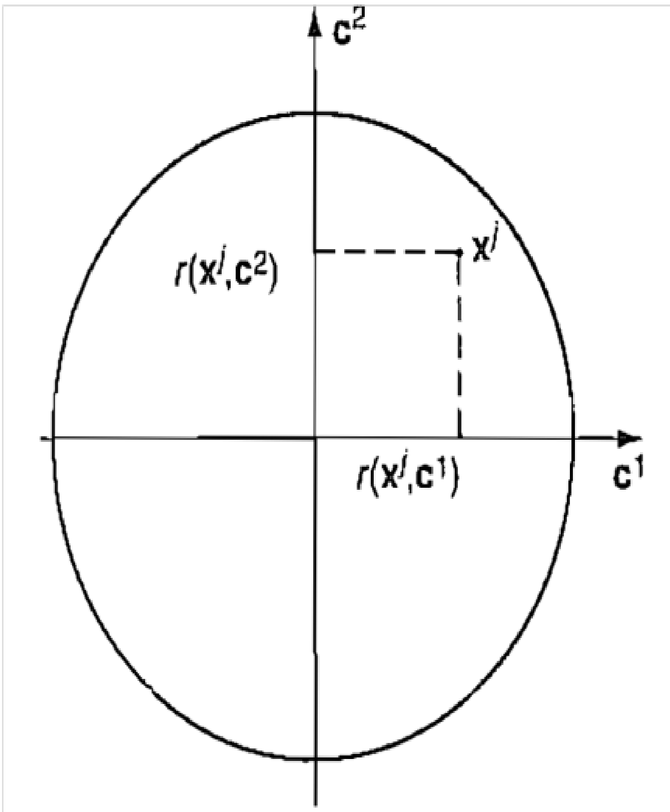

FIGURE 2.3 – *Cercle des corrélations*

-Interprétation de la position des individus

Dire que la première composante principale c^1 {est très corrélée avec une variable x^j signifie que les individus ayant une forte coordonnée positive sur l'axe 1 sont caractérisés par une valeur de x^j nettement supérieure à la moyenne puisque l'origine des axes principaux représente le centre de gravité du nuage. Il est aussi utile de calculer pour chaque axe la contribution apportée par les divers individus à cet axe ; la contribution de l'individu i à la composante c^k est défini par :

$$\frac{p_i(c^{k,i})^2}{\lambda_k}$$

Ou $c^{k,i}$ est la $i^{\text{ème}}$ coordonnée de la composante principale c^k.

Enfin, il est nécessaire de s'intéresser à la qualité de la représentation des individus sur les plans principaux : pour un individu e_i, elle se mesure gr$\widehat{a}$ce au cosinus de l'angle entre le plan principal et le vecteur e_i. Ainsi, si ce cosinus est grand, e_i est proche de plan, donc bien représenté.

2.11.2 Interprétation « externe »

-Variables et individus supplémentaires

Il est possible d'intégrer à une analyse des variables et des individus supplémentaires : ces éléments n'interviennent pas dans la détermination des composantes principales et des facteurs principaux, mais ils peuvent être placés a posteriori dans les représentations graphiques.
Donnons un exemple d'utilisation de variables supplémentaires : de nouvelles formes de délinquance ou de criminalité viennent d'apparaîte dans la société français ; compte tenu de leurs caractéristiques, on souhaite déterminer comment se situent les nouveaux comportements dans l'analyse comme « variables supplémentaires » : celles-ci ne participent pas activement à l'analyse mais il est possible de les positionner sur les divers plans de l'ACP et par conséquent, de juger de leurs ressemblances ou non avec les variables actives. On peut également utiliser les variables et les individus supplémentaires dans le cas suivant : un individu ou une variable a une très forte contribution sur un ou plusieurs axes principaux et un poids faible. Afin d'éviter cet effet « brouillage de l'information », cet individus ou cette variable peut être analysé en tant qu'individu ou variable supplémentaire.

Application sur le logiciel R 3

3.1 Définition du langage R

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

R a été crée par Ross Ihaka et Robert Gentleman en 1996 du département de statistique de l'Université d'Auckland, en Nouvelle Zélande, et est maintenant développé par la *R* développement Core Team. Il est conçu pour pouvoir étre utilisé avec les système d'exploitation Unix, Linux, Windows et MacOS.

Le *R* est une application n'offrant qu'une invite de commande il est basé sur la notion de vecteur, ce qui simplifie les calculs mathématique et réduit considérablement le recours aux structures itératives (boucles for, ...ec). Programmes courts, en général quelques lignes de code seulement. Temps de développement trés court.

3.2 Exemple d'application

Pour 15 villes de France, on dispose des moyennes des températures mensuelles calculées sur 30 ans (entre 1931 et 1960). Elles sont rassemblées dans le tableau (3.1), qui croise ces 15 villes en lignes (individus) et les 12 mois de l'année en colonnes (variables), (voirFrançois [2016]).

Différents packages et fonctions utilisés sont disponibles dans les bibliothèques standard de *R* :

Table des données

	Jan	Fev	Mar	Avr	Mai	Jui	Juil	Aou	Sep	Oct	Nov	Dec
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3
Lille	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1
Marseille	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5
Nantes	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4

TABLE 3.1 – *Températures mensuelles de 15 villes de France.*

Packages :

ade4, FactoMineR.

Fonctions :

read.table, colMeans, cov, cor, scale, dudi.pca, sum, barplot, abline, symbols, s.corcircle.

Programmation :

library(ade4) # Il Contient des fonctions danalyse des données.

library(FactoMineR) # Analyse exploratoire multidimensionnelle des données.

setwd("C :/ACP")

X=read.table("data.txt",h=T) # Importer le tableau

g =colMeans(X) # Centre de gravité g

round(g, 3)

Jan	Fev	Mar	Avr	Mai	Jui	Juil	Aou	Sep	Oct	Nov	Dec
3.857	4.707	8.086	10.850	14.336	17.729	19.757	19.464	16.871	12.214	7.843	4.750

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

round(V, 3)

$$V = \begin{bmatrix} 4.029 & 3.651 & 2.564 & 1.737 & 1.095 & 1.220 & ... \\ 3.651 & 3.491 & 2.660 & 2.012 & 1.537 & 1.757 & ... \\ 2.564 & 2.660 & 2.338 & 1.989 & 1.764 & 2.075 & ... \\ 1.737 & 2.012 & 1.989 & 2.002 & 2.028 & 2.391 & ... \\ 1.095 & 1.537 & 1.764 & 2.028 & 2.264 & 2.682 & ... \\ 1.220 & 1.757 & 2.075 & 2.391 & 2.682 & 3.217 & ... \\ \vdots & & & & & & \\ . & . & .. & . & . & & \end{bmatrix}$$

R = cor(X) # Matrice de corrélation R :
round(R, 3)

$$R = \begin{bmatrix} 1.000 & 0.973 & 0.835 & 0.611 & 0.363 & 0.339 & ... \\ 0.973 & 1.000 & 0.931 & 0.761 & 0.547 & 0.524 & ... \\ 0.835 & 0.931 & 1.000 & 0.920 & 0.767 & 0.757 & ... \\ 0.611 & 0.761 & 0.920 & 1.000 & 0.953 & 0.942 & ... \\ 0.325 & 0.515 & 0.749 & 0.954 & 1.000 & 0.994 & ... \\ 0.304 & 0.495 & 0.744 & 0.947 & 0.994 & 1.000 & ... \\ \vdots & & & & & & \\ . & . & .. & . & . & & \end{bmatrix}$$

Z = scale(X) # Tableau standard Z :
round(Z, 3)

$$Z = \begin{bmatrix} 0.810 & 0.946 & 1.352 & 1.286 & 0.908 & 0.818 & ... \\ 1.059 & 0.517 & -0.283 & -1.258 & -1.883 & -1.914 & \\ -0.684 & -0.607 & -0.480 & -0.481 & -0.421 & -0.297 & \\ -1.232 & -0.874 & -0.349 & -0.269 & 0.044 & -0.019 & \\ -0.784 & -1.035 & -1.461 & -1.470 & -1.351 & -1.413 & \\ -0.933 & -0.821 & -0.349 & -0.057 & 0.310 & 0.372 & \\ \vdots & & & & & & \\ . & . & .. & . & . & & \end{bmatrix}$$

acp=dudi.pca(X,center=T,scale=T,nf=2,scannf=F)

vp=acp$eig # Valeurs propres λ

round(vp,3)

9.582	2.276	0.070	0.040	0.014	0.008	0.006	0.002	0.001	0.000	0.000	0.000

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

round(pvp, 3)

79.848	18.970	0.583	0.33)	0.117	0.067	0.050	0.015	0.012	0.004	0.002	0.000

barplot(pvp,ylab="% d'inertie",names.arg=(round(pvp, 3)), col =1) Histogram des vps.

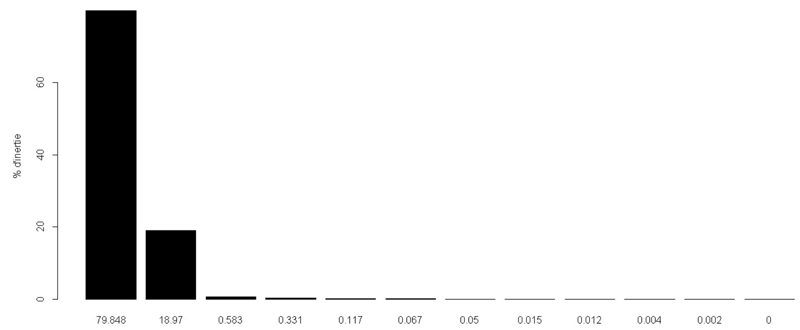

Figure 3.1 – *Eboulis des valeurs propres en %*

Commentaire : L'inertie expliquée par la 1^{re} dimension est de $79,848\%$, la 2ème dimension est de $18,970\%$...ect. En assemblant ces deux premiers pourcentage on obtient environ $98,818$ % d'inertie totale égale à $12 = I_g$ i.e une bonne qualité sur ce plan.

c1=acp$co[,1] # 1ère composante principale c1 :

round(c1, 3)

c2=acp$co[,2] # 2ème composante principale c2 :

round(c2, 3)

contrib=inertia.dudi(acp,row.inertia=T,col.inertia=T) # Contribution $CTR_i(x)_j$:

contribc=contrib$col.abs

round(contribc,3)

plot(c1,c2,type="n",ylab="comp1 :79.848%",xlab="comp2 :18.970%",main="Mois",

xlim=c(−1, 1),ylim=c(−1, 1),col=1)

abline(h=0,v=0)

text(c1,c2,row.names(acp$co),col=1) #Tracer le graphe des deux composantes c_j et $c_j{}'$.

symbols(0, 0,circles=1,ylab="comp1 : 79.848%",xlab="comp2 :18.970%",inches=F,add=T)

for(i in 1 : 12)

arrows(0, 0,c1[i],c2[i],angle=20,length=0.15)

s.corcircle(acp$co) # Cercle des correlations.

	Coordonnées		Contribution	
Mois	c1	c2	c1	c2
Jan	-0.761	0.644	6.048	18.238
Fev	-0.880	0.469	8.090	9.666
Mar	-0.969	0.156	9.795	1.069
Avr	-0.969	-0.204	9.806	1.822
Mai	-0.873	-0.475	7.950	9.899
Jui	-0.864	-0.499	7.783	10.953
Jui	-0.842	-0.531	7.391	12.406
Aou	-0.899	-0.430	8.427	8.120
Sep	-0.974	-0.208	9.901	1.902
Oct	-0.980	0.170	10.026	1.276
Nov	-0.904	0.414	8.524	7.527
DEC	-0.774	0.624	6.258	17.121

Table 3.2 – *Composantes et Contribution des variables.*

Commentaire : On observe que tout les coordonnées sur le $1^{ére}$ axe proche de 1 en valeur absolue i.e la valeur de corrélation entre ces variables et cet axe est fortement et positivement donc les variables ce bien représent sur cet axe. Sur le même tableau (3.2), on observe que la valeur de corrélation entre ces variables et ce $2^{ème}$ axe est faible.

On peut dire que les variables sont bien représentes sur le $1^{ére}$ plan principal [Voir la représentation des variables]

co1=acp$li[,1]   #$1^{ère}$ composante principale co1 :

round(co1, 3)

co2=acp$li[,2]   #$2^{ème}$ composante principale co2 :

round(co2, 3)

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

contribl=contrib$row.abs   # Contribution $CTR_i(e_j)$

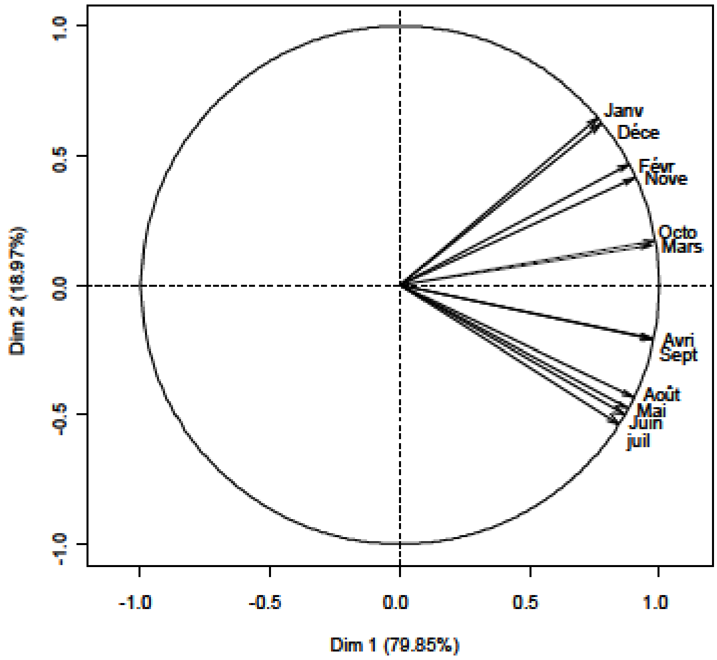

FIGURE 3.2 – *Représentation des variables*

round(contribl,3)

plot(co1,co2,ylab="axe1 :79.848%",xlab="axe2 :18.970%",xlim=c($-7, 7$),ylim=c($-4.5, 4.5$),col=1)

abline(h=0,v=0)

text(co1,co2,row.names(acp$li),col=1,cex=1) #Tracer le graphe des villes celons les 2 axes.

range(co1) # Borne du 1$^{\text{ère}}$ axe.

-6.007	4.217

range(co2) #Borne du 2$^{\text{ème}}$ axe.

-2.172	4.093

	Coordonnées		Contribution	
Villes	co1	co2	co1	co2
Bordeaux	-3.121	0.109	6.776	0.035
Brest	2.268	4.093	3.579	49.069
Clermont	1.726	-0.593	2.073	1.028
Grenoble	1.529	-1.688	1.627	8.344
Lille	4.217	0.595	12.372	1.037
Lyon	0.835	-1.788	0.485	9.365
Marseille	-4.833	-0.829	16.250	2.012
Montpellier	-4.147	-0.435	11.967	0.555
Nantes	0.281	1.115	0.055	3.638
Nice	-6.007	0.789	25.106	1.825
Paris	1.242	-0.156	1.073	0.072
Rennes	1.439	1.671	1.440	8.178
Strasbourg	4.106	-2.172	11.728	13.819
Toulouse	-1.736	-0.136	2.097	0.054
Vichy	2.201	-0.575	3.372	0.969

TABLE 3.3 – *Composantes et Contribution des individus.*

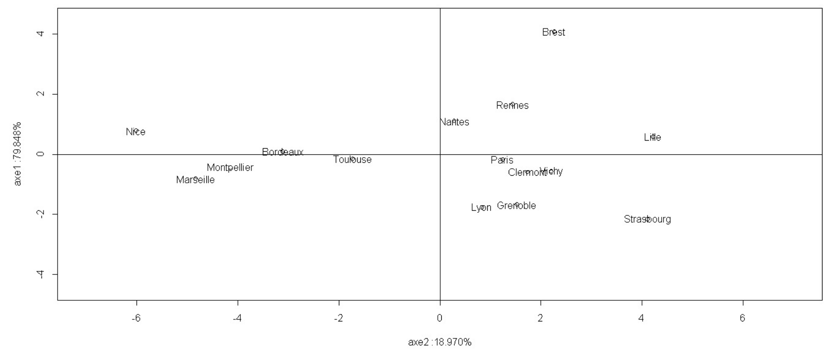

FIGURE 3.3 – *Représentation de nuage des individus.*

Commentaire :

On compare les coordonnées de la $1^{\text{ére}}$ composante principale à la racine carrée de la $1^{\text{ére}}$ vp, i.e : $\sqrt{\lambda_1} = \sqrt{9.582} = 3.095$, on prend seulement les individus qui ont des coordonnées supérieurs ou égales à $\sqrt{\lambda_1}$, puis on regroupe d'après ces signes. Le tableau suivant contient six villes devisées sur 2 groupes qui sont bien représentées sur le premier axe :

-	+
Bordeaux	Lille
Marseille	Strasbourg
Montpellier	
nice	

On a Lille, Strasbourg, Bordeaux, Marseille, Montpellier, Nice sont bien représentées sur l'axe 1.

De la même manière on compare les coordonnées des individus par la $2^{\text{ème}}$ composante principale à la racine carrée de la $2^{\text{éme}}$ vp i.e :

$\sqrt{\lambda_2} = \sqrt{2.276} = 1,509$, ou on prend seulement les individus qui ont des coordonnées supérieur ou égale à $\sqrt{\lambda_2}$ en valeur absolue puis on va regrouper d'après ces signes :

-	+
Grenoble	Brest
Lyon	Rennes
Strasbourg	

On a Brest, Rennes, Grenoble, Lyon, Strasbourg sont bien représentées sur l'axe 2,

Remarque 3.1 *Clermont, Nantes, Paris, Toulouse et Vichy sont bien représentées sur le plan principal.*

4 Annexe

4.1 Exécution du code R

Avant de démarrer :

1. Aller dans le disque local C :
2. Créer un dossier nommé ACP
3. Aller dans le dossier C :
4. Mettre à l'intérieur le fichier “data.txt”

Remarque 4.1 *S'assurer que les packages* ***ade4,FactoMineR*** *sont bien installés*

Centre de gravité g

g =colMeans(X)

round(g ; 3)

Jan	Fev	Mar	Avr	Mai	Jui	Juil	Aou	Sep	Oct	Nov	Dec
3.857	4.707	8.086	10.850	14.336	17.729	19.757	19.464	16.871	12.214	7.843	4.750

Matrice de covariance V

S = cov(X)

round(V ; 3)

	Janv	Fevr	Mars	Avri	Mai	Juin	Juil	Aôut	Sept	Octo	Nove	Dec
Janv	4.029	3.651	2.564	1.737	1.095	1.220	1.269	1.636	2.240	3.124	3.443	3.907
Fevr	3.651	3.491	2.660	2.012	1.537	1.757	1.950	2.210	2.623	3.215	3.323	3.550
Mar	2.564	2.660	2.338	1.989	1.764	2.075	2.349	2.455	2.573	2.710	2.558	2.496
Avri	1.737	2.012	1.989	2.002	2.028	2.391	2.748	2.701	2.558	2.344	1.996	1.716
Mai	1.095	1.537	1.764	2.028	2.264	2.682	3.143	2.968	2.620	2.121	1.604	1.122
Juin	1.220	1.757	2.075	2.391	2.682	3.217	3.786	3.568	3.117	2.484	1.852	1.269
Juil	1.269	1.950	2.349	2.748	3.143	3.786	4.532	4.243	3.670	2.878	2.101	1.350
Aôut	1.636	2.210	2.455	2.701	2.968	3.568	4.243	4.048	3.607	2.990	2.310	1.695
Sept	2.240	2.623	2.573	2.558	2.620	3.117	3.670	3.607	3.416	3.130	2.654	2.253
Octo	3.124	3.215	2.710	2.344	2.121	2.484	2.878	2.990	3.130	3.349	3.183	3.103
Nov	3.443	3.323	2.558	1.996	1.604	1.852	2.101	2.310	2.654	3.183	3.254	3.394
Dec	3.907	3.550	2.496	1.716	1.122	1.269	1.350	1.695	2.253	3.103	3.394	3.836

Figure 4.1 –

Matrice de corrélation R

R =cor(X)

round(R ; 3)

	Janv	Fevr	Mars	Avri	Mai	Juin	Juil	Août	Sept	Octo	Nove	Dece
Janv	1.000	0.973	0.835	0.611	0.363	0.339	0.297	0.405	0.604	0.850	0.951	0.994
Fevr	0.973	1.000	0.931	0.761	0.547	0.524	0.490	0.588	0.760	0.940	0.986	0.970
Mars	0.835	0.931	1.000	0.920	0.767	0.757	0.722	0.798	0.911	0.968	0.927	0.834
Avr	0.611	0.761	0.920	1.000	0.953	0.942	0.912	0.949	0.978	0.905	0.782	0.619
Mai	0.363	0.547	0.767	0.953	1.000	0.994	0.981	0.980	0.942	0.770	0.591	0.381
Juin	0.339	0.524	0.757	0.942	0.994	1.000	0.992	0.989	0.940	0.757	0.572	0.361
Juil	0.297	0.490	0.722	0.912	0.981	0.992	1.000	0.991	0.933	0.739	0.547	0.324
Aôut	0.405	0.588	0.798	0.949	0.980	0.989	0.991	1.000	0.970	0.812	0.637	0.430
Sept	0.604	0.760	0.911	0.978	0.942	0.940	0.933	0.970	1.000	0.926	0.796	0.622
Octo	0.850	0.940	0.968	0.905	0.770	0.757	0.739	0.812	0.926	1.000	0.964	0.866
Nove	0.951	0.986	0.927	0.782	0.591	0.572	0.547	0.637	0.796	0.964	1.000	0.961
Dec	0.994	0.970	0.834	0.619	0.381	0.361	0.324	0.430	0.622	0.866	0.961	1.000

FIGURE 4.2 –

Matrice centré réduite Z

Z =scale(X)

round(Z ; 3)

	Janv	Fevr	Mars	Avri	Mai	Juin	Juil	Aôut	Sept	Octo	Nove	Dece
Bordeaux	0.810	0.946	1.352	1.286	0.908	0.818	0.501	0.712	0.873	0.809	0.650	0.691
Brest	1.059	0.517	-0.283	-1.258	-1.883	-1.914	-1.988	-1.773	-1.237	-0.175	0.595	1.100
Clermon	-0.684	-0.607	-0.480	-0.481	-0.421	-0.297	-0.204	-0.232	-0.426	-0.612	-0.736	0.637
Grenob	-1.232	-0.874	-0.349	-0.269	0.044	-0.019	0.125	-0.033	-0.155	-0.503	-0.791	-1.300
Lille	-0.784	-1.035	-1.461	-1.470	-1.351	-1.413	-1.284	-1.226	-1.237	-1.049	-1.013	-0.688
Lyon	-0.933	-0.821	-0.349	-0.057	0.310	0.372	0.407	0.265	-0.047	-0.503	-0.680	-0.892
Marseille	0.761	0.946	1.155	1.428	1.573	1.654	1.628	1.607	1.576	1.464	1.260	1.048
Montpelli	0.810	0.999	1.090	1.286	1.174	1.264	1.347	1.359	1.252	1.246	1.149	0.844
Nantes	0.511	0.250	0.109	-0.127	-0.354	-0.353	-0.485	-0.480	-0.317	-0.066	0.152	0.334
Nice	1.757	1.962	1.679	1.640	1.506	1.264	1.347	1.458	1.793	2.011	1.981	1.712
Paris	-0.286	-0.392	-0.414	-0.198	-0.089	-0.186	-0.344	-0.431	-0.534	-0.503	-0.458	-0.279
Rennes	0.412	0.250	-0.218	-0.622	-0.886	-0.911	-0.908	-0.878	-0.696	-0.393	-0.070	0.283
Strasb	-1.780	-1.784	-1.722	-0.834	-0.288	-0.353	-0.391	-0.630	-1.021	-1.541	-1.678	-1.811
Toulouse	0.362	0.410	0.632	0.438	0.310	0.483	0.501	0.663	0.711	0.536	0.373	0.334
Vichy	-0.784	-0.767	-0.741	-0.763	-0.554	-0.409	-0.251	-0.381	-0.534	-0.721	-0.736	-0.739

FIGURE 4.3 –

Représentation des individus

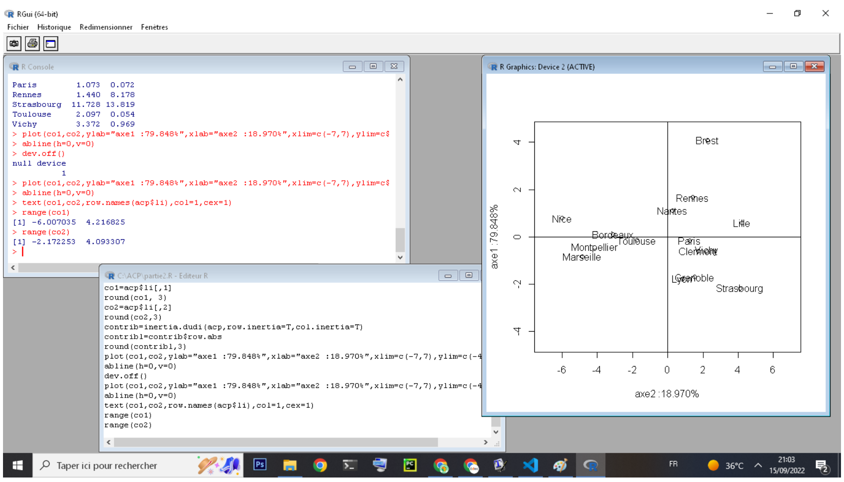

FIGURE 4.4 –

Représentation des variables

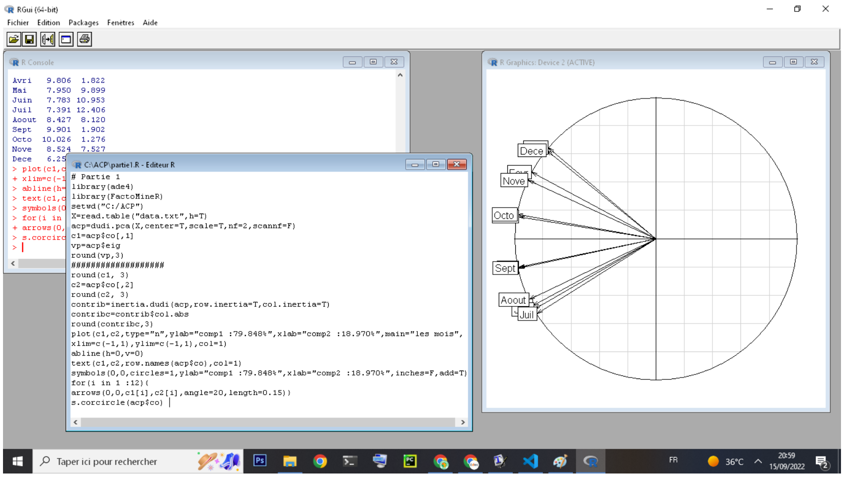

FIGURE 4.5 –

Conclusion générale

L'ACP est une technique de statistique descriptive dont le principe est simple mais qui met en oeuvre des calculs numériques importants, pour cette raison elle n'a pu se d´evelopper qu'avec l'apparition des ordinateurs.
L'ACP est à conseiller pour un premier examen, une mise en forme ou une présentation synthétique de données abondantes croisant des individus avec des variables quantitatives.

On n'omettra cependant pas d'examiner préalablement les données par les méthodes statistiques usuelles (moyenne, écart-type, graphiques, corrélation, etc.).

Un reproche fréquemment adressé à l'ACP et aux techniques connexes est qu'elles ne révéleraient que des évidences. Le propos est injuste, mais il est rassurant que souvent les premiers axes retrouvent et confirment ce qui était déjà connu.

Comme avec les autres méthodes descriptives, il faut être trés prudent pour inférer des modéles explicatifs ou causals à partir des configurations obtenues.

Notations

p	:Nombre de variables quantitatives
n	:nombre des individus
X	:tableaux de données
X^J	:j$^{\text{éme}}$ variable
D	:Matrice des poids
I_n	:la matrice identité (diagonale) d'ordre n
g	:le centre de gravité
$\overline{x^j}$	:la moyenne arithmétique de la j^{me} variable
R^n	:l'espace des variables
R^P	:l'espace des individus
Y	:Tableau de données centrés
e_i	:l'individu i
e_{ic}	:l'individu i de valeurs centrées
V	:Matrice variances-covariances
$V(X^j)$	:la variance de la variable X^j
cov	:covariance
$D_{\frac{1}{S}}$	:matrice diagonale des inverses des écart-type
$D_{\frac{1}{S^2}}$	:matrice diagonale des inverses des variances
Z	:tableaux de données centrés réduite
R	:matrice des corrélations
F	:l'espace vectoriel
$d^2_M(e_i, e_j)$	:La M distance entre les 2 individus $\left(e_i, e_j\right)$
S_j	:l'écart –type de la variable j
${S_j}^2$	:la variance de la variable j

l_n	:(1,. . . . . .,1) un vecteur de R^n
I_g	:inertie au centre de gravité g
I_a	:inertie en un point a
X_i^j	:la valeur de l'individus i pour la variable j
P	:matrice de projection
MVM	:matrice d'inertie du nuage
F_k	:le sous-espace de dimension k
M^{-1}	:la matrice inverse de M
P_i	:poids
θ_{jk}	:l'angle entre deux variables centrées
proj	:Projection
f_i	:projection de l'individu e_i
r	:le confficient de corrélation

Bibliographie

R. Boumaza. Analyse des données-acp, 2007.

J.M. Bouroche, G. Saporta. L'analyse des données, 1992.

F. Caillez. Introduction à l'analyse des données,smash, paris, 1982.

D. Chessel. Composantes principales, 2008.

C. Duby, S. Robin. Analyse en composantes principales . tna .paris grrignon, 2006.

Y. Escofier. Cours d'analyse de données, crig montpellier, 1979.

Husson. François. Laboratoire de mathématiques appliquées - agrocampus rennes, 2016.

J. Gergaid. Algèbre linéaire : une application l'analyse en composantes principales., octobre 2006.

L. Lebart. Statistique exploratoire multidimensionnelle, 1995.

B. Rousselet. Aide mémoire d'analyse de données, laboratoire de mathématiques,parc valrose, 2005.

G. Saporta. Probabilités, théories et méthodes de la statistique. technip, paris, 1980.

M. Volle. L'analyse des données, economica, paris $2^{\text{ème}}$ ed, 1980.