الجـــــــــــــــــــمـهـوريـــــة الجـــــــــــــزائـريـة الديمـــــــقـراطيـة الـشـــــــــعبيـة

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التـــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـــــــــــــامعة أبي بكـــر بـلـقـايد– تـلمســـــــــــــان –

Université Aboubakr Belkaïd– Tlemcen –
Faculté des SCIENCES

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : MATHEMATIQUE

Spécialité : Statistiques et probabilité approfondie

Par :
BOUNKHALA ASMA

Sujet

Méthodes ACP et AFC en statistiques et leurs applications

Soutenu publiquement, septembre 2017, devant le jury composé de :

Mr. MOURID. T	Professeur à l'Université de Tlemcen	Président
Mr. BENMANSOUR. D	Professeur à l'Université de Tlemcen	Examinateur
Mr. LABBES. A	M.C à l'Université de Tlemcen	Examinateur
Mme. BESSADAT. N	M.C à l'Université de Tlemcen	Encadreur

Année universitaire 2016-2017.

Remerciements

Je tiens tout d'abord à remercier Dieu le tout puissant et miséricordieux qui m'a donné la force et la patience d'accomplir ce modeste travail .

Je tiens à remercier avec ma plus grande gratitude ***Mr T. MOURID****, Professeur à l'université de Tlemcen de l'honneur qu'il me fait d'avoir accepté de présider le jury de ce mémoire.*

Je tiens à remercier mon encadreur ***Mme. N. BESSADAT****, M.C à l'université de Tlemcen pour ses précieux conseils et ses aides durant toute la période du travail.*

Mes vifs remerciements vont également aux membre de jury pour l'intérêt qu'ils ont porté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.

Je remercie également ***Mr D. BENMANSOUR*** *Professeur à l'université de Tlemcen et* ***Mr A. LABBES****, M. C à l'université de Tlemcen d'avoir accepté de se joindre à ce jury comme examinateurs.*

Je n'oublie pas de remercier vivement ***Dr M. HAMMIDOU*** *pour son soutien, ses conseils judicieux, et son aide précieuse.*

Enfin, je tiens également à remercier toutes les personnes qui ont participé de prés ou de loin à la réalisation de ce travail.

Dédicaces

Au nom du Dieu clément et misécordieux

J'ai l'immense honneur de dédier ce modeste travail

A mon cher père

Pour l'amour et l'éducation qu'il m'a donné

A ma chère mère

Pour son grand amour,ses sacrifices et toute l'affection qu'elle m'a toujours offerte

A ma chère grand mère

Qui m'a accompagné par ses prières, sa douceur, puisse Dieu lui prêter longue vie

A la memoire de mon cher oncle Arbi

Pour son soutien j'aurais tant aimé que vous soyez présents.

A ma chère soeur Hanifa et son mari Yacoub

A mon cher frère Mohammed et sa femme Sara

A mon cher frère Anes

En leurs souhaitant tout le succés...tout le bonheur

A ma chère nièce Lina

A mon cher neveu Imad Eddine

Aucune dédicace ne saurait exprimer tout l'amour que j'ai pour vous, Votre joie et votre gaieté me comblent de bonheur.

A mes chères amies Fatima Soumia Mehdia et Fatima

Pour les bons moments passés ensemble

Table des matières

Introduction générale

L'analyse des données est un sous domaine des statistiques qui se préoccupe de la description des données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différents données ainsi qu'à en tirer une information statistique qui sert à décrire les pricipales informations contenues dans ces données.

Dans le premier chapitre nous proposons un bref compte rendu de l'algèbre linéaire afin de bien fixer les idées générales autour desquelles s'articule la technique. Ainsi nous allons citer les différentes coefficients de corrélation qui servent à étudier les liaisons entre les variables observées.

Dans le chapitre suivant on va traiter une méthode de l'analyse des données qui est l'analyse en composantes principales. Cette méthode permet au praticien de résumer l'information en un nombre de composantes plus limités que le nombre d'origine de variables.

L'analyse factorielle des correspondances (AFC) est une méthode descriptive d'analyse proposée par **J.P.Benzekri**(1970-1990) permettant d'étudier un tableau de contingence conduisant à une représentation graphique. C'est un outil qui permet de réduire la dimension des données en conservant le plus d'information possible. C'est ce que nous allons aborder dans le troisième chapitre.

Enfin, on termine notre travail par une application des deux techniques ACP et AFC sur des données réelles sur un groupe de cent cinquante personnes portant sur l'indemnisation des assurances suites à des accidents corporels.

Chapitre 1

Notions de base

1.1 Espace vectoriel :

Dans ce qui suit K est un corp que l'on supposera égale à R (ou C,ou Q).
définition 1.1 : On appelle K-espace vectoriel tout ensemble E muni de deux lois :

1. $+_E : E \times E \mapsto E$, appelé "addition" ou "loi interne".

2. $\times_E : E \times E \mapsto E$, appelé "multiplication" ou "loi externe".

qui satisfont aux axiomes suivants :
— La loi $+_E$ est associative et commutative.

— La loi $+_E$ admet un élément neutre(à droite et à gauche).

— Tout élément de E a un opposé pour la loi $+_E$.

Autrement dit $+_E$ est une loi de groupe commutatif ou Abélein sur E.

— La loi $\times_E$ est associative.
— La multiplication par l'unité du corp est l'identité de E.
— La loi $\times_E$ est distributive par rapport à $+_E$.
— La multiplication par un élément de E est linéaire de K dans E.
Les éléments d'un éspace vectoriel sont appelés vecteur et les éléments de K sont appelés scalaires.

Définition 1.2 : [1]
Soit E un espace vectoriel sur K et soit F un sous ensemble de E. On dit que F est un sous espace vectoriel de E si F possède les propriétés suivantes :

1. 0_F.
2. $\forall x, y \in F, x + y \in F$ (F est stable par addition).
3. $x \in F, \forall \lambda \in K, \lambda x \in F$(F est stable par la multiplication par scalaire).

corollaire 1.3 : [1]
Soit E un espace vectoriel et $F \subset E$ si F vérifie les propriétés 1et 2 suivantes alors F est un sous espace vectoriel de$E$

1. F est non vide (F contient l'élément neutre de E).
2. $\forall (x, y) \in F \times F, \forall (\lambda, \mu) \in K \times K$, alors $\lambda x + \mu y \in F$.

Soient maintenant E un K espace vectoriel et $\mathscr{F} = (e_i)_{1\leq i\leq n}$ une famille finie de vecteurs de E.

définition 1.4 :
On appelle espace vectoriel engendré par la famille
$\mathscr{F} = (e_i)_{1\leq i\leq n}$ le sous espace vectoriel engendré par $\{e_1, e_2, ..., e_n\}$. On le note $vect(\mathscr{F})$ donc un elément x de E appartient à $vect(\mathscr{F})$ si et seulement si $\exists e_1, e_2, ..., e_n \in \mathscr{F}$ et des scalaires $\lambda_1, \lambda_2, ..., \lambda_n$
tels que $x = \lambda_1 e_1 + \lambda_2 e_2 + ... + \lambda_n e_n$

définition 1.5 : [1]
On dit qu'une famille $\mathscr{F} = (e_i)_{1\leq i\leq n}$ de vecteur de E est
génératrice de E si tout vecteur x de E s'écrit comme combinaison linéaire des vecteurs de la famille $\mathscr{F}$, c'est à dire :
$\forall x \in E, \exists(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n) \in K^n \backslash x = \lambda_1 e_1 + ... + \lambda_n e_n = \sum_{1\leq i\leq n} \lambda_i e_i$.

définition 1.6 : [1]

1. On dit qu'une famille$(e_1, .., e_n)$ de vecteur de E est linéairement indépendant ou libre si elle vérifie

$$\forall \lambda_1, ..., \lambda_n \in K, \lambda_1 e_1 + ... + \lambda_n e_n = 0 \Rightarrow \lambda_1 = \lambda_2 = ... = \lambda_n = 0$$

2. On dit que la famille $(e_1, .., e_n)$ est liée si elle n'est pas libre.

définition 1.7 : [1]
On dit qu'une famille $\mathscr{B} = (e_i)_{1\leq i\leq n}$ de vecteur de E est une base de E si celle ci est libre et génératrice.

Théorème 1.8 : [1]
Si $\mathscr{B} = (e_i)_{1\leq i\leq n}$ est une base d'un K_espace vectoriel de E alors $\forall x \in E$,
$\exists(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n) \in K^n, x = \lambda_1 e_1 + ... + \lambda_n e_n$ avec $\lambda_1, \lambda_2, \lambda_3, ..., \lambda_n$ sont appelés les composantes de x dans la base $\mathscr{B}$.

définition 1.9 : [1]
Soit F et G deux sous-espaces vectoriels de E.
On dit que F et G sont en somme directe si tout vecteur w de $F + G$ s'écrit de façcon unique sous la forme $w = u + v$ avec $u \in F$ et $v \in G$. On écrit alors

$$E = F \oplus G$$

.
On dit que F et G sont supplémentaires .

Théorème 1.10 : [1]
Les sous-espaces vectoriels F et G sont en somme directe si et seulement si $F \cap G = \{\overrightarrow{0}\}$
Par conséquent F et G sont supplémentaires si et seulement si $F + G = E$ et $F \cap G = \{\overrightarrow{0}\}$
Preuve :
Supposons que F et G sont en somme directe. On a $\overrightarrow{0} \in F$ et $\overrightarrow{0} \in G$,
donc $\overrightarrow{0} \in F \cap G$

Soit $u \in F \cap G$ .On peut écrire :

$$u = \underbrace{u}_{\in F} + \underbrace{\overrightarrow{0}}_{\in G} = \underbrace{\overrightarrow{0}}_{\in F} + \underbrace{u}_{\in G}$$

Par unicité de la décomposition, on a $u = \overrightarrow{0}$ .Donc $F \cap G = \{\overrightarrow{0}\}$.
Réciproquement supposons que $F \cap G = \{\overrightarrow{0}\}$ .Soit $u, u' \in F$ et $v, v' \in G$ tels que $u+v = u'+v'$ alors $u - u' = v' - v$ .De plus $u - u' \in F$ et $v' - v \in G$ donc si on pose $w = u - u' = v' - v$ on a $w \in F \cap G = \{\overrightarrow{0}\}$.
On en déduit que $w = \overrightarrow{0}$,donc $u = u'$ et $v = v'$ .Par conséquent tout vecteur de $F + G$ s'écrit de façon unique comme somme d'un vecteur de F et d'un vecteur de G autrement dit F et G sont en somme directe.
Le deuxième point du théorème découle directement du premier point et la définition de supplémentaire.

1.2 Les matrices :

1.2.1 définitions :

Soient $n, p \in N^*$, une matrice (n,p) à coefficient dans K est une application de $[1, n] \times [1, p]$ dans K.
On représente la matrice M par un tableau à n lignes et p colonnes, en mettant l'élément $m_{i,j}$ à l'intersection de la i ème ligne et la j ème colonne.

$$M = \begin{pmatrix} m_{1,1} & . & . & m_{1,p} \\ . & . & . & . \\ . & . & . & . \\ m_{n,1} & . & . & m_{n,p} \end{pmatrix}$$

L'ensemble des matrices de type (n,p) à coefficients dans K est noté $M_{n,p}(K)$.
Lorsque n=p ,on notera $M_n(K)$.
Une matrice $M_n(K)$ est dite symétrique (resp.antisymétrique) lorsque
$a_{i,j} = a_{j,i}$ (resp. $a_{i,j} = -a_{i,j}$).
Une matrice $M_n(K)$ est dite triangulaire supérieur(resp.inférieur) lorsque $a_{i,j} = 0$ pour tout couple (i,j) telle que $i \geq j$ resp($i \leq j$).
Une matrice $M_n(K)$ est dite diagonale lorsque $a_{i,j} = 0$ pour tout couple (i,j) telle que $i \neq j$.

1.2.2 Les opérations sur les matrices :

L'addition :
Si $A = (a_{i,j})$ et $B = (b_{i,j})$ sont des matrices $m \times n$ alors $A + B$ est une matrice $m \times n$ dont les coefficients sont donnés par :

$$c_{i,j} = a_{i,j} + b_{i,j} \quad [1]$$

. **La multiplication par un scalaire :**
Si $A = (a_{i,j})$ est une matrice $m \times n$ et $\alpha \in R$ alors αA est une matrice $m \times n$ dont les coefficients sont donnés par :

$$C = c_{i,j} = \alpha a_{i,j} \quad [1]$$

.

La multiplication de deux matrices : [1]
Si A=($a_{i,j}$) et $B = (b_{i,j})$ sont des matrices $m \times n$ et $n \times p$ alors AB est une matrice $m \times p$ dont les coefficients sont donnés par :

$$c_{i,j} = \sum_{1 \leq k \leq n} a_{i,k} b_{k,j}$$

.
Remarque :
il faut bien noter que la multiplication de deux matrices n'est définie que si le nombre de colonnes de la première matrice est égale au nombre de lignes de la seconde matrice.
proposition :
Les opérations sur les matrices vérifient les propriétés suivantes : [1]

1. $A(B + C) = AB + AC$ et $(A + B)C = AC + BC$(distributivité).
2. $(AB)C = A(BC)$(associativité).
3. $\alpha(AB) = (\alpha A)B = A(\alpha B)$ avec $\alpha \in K$ (compatibilité).

1.2.3 Transposition d'une matrice :

Définition 2.3.1 : [1]
Pour $A \in M_{n,p}(K)$, la matrice transposée de M noté M^t désigne la matrice $N \in M_{p,n}(K)$ telle que :

$$\forall (i,j) \in [1,p] \times [1,n] \ , \ N_{i,j} = M_{j,i}$$

proposition 2.3.2 [1]

— Si $A \in M_{n,p}(K)$ et $B \in M_{p,q}(K)$ alors $(AB)^t = B^t A^t$.

— Si $M \in M_{n,p}(K)$ alors $(M^t)^t = M$.

1.2.4 Le déterminant d'une matrice :

Le calcul du déterminant d'une matrice est un outil nécessaire tant en algèbre linéaire pour vérifier une inversibilité ou calculer l'inverse d'une matrice.

Définition 2.4.1 : [1]
Le déterminant d'une matrice carré A

$$A = \begin{pmatrix} a_{1,1} & . & . & a_{1,n} \\ . & . & . & . \\ . & . & . & . \\ a_{n,1} & . & . & a_{n,n} \end{pmatrix}$$

que l'on dénote par det(A) ou encore $\mid A \mid$ est donné par la formule de *Leibniz*

$$det(A) = \begin{vmatrix} a_{1,1} & . & . & a_{1,n} \\ . & . & . & . \\ . & . & . & . \\ a_{n,1} & . & . & a_{n,n} \end{vmatrix} = \sum_{\sigma \in \varsigma_n} \varepsilon(\sigma) \prod_{i=1}^{n} a_{\sigma_{(i)}} \quad [1]$$

où ς_n designe l'ensemble des permutations de $\{1, 2, 3, ..., n\}$
et $\varepsilon(\sigma)$ la signature de la permutation σ et elle est définie par :

$$\varepsilon(\sigma) = \prod_{\{i,j\}\in\mathscr{P}} \frac{\sigma(j) - \sigma(i)}{j - i}$$

où $\mathscr{P}$ désigne l'ensemble des paires d'entiers compris entre 1 et n.

Remarques :
- Soient $i < j$ deux entiers compris entre 1 et n. On dit que la paire{i, j} est en inversion pour σ si $\sigma(i) > \sigma(j)$.
- Une permutation est dite paire si elle présente un nombre pair d'inversions, impaire sinon.
- La signature d'une permutation paire est 1, celle d'une permutation impaire est –1.

Exemple :

$$A = \begin{pmatrix} -2 & 2 & -3 \\ -1 & 1 & 3 \\ 2 & 0 & -1 \end{pmatrix}$$

$det(A) = (-2).1(-1) + (-3).0.(-1) + 2.3.2 - (-3).1.2 - (-2).3.0 - 2(-1).(-1)$
$= 2 + 0 + 12 - (-6) - 0 - 2 = 18$

On peut aussi calculer le déterminant d'une matrice de taille n à l'aide de n déterminants de matrices de taille n - 1 .Si A est la matrice, pour tout i et j, on note $A_{i,j}$ la matrice obtenue en enlevant à A sa i-ième ligne et sa j-ième colonne.

$$A_{i,j} = \begin{pmatrix} a_{1,1} & . & . & . & a_{1,j-1} & a_{1,j+1} & . & . & . & a_{1,n} \\ . & & & & & & & & & \\ . & & & & & & & & & \\ . & & & & & & & & & \\ a_{i-1,1} & . & . & . & a_{i-1,j-1} & a_{i-1,j+1} & . & . & . & a_{i-1,n} \\ a_{i+1,1} & . & . & . & a_{i+1,j-1} & a_{i+1,j+1} & . & . & . & a_{i+1,n} \\ . & & & & & & & & & \\ . & & & & & & & & & \\ . & & & & & & & & & \\ a_{n,1} & . & . & . & a_{n,j-1} & a_{n,j+1} & . & . & . & a_{n,n} \end{pmatrix}$$

On peut alors développer le calcul du déterminant de A suivant une ligne ou une colonne.
Développement suivant la ligne i :

$$det(A) = \sum_{j=1}^{n} a_{i,j}(-1)^{i+j} det(A_{i,j}) \quad [1]$$

On appelle la quantité $(-1)^{i+j} det(A_{i,j})$ le cofacteur de $a_{i,j}$ que l'on dénote par $\triangle_{i,j}$.

Exemple :

le déterminant de la matrice précédente se développe aisément suivant la deuxième colonne, la plus avantageuse pour la disposition des zéros.

$$det \begin{pmatrix} -2 & 2 & -3 \\ -1 & 1 & 3 \\ 2 & 0 & -1 \end{pmatrix} = 2.(-1)^{1+2} \begin{vmatrix} -1 & 3 \\ 2 & -1 \end{vmatrix} + 1.(-1)^{2+2} \begin{vmatrix} -2 & -3 \\ 2 & -1 \end{vmatrix}$$

$$= (-2).((-1)(-1) - 2.3) + 1.((-2)(-1) - 2.(-3)) = (-2)(-5) + 8 = 18$$

Propriétés 2.4.2 : [1]

- Si on permute deux lignes ou deux colonnes, le déterminant change de signe .
- Si deux lignes ou deux colonnes sont identiques, le déterminant est nul .
- Si on multiplie tous les termes d'une même ligne ou d'une même colonne par un réel k, le déterminant est multiplié par k.
- Si une ligne ou une colonne est nulle, le déterminant est nul.
- $$det(A.B) = det(A).det(B)$$
- Le déterminant d'une matrice triangulaire et diagonale est le produit des coefficients diagonaux.
- Le déterminant d'une matrice triangulaire par blocs est le produit des déterminants des blocs diagonaux
$$det\begin{pmatrix} A & B \\ 0 & C \end{pmatrix} = det(A).det(C)$$

1.2.5 L'inverse d'une matrice :

Définition 2.5.1 :
Une matrice carrée A d'ordre n est dite inversible ou régulière s'il existe une matrice B d'ordre n, appelée matrice inverse de A et notée A^{-1} telle que :

$$AB = BA = I$$

.

Propriétés 2.5.2 : [1]

- Deux matrices A et B inversibles, leur produit est inversible et il est égale au produit des inverses effectué dans l'ordre inverse :
$$(AB)^{-1} = B^{-1}.A^{-1}$$
- L'inverse de la transposé est définie par :
$$[A']^{-1} = [A^{-1}]'$$
- Le déterminant de l'inverse est définie par :
$$| A^{-1} | = \frac{1}{| A |}$$

Remarque :
Une matrice carrée admettant un inverse est dite inversible ou régulière sinon elle est singulière.

La somme de deux matrices inversibles n'est pas toujours une matrice inversible.
exemple :

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Théorème 2.5.3 : [1]

soit A une matrice carrés $\in M_n(K)$

$$\boxed{\text{A est inversible} \Leftrightarrow \det(\text{A}) \neq 0}$$

Preuve :
" $\Rightarrow$ " Par multiplicativité du déterminant :

$$det(A)det(A^{-1}) = det(AA^{-1}) = det(I) = 1$$

Donc on a nécéssairement $det(A) \neq 0$
" $\Leftarrow$ "
Si $det(A) = 0$ alors pour tout matrice B on a $det(AB) = det(A)det(B) = 0$
et on ne peut pas avoir $AB = I$ donc A n'est pas inversible.

Définition 2.5.4 : [1]
Soit A une matrice carrée inversible alors son inverse est égale à :

$$A^{-1} = \frac{1}{det(A)} \begin{pmatrix} \triangle_{1,1} & . & . & \triangle_{1,n} \\ & . & & \\ & . & & \\ & . & & \\ \triangle_{n,1} & . & . & \triangle_{n,n} \end{pmatrix}'$$

Avec $\triangle_{i,j}$ sont les cofacteurs de $a_{i,j}$.

1.2.6 La trace d'une matrice :

Définition 2.6.1 : [1]
Soit $A = (a_{i,j})_{1 \leq i,j \leq n} \in M_n(K)$ une matrice carrée. On définie la trace de A par :

$$TrA = \sum_{1 \leq i \leq n} a_{i,i}$$

Proposition 2.6.2 : [1]

— $Tr(\alpha A + \beta B) = \alpha Tr(A) + \beta Tr(B)$, pour $A, B \in M_n(K), \alpha, \beta \in K$.
— $Tr(AB) = Tr(BA)$ pour $A, B \in M_n(K)$.
— Deux matrices semblable ont la même trace.

Remarque : A et B sont semblable ssi il existe une matrice P telle que $A = P^{-1}BP$

1.2.7 Matrice de passage :

Définition 2.7.1 : [1]
Soit E un K espace vectoriel muni d'une base B.
On appelle matrice des composantes dans B du vecteur x la matrice colonne de $M_{n,1}(K)$ telle que ses coeficients $\alpha_1, \alpha_2, ..., \alpha_n$ sont les composantes de x dans la base B on la note

$$Mat_B(x) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_n \end{pmatrix}$$

Définition 2.7.2 : [1]
Soit $F = (x_1, x_2, ..., x_n)$ une famille de vecteurs d'un K espace vectoriel E muni d'une base $B = (e_1, e_2, ..., e_n)$. Pour tout $1 \leq i \leq p$ notons c_i la colonne des composantes dans B du vecteur x_i
On appelle matrice des composantes dans la base B de la famille des vecteurs F la matrice de $M_{n,p}(K)$ dont les colonnes sont $c_1, c_2, ..., c_n$, on la note $Mat_B(F)$

Définition 2.7.3 : [1]
Soit E un K-espace vectoriel de dimension n muni de deux bases
$B = (e_1, e_2, ..., e_n)$ et $B' = (e'_1, e'_2, ..., e'_n)$.
On appelle matrice de passage de la base B à la base B' la matrice $P = Mat_B(B') = Mat_B(e'_1, e'_2, ..., e'_n)$

Exemple :
Soit l'espace vectoriel R^3 muni de la base canonique $B = (e_1, e_2, ..., e_n)$ et de la base $B' = (e'_1, e'_2, ..., e'_n)$ où $e'_1 = e_1 - e_2 + e_3$,$e'_2 = e_2 - e_3$ et $e'_3 = -2e_1 + 2e_2 - e_3$.
La matrice de passage de la base B à la base B' est :

$$Mat_B(B') = \begin{pmatrix} 1 & 0 & -2 \\ -1 & 1 & 2 \\ 1 & -1 & -1 \end{pmatrix}$$

1.2.8 Matrice de projection orthogonale :

Théorème : [1]
Soit $A \in M_{m,n}(K)$ et soit $y \in R^m$. Si la matrice A^tA est inversible, alors la projection orthogonale Y^* de Y sur le sous espace de R^m engendré par les colonnes de A est :

$$Y^* = PY$$

avec $P = A(A^tA)^{-1}A^t$
On appelle P la matrice de projection orthogonale sur le s.e.v de R^m et elle est symétrique et idempotente ($P^2 = P$).

1.2.9 Le rang d'une matrice :

Définition 2.9.1 :
Si $F = (x_1, x_2, ..., x_n)$ est une famille de vecteur d'un $K_e space$ vectoriel E alors on appelle rang de la famille F la dimension de l'espace engendré par F. On note $RgF = \dim vect(x_1, x_2, ..., x_n)$

Définition 2.9.2 : [1]
Soit $A = (a_{ij})_{(1 \leq i \leq n),(1 \leq j \leq p)} \in M_{np}(K)$ de colonnes $c_1, c_2, ..., c_p$ on appelle rang de A le rang de la famille $(c_1, c_2, ..., c_p)$.
On note $Rg(A) = Rg(c_1, c_2, ..., c_p)$.

1.3 Vecteurs et valeurs propres :

Définition 3.1 : [1]
Soit A une matrice carré de $M_n(K)$ et $\lambda \in K$, s'il existe un vecteur $X \in K^n$ telle que $AX = \lambda X$ alors :

1. λ est une valeur propre de A.

2. X est un vecteur propre de A.

Dans ce cas $E_\lambda = \{X \in M_{n,1}(K) \backslash AX = \lambda X\}$ est l'espace propre associé à la valeur propre λ.

Proposition 3.2 : [1]
Soient $A \in M_n(K)$ et $\lambda \in K$. Les conditions suivantes sont équivalentes :
i- λ est une valeur propre de A.
ii- $A - \lambda I_n$ n'est pas inversible.
iii- $det(A - \lambda I_n) = 0$
Preuve :
Il suffit de démontrer que $i \Rightarrow ii$
On a λ est une valeur propre de A, donc $\exists X \in M_{n,1}(K)$ un vecteur propre non nul associé à la valeur propre λ.
On a $AX - \lambda X = (A - \lambda I_n)X = 0$ si on pose $A - \lambda I_n$ est inversible,
on a $(A - \lambda I_n)^{-1}(A - \lambda I_n)(X) = 0$
Ceci implique que X est nul ce qui est absurde.

1.4 Diagonalisation d'une matrice carrée :

Diagonaliser une matrice carrée A, c'est trouver une matrice de passage P et une matrice diagonale D telle que :

$$P^{-1}AP = D \Rightarrow A = PDP^{-1} \text{ [1]}$$

avec les vecteurs colonnes de la matrice de passage P sont les vecteurs propres de A avec P inversible.

Calcul des vecteurs et valeurs propres :
D'aprés la proposition précédente on a :
λ est une valeur propre de A si et seulement si $det(A - \lambda I_n) = 0$.
En calculant ce déterminant on obtient un plynôme en λ de degré n de coefficient dominant $(-1)^n$.
Ce polynôme est appellé le polynôme caractéristique de A.

Théorème : [1]
λ est une valeur propre de $A \Leftrightarrow \lambda$ est une racine de polynôme caractéristique.

Exemple d'application :
On prend la matrice carrée A suivante :

$$\begin{pmatrix} 5 & -6 & -6 \\ -1 & 4 & 2 \\ 3 & -6 & -4 \end{pmatrix}$$

En calculant $det(A - \lambda I_3) = 0$ on obtient le polynôme caractéristique

$$P(\lambda) = -\lambda^3 + 5\lambda^2 - 8\lambda + 4$$

Une racine évidente est 1, pour trouver les autres racines on fait la division euclidienne de$-\lambda^3 + 5\lambda^2 - 8\lambda + 4$ par $\lambda - 1$
On trouve que le quotient est égale à

$$-\lambda^2 + 4\lambda - 4 = -(\lambda - 2)^2$$

donc A a deux valeurs propres qui sont 1 et 2 .
Cherchons maintenant une base des sous espaces propres E_1et E_2.
Pour E_1 c'est donc chercher les vecteurs V de E telle que
$V = (x, y, z)^t \in R^3$ satisfait au système linéaire $AV = V$ ou encore :

$$\begin{cases} 5x - 6y - 6z = x \\ -x + 4y + 2z = y \\ 3x - 6y - 4z = z \end{cases}$$

et de facons équivalente

$$\begin{cases} 4x - 6y - 6z = 0 \\ -x + 3y + 2z = 0 \\ 3x - 6y - 5z = 0 \end{cases}$$

On a un système homogène, qui a évidement une solution non triviale, car le déterminant de $A - I_3$ est nul (puisque 1 est une valeur propre).
Pour trouver toutes les solutions on détermine d'abord le rang de la matrice $A - I_3$ du système.
Les deux premières ligne et colonne fournissent le déterminant d'ordre 2
non nul $\begin{vmatrix} 4 & -6 \\ -1 & 3 \end{vmatrix} = 18$
Le rang est donc 2 et de plus on peut prendre pour équations principales les variables x et y et on résoud le système de Cramer

$$\begin{cases} 4x - 6y = 6z \\ 3x - 6y = 5z \end{cases}$$

et on obtient $x = z$ et $y = -\frac{z}{3}$.
Nous avons ainsi paramétré l'espace propre par l'inconnu non principale et on remarque que E_1 est de dimension 1.
Une base de E_1 est obtenue en prenant par exemple $z = 3$ et cette base est formée de l'unique vecteur $V = (3, -1, 3)^t$
Passons à E_2. On cherche donc les vecteurs V de E avec
$V = (x, y, z)^t \in R^3$ satisfait au système linéaire $AV = 2V$ ou encore $(A - 2I_3) = 0$, ou encore :

$$\begin{cases} 3x - 6y - 6z = 0 \\ -x + 2y + 2z = 0 \\ 3x - 6y - 6z = 0 \end{cases}$$

Cette fois ci, le rang de $A - 2I_3$ est 1.
On peut prendre la première équation principale, et x pour inconnu principale.
Le système de Cramer se réduit à l'unique équation

$$3x = 6y + 6z$$

. L'espace propre E_2 est de dimension 2 est paramétré par les deux inconnus non principales y et z c'est un plan d'équation

$$x = 2(y + z)$$

.
Une base de E_2 est donc obtenue par exemple en prenant (y,z)=(1,0) et (0,1), fournissant les vecteurs

$$f_2 = 2e_1 + e2$$

et

$$f_3 = 2e_1 + e3$$

.
Il est clair f_2 et f_3 sont indépendants
On a donc :

$$P = \begin{pmatrix} 3 & 2 & 2 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}$$

et

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Si on calcule P^{-1} par les techniques standard on obtient

$$P^{-1} = \begin{pmatrix} -1 & 2 & 2 \\ -1 & 3 & 2 \\ 3 & -6 & -5 \end{pmatrix}$$

et on a bien

$$A = PDP^{-1} = \begin{pmatrix} 3 & 2 & 2 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \quad \begin{pmatrix} -1 & 2 & 2 \\ -1 & 3 & 2 \\ 3 & -6 & -5 \end{pmatrix}$$

En statistique, étudier les phénomènes aléatoires revient parfois à étudier les liaisons entre différentes variables observées. L'étude qui met en évidence ces liens est ce qu'on appelle communément l'étude des corrélations.

1.5 coefficients de corrélation :

1.5.1 Le coefficient de corrélation linéaire :

Soient X et Y deux variables quantitatives de R^n.
On dit qu'il y a une corrélation s 'il y a une dépendance en moyenne c'est à dire à $X = x$ fixé, la moyenne $\bar{Y}$ est en fonction de X .Si cette liaison est linéaire on se trouve dans le cas de la corrélation linéaire.

Définition : [3]
Le coefficient de corrélation linéaire ou bien le coefficient de"Bravais Pearson" sert à caractériser une liaison linéaire positive ou négative, il est définie par :

$$\boxed{\mathrm{r}_{XY} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}}$$

avec σ_X et σ_Y sont les écart type de X et Y,$\sigma_{X,Y}$ la covariance entre X et Y.
Si on observe un échantillon de couples (X_i, Y_j) ,on estime le coefficient de corrélation par la formule suivante :

$$\boxed{\hat{r}_{X,Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}} \quad [3]$$

On appelle $\hat{r}_{X,Y}$ le coefficient de corrélation empirique et $\bar{X}$, $\bar{Y}$ la moyenne empirique de X et Y.
Ce coefficient varie entre -1 et 1, et l'intensité de la liaison est définie par la valeur absolue du coefficient.

- Si $r_{X,Y}$ est proche de 1, il y a une relation linéaire forte et croissante entre X et Y.
- Si $r_{X,Y}$ est proche de -1, il y a une relation linéaire forte et décroissante entre X et Y.
- Si $r_{X,Y}$ est proche de 0, il n' y a pas une relation linéaire entre X et Y.

Interprétation géométrique :
Si on considère dans R^n deux vecteurs centrés X' et Y'

$$X' = \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ . \\ . \\ X_n - \bar{X} \end{pmatrix} et \quad Y' = \begin{pmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ . \\ . \\ Y_n - \bar{Y} \end{pmatrix}$$

alors $r_{X,Y}$ est le cosinus de l'angle formé par ces deux vecteurs.En effet :

$$\langle X', Y' \rangle = \|X'\|\|Y'\| cos(\alpha)$$

avec

$$\|X'\| = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\|Y'\| = \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$$\langle X', Y' \rangle = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}$$

D'aprés la formule donnée dans la définition précédente :

$$\boxed{\cos(\alpha) = \frac{\langle X', Y' \rangle}{\|x'\|\|y'\|} = r_{XY}}$$

- Si r=1 $\Rightarrow \alpha = 0^o$,les vecteurs X' et Y' sont colinéaire.
- Si r= 0 $\Rightarrow \alpha = 90^o$, les vecteurs X' et Y' sont orthogonaux.
- Si r= -1 $\Rightarrow \alpha = 180^o$,les vecteurs X' et Y' sont colinéaire de sens opposé.

1.5.2 Le coefficient de corrélation partielle :

Souvent, deux variables semblent faiblement lié, mais on se rend compte après que la liaison repose sur l'intervention d'une troisième variable.

La corrélation partielle corrige cet inconvénient, elle mesure la liaison en annulant l'effet d'une ou plusieurs variables.

Définition : [3]
Soient X,Y et Z des variables aléatoires dans R^n, Le coefficient de corrélation partielle $r_{XY,Z}$ est définie à partir des corrélations de Pearson par :

$$\boxed{\mathrm{r}_{XY,Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1-r_{XZ}^2}\sqrt{1-r_{YZ}^2}}}$$

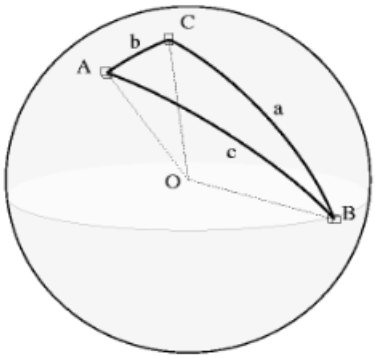
L'idée est de retrancher de r_{XY} le double effet des corrélations qu'entretiennent X et Y avec Z, puis un terme de normalisation est introduit de manière à ce que $-1 \leq r_{XY,Z} \leq 1$.
Remarque :

- L'orsque Z est independant de X et Y ($r_{XZ} = r_{YZ} = 0$) $r_{XY,Z} = r_{XY}$ c'est à dire Z ne pèse en aucune manière dans la relation entre X et Y.
- L' orsque Z est fortement lié positivement avec X et Y ,on peut aboutir au résultat $r_{XY,Z} \approx 0$ cela veut dire que dans la relation (X, Y) tout est expliqué par Z.

L'estimation de la corrélation partielle passe simplement par l'introduction des estimateurs des corrélations linéaires dans la formule précédente

$$\hat{r}_{XY,Z} = \frac{\hat{r}_{XY} - \hat{r}_{XZ}\hat{r}_{YZ}}{\sqrt{1-\hat{r}_{XZ}^2}\sqrt{1-\hat{r}_{YZ}^2}}$$

Preuve :
Une manière de la démontrer consiste à s'appuyer sur l'interprétation géométrique de la corrélation (cosinus) : Les séries d'observations $A = X$, $B = Y$ et $C = Z$, une fois centrées réduites, sont des vecteurs centrés OA, OB, OC de longueur unité.Leurs extrémités déterminent un triangle sphérique ABC, dont les côtés "a", "b" et "c" sont les arcs de grand cercle BC, AC et AB.



Les corrélations entre ces vecteurs sont :

$$r_{BC} = cos(a)$$

$$r_{AC} = cos(b)$$

$$r_{AB} = cos(c)$$

Alors la loi fondamentale des triangles sphériques donne pour l'angle C la relation suivante entre les cosinus

$$cos(C) = \frac{cos(c) - cos(a)cos(b)}{sin(a)sin(b)} = \frac{cos(c) - cos(a)cos(b)}{\sqrt{1 - cos(a)^2}\sqrt{1 - cos(b)^2}}$$

De même que "c" est l'angle entre les points A et B, vus du centre de la sphère, "C" est l'angle sphérique entre les points A et B, vus du point "C" à la surface de la sphère est $r_{AB.C} = cos(C)$ est la corrélation partielle entre A et B quand C est fixé.

1.5.3 Le coefficient de corrélation multiple :

Définition :

Soient une variable numérique Y et un ensemble de p variables numériques $X^1, X^2, ..., X^p$.
Le coefficient de corrélation multiple R est alors la valeur maximale prise par le coefficient de corrélation linéaire entre Y et une combinaison linéaire de X^i :

$$\boxed{\text{R}= \sup_{a_1,a_2,.....,a_n} r(Y; \sum_{i=1}^{p} a_i X^i)} \quad [3]$$

On a donc toujours $0 \leq R \leq 1$.
En d'autre termes, si on pose $Y^* = b_0 + b_1X^1 + ... + b_pX^p$, on désire que Y^*soit le plus proche possible de Y.
Alors si l'espace des variables R^n est muni de la métrique D, on exigera que $\|Y - Y^*\|^2$ soit minimale.
Donc $R = 1$ s'il existe une combinaison linéaire X^i telle que :

$$Y = a_0 + \sum_{i=1}^{p} a_i X^i$$

Interprétation géométrique :

Rappelons que le coefficient de corrélation est le cosinus de l'angle formé dans R^n par des variables centrées. R est donc le cosinus du plus petit angle formé par Y (centrée) et une combinaison linéaire des X^i centrées.

Considérons le sous-espace W de R^n engendré par les combinaisons linéaires des X^i et la constante 1.
R est alors le cosinus de l'angle formé par $Y - \bar{Y}$ et sa projection orthogonale $Y^* - \bar{Y}$ sur W.
On a donc :

$$R = \frac{cov(Y, Y^*)}{\sigma_Y \sigma_{Y^*}} \quad [3]$$

Et puisque :

$$cov(Y, Y^*) = \|Y - \bar{Y}\|\|Y^* - \bar{Y}\|cos(Y - \bar{Y}, Y^* - \bar{Y})$$
$$\sigma_Y = \|Y - \bar{Y}\| \qquad \sigma_{Y^*} = \|Y^* - \bar{Y}\|$$

Alors

$$R = \frac{cov(Y, Y^*)}{\sigma_Y \sigma_{Y^*}} = cos(Y - \bar{Y}, Y^* - \bar{Y})$$

Comme tout coefficient de corrélation linéaire, son carré s'interprète en terme de variance expliquée :

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Soit P la matrice de projection orthogonale sur W , alors on a :

$$\|Y^* - \bar{Y}\|^2 = (Y^* - \bar{Y})'(Y^* - \bar{Y}) = (PY - \bar{Y})'(PY - \bar{Y})$$
$$= Y'P'PY - Y'A'\bar{Y} - \bar{Y}'PY + \bar{Y}'\bar{Y}$$
$$= YP'Y - Y'P\bar{Y} - \bar{Y}'PY + \bar{Y}'P\bar{Y} = (Y - \bar{Y})'P(Y - \bar{Y})$$

Donc on a bien :

$$R^2 = \frac{(Y - \bar{Y})'P(Y - \bar{Y})}{\|Y - \bar{Y}\|^2}$$

En particulier si Y est centré, c'est-à-dire $\bar{Y} = 0$ on a :

$$R^2 = \frac{Y'PY}{\|Y\|^2}$$

1.5.4 Coefficients de corrélation sur les rangs :

Souvent, on ne dispose que d'un ordre sur un ensemble d'individus et non de valeurs numériques d'une variable mesurable : soit par ce qu'on a que des données du type classement (classement A, B, C, D, E...etc.), ou bien par ce que les valeurs numériques d'une variable n'apportent que peu comparé à leur ordre (notes de copies,...etc.) .

L'idée est d'associer à chaque individu de 1 à n son rang selon une variable. Nous créons donc deux nouvelles colonnes $r_i = rang(X_i)$
et $s_i = rang(Y_i)$, avec $r_i$et s_i sont des permutations différentes des n premiers entiers.

Le coefficient de Spearman :

Le coefficient de corrélation de Spearman (nommé daprès Charles Spearman), ou rho de Spearman ρ,mesure la liaison monotone entre 2 variables X et Y.
Il est calculé sur les rangs (r_i, s_i) des variables (X_i, Y_i).Autrement dit ,il n' est ni plus ni moins que le coefficient de Pearson calculé sur les rangs.
alors on a :

$$r_s = \frac{cov(r,s)}{\sigma_r \sigma_s} = \frac{\sum_{i=0}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=0}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=0}^{n}(s_i - \bar{s})^2}}$$

Compte tenu du fait que les rangs soient des permutations de $[1....n]$ et de l'absence d'ex aequo on a :

$$\bar{r} = \bar{s} = \frac{1}{n}\sum_{i=0}^{n} i = \frac{1}{n}\frac{n}{2}(n+1) = \frac{n+1}{2}$$

$$\sigma_s^2 = \sigma_r^2 = \frac{1}{n}\sum_{i=0}^{n}(r_i - \bar{r})^2 = \frac{n^2-1}{12}$$

$$r_s = \frac{\frac{1}{n}\sum_{i=0}^{n} r_i s_i - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}} \quad [3]$$

En posant $d_i = r_i - s_i$ différence des rangs d'un même objet selon les deux classements, on a :

$$\sum_{i=0}^{n} r_i s_i = -\frac{1}{2}\sum_{i=0}^{n} -2r_i s_i = -\frac{1}{2}\sum_{i=0}^{n}[(r_i - s_i)^2 - r_i^2 - s_i^2]$$

$$= -\frac{1}{2}\sum_{i=0}^{n}(r_i - s_i)^2 + \frac{1}{2}\sum_{i=0}^{n} r_i^2 + \frac{1}{2}\sum_{i=0}^{n} s_i^2$$

mais :

$$\sum_{i=0}^{n} r_i^2 = \sum_{i=0}^{n} s_i^2 = \frac{n(n+1)(2n+1)}{6}$$

d'où :

$$r_s = \frac{-\frac{1}{2n}\sum_{i=0}^{n} d_i^2 + \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}}{\frac{(n^2-1)}{12}}$$

$$= -\frac{6\sum_{i=0}^{n} d_i^2}{n(n^2-1)} + \frac{2(n+1)(2n+1) - 3(n+1)^2}{n^2-1}$$

et on a l'expression équivalente pratique :

$$r_s = 1 - \frac{6\sum_{i=0}^{n} d_i^2}{n(n^2-1)}$$

La définition de r_s comme coefficient de corrélation linéaire sur des rangs nous indique que :
$r_s = 1 \Rightarrow$ les deux classements sont identiques.
$r_s = -1 \Rightarrow$ les deux classements sont inverses l'un de l'autre.
$r_s = 0 \Rightarrow$ les deux classements sont indépendants.

Le coefficient de corrélation des rangs τ de M.G.Kendall :

Le τ de Kendall est définie pour mesurer l'association entre variables ordinales en reposant sur la notion de paires discordantes et concordantes.
Soient (X_i, Y_i) et (X_j, Y_j) des réalisations indépendantes du couple(X,Y)
1. On dit que les paires observations i et j sont concordantes si et seulement si ($X_i > X_j$ alors $Y_i > Y_j$) ou ($X_i < X_j$ alors $Y_i < Y_j$). Nous pouvons simplifier l'écriture avec $(X_i - X_j)(Y_i - Y_j) > 0$
2. On dit que les paires sont discordantes lorsque ($X_i > X_j$ alors $Y_i < Y_j$) ou ($X_i < X_j$ alors $Y_i > Y_j$).En d'autres termes $(X_i - X_j)(Y_i - Y_j) < 0$

Le τ de Kendall théorique, calculé sur la population, est défini par :

$$\tau = 2.P((X_i - X_j)(Y_i - Y_j) > 0) - 1$$

Pour un échantillon de taille n,le τ de Kendall est définie de la manière suivante :

$$\hat{r} = \frac{P - Q}{\frac{1}{2}n(n-1)} \quad [3]$$

Avec P (resp Q) représente le nombre de paires concordantes (resp. discordantes).
Le dénominateur représente le nombre total de paires possibles :

$$\frac{1}{2}n(n-1) = C_n^2$$

Interprétation :
le τ de Kendall s'interprète comme le degré de correspondance entre 2

classements.
- Si $\tau = 1 \Rightarrow$ toutes les paires sont concordantes c.à-d. le classement selon x concorde systématiquement avec le classement selon Y.
- Si $\tau = -1 \Rightarrow$ toutes les paires sont discordantes.
- Si $\tau = 0 \Rightarrow$ les deux classements sont totalement indépendants.

Calcul pratique :
La manière la plus simple de calculer τ est de trier les données selon X puis de comptabiliser la quantité suivante :

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \nu_{ij} \quad [3]$$

où

$$\nu_{ij} = \begin{cases} +1 \ , si \ \ Y_i < Y_j \\ -1 \ , si \ \ Y_i > Y_j \end{cases}$$

S est donc l'écart entre le nombre total de paires concordantes, et le nombre total de paires discordantes $S = P - Q$.
Nous pouvons dès lors ré-écrire le coefficient de Kendall :

$$\hat{r} = \frac{2S}{n(n-1)}$$

Coefficient de Daniels et de Guttman :

Les trois coefficients de corrélation (Pearson, Speannan, Kendall) peuvent être présentés comme 3 cas particuliers d'une même formule, dite formule de Daniels. En effet si on considère pour toute paire d'individus i,j deux indices a_{ij} et b_{ij} le premier associé à la variable X et le deuxième associé à la variable Y ,et on définit le coefficient suivant :

$$\frac{\sum\sum a_{ij}b_{ij}}{\sqrt{\sum\sum a_{ij}^2 \sum\sum b_{ij}^2}} \quad [3]$$

qui varie entre - l et + 1 d'après l'inégalité de Schwarz.
-si on pose $a_{ij} = x_i - x_j$ et $b_{ij} = y_i - y_j \Rightarrow$ On trouve r le coefficient de Pearson.
-si on pose $a_{ij} = r_i - r_j$ et $b_{ij} = s_i - s_j \Rightarrow$ On obtient le coefficient de Spearman.
-si on pose $a_{ij} =$ le signe $(x_i - x_j)$ et $b_{ij} =$ le signe$(y_i - y_j) \Rightarrow$ On trouve le coefficient τ de Kendall.

1.6 Matrice de covariance :

Lorsque l'on observe les valeurs numériques de p variables sur n individus on se trouve en présence d'un tableau X à n lignes et p colonnes.
x_i^j est la valeur prise par la variable j sur le i^{me} individu.
Le tableau des données centrées Y s'obtient en utilisant l'opérateur de centrage :

$$A = I - \frac{11'}{n}$$

qui est une matrice de $n \times n$ de terme générale :
Avec 1 un vecteur de R^n dont toutes les composantes sont égale à 1.

$a_{ii} = 1 - \frac{1}{n}$ $\quad a_{ij} = -\frac{1}{n}$ si $i \neq j$
Alors on a bien :

$$Y = AX$$

La matrice des variances et covariances des p variables :

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & . & \sigma_{1p} \\ . & & & \\ . & & & \\ . & & & \\ . & . & . & \sigma_p^2 \end{pmatrix}$$

où : $\sigma_{kl} = \frac{1}{n} \sum_{i=1}^{n} x_i^k x_i^l - \bar{x}^k \bar{x}^l$
Alors la matrice de covariance V est définie par :

$$V = \frac{1}{n} Y'Y \quad [3]$$

Exemple :
Le US Census Bureau recueille des statistiques comparant les différents 50 états. Le tableau suivant montre le taux de pauvreté (% de la population en dessous du seuil de pauvreté) et le taux de mortalité infantile pour 1 000 naissances vivantes) par état, le taux de crime, de medecin ... etc. [5]
Les données pour les premiers états sont décrites dans la figure suivante :

Etats :	pauvreté	Enfants morts	Pourcentage des blancs	Crimes	médecins	Les victimes des accidents	universités	chômage	revenu
Alabama	15.7	9	71	448	218.2	1.81	22	5.0	42.666
Alaska	8.4	6.9	70.6	661	228.5	1.63	27.3	6.7	68.460
Arizona	14.7	6.4	86.5	483	209.7	1.69	25.1	5.5	50.958
Arkansas	17.3	8.5	80.8	529	203.4	1.96	18.8	5.1	38.815
Californie	13.3	5.0	76.6	523	268.7	1.21	29.6	7.2	61.021
Colorado	11.4	5.7	89.7	348	259.7	1.14	35.6	4.9	56.993
Connecticut	9.3	6.2	84 .3	256	376.4	0.86	35.6	5.7	68.595
Delaware	10	8.3	74.3	689	250.9	1.23	27.5	4.8	57.989
Florida	13.2	7.3	79.8	723	247.9	1.56	25.8	6.2	47.778
Georgia	14.7	8.1	65.4	493	217.4	1.46	27.5	6.2	50.861
Hawaii	9.1	5 .6	29.7	273	317	1.33	29.1	3.9	67.214
Idaho	12.6	6.8	94.6	239	168.8	1.6	24	4.9	47.576

.
On propose d'étudier la relation existant entre les variables suivantes : pauvreté, enfants, les blancs, crimes et médecin du tableau :

$$\begin{pmatrix} & pauvrete & enfants & blancs & crimes & medecin \\ pauvrete & 8.19659 & 1.74500 & 13.98022 & 58.83864 & -94.394 \\ enfants & 1.74500 & 1.630606 & 1.27409 & 86.33788 & -33.794242 \\ blancs & 13.98022 & 1.27409 & 277.94750 & 115.49318 & -316.28045 \\ crimes & 58.83863 & 86.33787 & 115.49318 & 28065.5378 & -2670.75606 \\ medecins & -94.39409 & -33.79424 & -316.28045 & -2670.75606 & 3067.93969 \end{pmatrix}$$

1.7 Matrice de corrélation :

On appelle matrice de corrélation la matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux qu'on la note R

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

En posant :

$$D_{\frac{1}{\sigma}} = \begin{pmatrix} \frac{1}{\sigma_1} & . & . & . & 0 \\ . & \frac{1}{\sigma_2} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & \frac{1}{\sigma_p} \end{pmatrix}$$

la matrice diagonale des inverses des écarts types. On a :

$$R = D_{\frac{1}{\sigma}} V D_{\frac{1}{\sigma}} \quad [3]$$

Exemple :
On calcule la matrice de correlation des variables de l'exemple précédent :

$$\begin{pmatrix} & pauvrete & enfants & blancs & crimes & medecin \\ pauvrete & 1.00000 & 0.47731 & 0.29289 & 0.12267 & -0.59525 \\ enfant & 0.47731 & 1.00000 & 0.05984 & 0.40358 & -0.47779 \\ blancs & 0.29289 & 0.05984 & 1.00000 & 0.04135 & -0.34250 \\ crimes & 0.12267 & -0.40358 & 0.04135 & 1.00000 & -0.28782 \\ medecin & -0.59525 & -0.47779 & -0.34250 & -0.28782 & 1.00000 \end{pmatrix}$$

Conclusion :
On remarque que les variables ne sont pas trés corrélées entre elles.

1.8 Tests sur la corrélation :

Le coefficient de corrélation mesure l'intensité de la relation existant entre deux variables X et Y. le problème posé est la significativité de ce coefficient qui peut être vérifiée statistiquement par le test des hypothèses suivantes :
$H_0 : \rho_{xy} = 0$ contre $H_1 : \rho_{xy} \neq 0$

1.8.1 coefficient de corrélation de Pearson :

test de significativité :
On sait bien que si dans une population deux variables ne sont pas corrélées leur coefficient de corrélation n'est pas exactement 0,donc il faut déterminer si r_{xy} est significativement différent de 0.Pour cela il faut faire un test d'indépendance en procédant de la façon suivante :
$H_0 = r_{xy} = 0$ absence de liaison entre X et Y.
$H_1 = r_{xy} \neq 0$ ou bien $H_1 = r_{xy} >$ ou $r_{xy} \leq 0$
Le test étudié dans cette section est paramétrique .On suppose que les observations proviennent d'un couple gaussien .
Sous H_0 la statistique du test est définie par :

$$T = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \quad [3]$$

T suit une loi de student à (n - 2) degrés de liberté avec R est le coefficient de corrélation empirique.
Ce test a comme région critique au risque α :

$$D = \{\mid T \mid > T_{1-\frac{\alpha}{2}}(n-2)\}$$

où $T_{1-\frac{\alpha}{2}}(n-2)$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de student à (n - 2) degrés de liberté.
Exemple numérique :

Toujours avec les données numériques des 12 états :
Le coefficient de corrélation entre la pauvreté et le taux de mortalité des enfants vaut :$r = 0.47731$
On souhaite tester sa significativité au risque $\alpha = 0.05$
Nous devons calculer les éléments suivants :
- La statistique du test : $T = \frac{0.47731}{\sqrt{1-0.47731^2}}\sqrt{(12-2)} = 1.71768$
- Le seuil théorique au risque α est : $t_{0.975(12-2)} = 2.228$

On accepte l'hypothèse nulle car $T < t_{0.975(12-2)}$ donc il n'existe pas une liaison linéaire entre les deux variables .

Test asymptotique :
Lorsque n est assez grand ($n > 100$) la loi de R suit approximativement une loi normale $N(0, \frac{1}{\sqrt{n-1}})$ [3]
.

1.8.2 Signification d'un coefficient de corrélation partielle :

Dans le cas gaussien, la loi du coefficient de corrélation partielle est même que celle d'un coefficient de corrélation simple mais avec un degré de liberté diminué de d le nombre de variables fixées.
Donc $\frac{R}{\sqrt{1-R^2}}\sqrt{n-d-2}$ suit une $T_{(n-d-2)}$, ce qui permet de tester le caractère significatif d'une liaison partielle.
Exemple numérique : [6]
Sur 33 parcelles forestières on a observé 8 variables qui sont :

X_1	altitude (en m)
X_2	pente (en degré)
X_3	nombre de pins moyens dans une placette de 5 ares
X_4	hauteur de l'arbre échantillonné au centre de la placette
X_5	diamètre de cet arbre
X_6	note de densité du peuplement
X_7	orientation de la placette (1 : sud, 2 : autres
X_8	hauteur (en m) des arbres dominants

Nous cherchons à caractériser la liaison entre "l'altitude" et "le nombre de pins", la variable de contrôle est "la pente".
La matrice de corrélation est la suivante :

$$\begin{pmatrix} & x_1 & x_2 & x_3 \\ x_1 & 1.000 & 0.121 & 0.538 \\ x_2 & 0.121 & 1.000 & 0.322 \\ x_3 & 0.538 & 0.322 & 1.000 \end{pmatrix}$$

$$\hat{r}_{x_1x_3,x_2} = \frac{0.538 - (0.121)(0.322)}{\sqrt{1-(0.121)^2}\sqrt{1-(0.322)^2}} = 0.531$$

Au seuil $\alpha = 5\%$ la valeur critique est :2.042
La valeur de la statistique t est :3.662
On a : $2.042 < 3.662$, donc la liaison est significative.

1.8.3 Test de significativité du coefficient de corrélation multiple :

Supposons que les n observations de Y vérifient les conditions de normalité et d'indépendance des X^i alors on a bien :

$$\frac{R^2}{1-R^2}\frac{(n-p-1)}{p} = \mathscr{F}(p, n-p-1) \quad [3]$$

Avec $\mathscr{F}(p, n-p-1)$ est une loi de fisher de (p,n - p - 1) degrés de liberté.

Remarque :
Si on pose $p = 1$ on trouve la loi de coefficient de corrélation linéaire.

1.8.4 Signification d'un coefficient de Spearman :

En présence d'un échantillon de n couple de rangs : $(r_1, s_1), (r_2, s_2), ..., (r_n, s_n)$ obtenus, soit par observation directe d'un couple de rangs (R, S), soit par transformation en rangs des valeurs d'un couple(X,Y) de valeurs réelles :
Lorsque $n \leq 100$, on se rapportera à la table du coefficient de corrélation de Spearman.
La région critique est $R_s > k$:
- Si $R_s > k$: il y a concordance de classements.
- Si $R_s < -k$: il y a discordance de classements.
Lorsque $n > 100$, on admet que R_s , est distribué comme une normale $N(0, \frac{1}{\sqrt{n-1}})$ et donc il suffit de comparer $R_s\sqrt{n-1}$ à la valeur critique lue dans une table de loi normale centrée réduite.
Lorsque les observations proviennent d'un couple normal (X, Y)
de corrélation ρ et que l'on calcule r_s à la place de r, si n est grand on a les relations approchées suivantes :

$$r_s = \frac{6}{\pi}\sin^{-1}(\frac{\rho}{2}) \quad [3]$$

où

$$\rho = 2\sin(\frac{\pi}{6}r_s)$$

Exemple :
Supposons que nous examinons 10 pays. On désire savoir s'il existe une corrélation entre l'espérance de vie et le produit intérieur brut (PIB) de ces pays.
En nous référant au CIA World Factbook nous obtenons la liste des 10 pays dont l'espérance de vie est la plus élevée.

Pays	Espérance de vie (en année)	Rang mondial
Macau	84,36	1
Andorre	82,51	2
Japon	82,12	3
Singapour	81,98	4
San Marino	81,97	5
Hong Kong	81,86	6
Australie	81,63	7
Canada	81,23	8
France	80,98	9
Suède	80,86	10

Nous devons maintenant utiliser les rangs concernant le PIB. Nous obtenons le tableau suivant pour ces mêmes 10 pays.

Pays	PIB ($ US)	Rang mondial	Rang corrigé
Macau	30,000	44	10
Andorre	42,500	15	3
Japon	34,200	36	8
Singapour	52,000	9	1
San Marino	41,900	16	4
Hong Kong	43,800	14	2
Australie	38,100	26	7
Canada	39,300	21	5
France	32,700	38	9
Suède	38,500	23	6

nous obtenons donc le tableau suivant :

Pays	Rang Mondial (Espérance de vie)	Rang Corrigé (PIB)	D (Différence entre les deux rangs)	D^2 (Différence au carré entre les deux rangs)
Macau	1	10	-9	81
Andorre	2	3	-1	1
Japon	3	8	-5	25
Singapour	4	1	+3	9
San Marino	5	4	+1	1
Hong Kong	6	2	+4	16
Australie	7	7	0	0
Canada	8	5	+3	9
France	9	9	0	0
Suède	10	6	+4	16
Total			0	158

On a :

$$r_s = 1 - \frac{6\sum_{i=0}^{n} d_i^2}{n(n^2-1)} = 0,0424$$

Au risque $\alpha = 5$, la valeur critique, d'après la table du coefficient de corrélation de Spearman, est :

$$k = 0,648$$

On a $r_s < k$ donc il y a discordance de classements.

1.8.5 Signification du coefficient de Kendall :

Dès que $n \geq 8$, et plus sûrement lorsque n $>$ 10 nous pouvons nous appuyer sur la normalité asymptotique de $\hat{r}$ sous l'hypothèse d'indépendance de X et Y.
Le test de significativité repose alors sur la statistique :

$$U = 3\hat{r}\sqrt{\frac{n(n-1)}{n(2n+5)}} \quad [3]$$

Sous H_0 $(\tau = 0)$ $U \sim N(0,1)$
La région critique du test pour un risque α s'écrit :

$$| U | > U_{1-\frac{\alpha}{2}}$$

Exemple :
Mettons en relation la taille et le poids de 8 personnes, les données sont triées selon la taille de la personne.

	Taille	Poids							
1	1.496	67.585	67.585						
2	1.500	58.068	-1	58.068					
3	1.539	55.000	-1	-1	55.000				
4	1.542	71.668	+1	+1	+1	71.668			
5	1.543	58.060	-1	-1	+1	-1	58.060		
6	1.557	61.689	-1	+1	+1	-1	+1	61.689	
7	1.577	70.060	+1	+1	+1	-1	+1	+1	70.060
8	1.621	68.039	+1	+1	+1	-1	+1	+1	-1
Somme			-1	+2	+5	-4	+3	+2	-1

On a : $\hat{r} = 0.214$ et $U = 0.741$
Pour un risque $\alpha = 5\%$,le seuil critique du test est : $U_{0,975} = 1,96$.
Donc il n y a pas de liaison significative entre les deux classements.

Chapitre 2

Analyse en composantes principales

Une ACP est une étude exploratoire appliquée à un tableau de données où on cherche les ressemblances entre les individus et les liaisons entre les variables en résumant l'ensemble des variables par un petit nombre de variables synthétiques appelées composantes principales.

2.1 Tableau des données et espace associée :

2.1.1 Tableau des données :

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

$$\begin{array}{c} \\ e_1 \\ e_2 \\ . \\ e_i \\ . \\ . \\ e_n \end{array} \begin{array}{c} \begin{array}{ccccccc} X^1 & . & . & X^j & . & . & X^p \end{array} \\ \begin{pmatrix} & & & . & & & \\ & & & . & & & \\ & & & . & & & \\ . & . & . & x_i^j & . & . & . \\ & & & . & & & \\ & & & . & & & \\ & & & . & & & \end{pmatrix} \end{array}$$

Avec :
L'individu e_i **:** élément de R^p.
La variable X^j **:** élément de R^n.
x_i^j **:** est la valeur prise par la variable j sur l'individu i.

En générale on note :

$$X^j = \begin{pmatrix} x_1^j \\ x_2^j \\ . \\ . \\ x_n^j \end{pmatrix}$$ pour la variable et $e_i^t = (x_i^1, x_i^2, x_i^3, ., ., ., x_i^p)$ pour l'individu.

Si $p = 3$ on peut représenter les individus mais lorsque la dimension est plus grande que 3, il est impossible de les visualiser.
Dans ce cas, on cherche une représentation des n individus $e_1, e_2, ., ., e_n$ dans un sous espace F_k de R^p de dimension k et c'est **le principe d'une ACP.**

Autrement dit on cherche à définir k nouvelles variables combinaisons linéaires de p variables initiales et qui nous ferons perdre le moindre
d'information possible.

$$C_l = a_{1l}X^1 + a_{2l}X^2 + a_{3l}X^3 + ... + a_{pl}X^p$$

- Les C_l sont appelées **composantes principales**.
- Les axes qu'elles déterminent sont appelés **axes principaux**.
- Les formes linéaires associées sont appelées **facteurs principaux**.

Remarque :
Les a_{ik} représentent les éléments du vecteur propre associé à C_l.

Matrice poids :

Dans la plupart des cas, les individus jouent le même rôle. Nous nous sommes situés implicitement dans cette situation,en affectant le même poids à chaque individu. Par commodité, on choisit ces poids tels que la masse totale de ces individus soit égale à 1 : à chaque individu on associe alors le poids $\frac{1}{n}$.

Or pour certaines applications on travaille avec des poids p_i éventuellement différents d'un individu à l'autre, ils sont regroupés dans une matrice
diagonale D :

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

Et on a : $D = \frac{1}{n}I$ dans le cas où les poids sont égaux.

Point moyen (centre de gravité) :

Le vecteur g des moyennes arithmétiques de chaque variable
$g^t = (\bar{X}^1, \bar{X}^2, ..., \bar{X}^p)$ [3]

définit le point moyen, ou centre de gravité du nuage.
Avec :

$$\bar{X}^k = \sum_{i=1}^{n} p_i x_i^k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Alors on aura :

$$g = X^t D 1_n$$

avec 1_n un vecteur de R^n dont toutes les composantes sont égales à 1.
Soit Y le tableau des variables aléatoires centrées avec : $y_i^k = x_i^k - \bar{X}^k$
Donc :

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

Matrice de variance-covariance et matrice de corrélation :

D'aprés ce qu'on a vu dans le chapitre précédent, la matrice de variance-covariance d'un tableau centré Y est définie par :

$$V = \frac{1}{n} Y^t Y$$

On peut écrire V sous la forme suivante :

$$\boxed{\text{V=X}^t DX - gg^t = Y^t DY} \quad [3]$$

En effet :

$$\begin{aligned} V &= Y^t DY = (X - 1_n g^t)^t D(X - 1_n g^t) \\ &= (X^t - g1_n^t) D(X - 1_n g^t) \\ &= X^t DX - X^t D1_n g^t - g1_n^t DX + g1_n^t D1_n g^t \\ &= X^t Dx - gg^t - gg^t + gg^t \\ &= X^t DX - gg^t \end{aligned}$$

Cette matrice est celle qu'on utilise pour les calculs numériques.

On note $D_{\frac{1}{s}}$ la matrice diagonale des inverses des écarts types.

$$D_{\frac{1}{s}} = \begin{pmatrix} \frac{1}{s_1} & 0 & . & 0 \\ 0 & \frac{1}{s_2} & . & . \\ . & . & . & . \\ 0 & . & . & \frac{1}{s_p} \end{pmatrix}$$

Ainsi on définit le tableau des données centrées réduites par Z tel que :

$$z^j = \frac{x_i^j - \bar{X}^j}{s_j}$$

donc :

$$Z = Y D_{\frac{1}{s}}$$

Ainsi cette matrice va définir la matrice de corrélation R

$$\boxed{\text{R}=Z^t DZ = D_{\frac{1}{s}} V D_{\frac{1}{s}}}$$

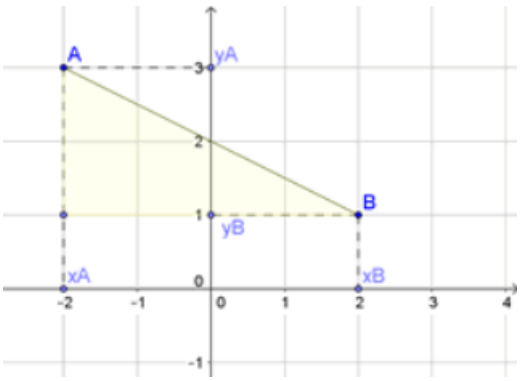
2.1.2 Espace des individus :

Un individu est représenté comme un point de l'espace vectoriel à p dimension noté R^p, dont chaque dimension représente une variable.
L'ensemble des individus constitue un nuage de point dans R^p .

Cet espace est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus.

La métrique :

Comment mesurer la distance entre deux individus ?
Soient $A(x_A, y_A)$ et $B(x_B, y_B)$ deux points dans R^2.



Pour mesurer la distance entre ces deux points on utilise la formule de Pythagore :

$$d(A,B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

En généralisant à p dimension, on a : .
Notons :
$e_i = (x_i^1, x_i^2, x_i^3, ..., x_i^p)$
$e_j = (x_j^1, x_j^2, x_j^3, ..., x_j^p)$

$$\begin{aligned} d^2(e_i, e_j) &= (x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + (x_i^3 - x_j^3)^2 + ... + (x_i^p - x_j^p)^2 \\ &= \sum_{k=1}^{p} (x_i^k - x_j^k)^2 \end{aligned}$$

Remarque :
- Pour avoir la même unité on remplace les données par des données centrées réduites (on travaille avec le tableau Z au lieu de X)
- La formule de Pythagore n'est valable que si les individus sont représentés dans un repère orthonormé or ceci n'est pas vrai en réalité, pour remédier à ce problème on va utiliser une formule générale.

On munit notre espace des individus par un produit scalaire associé à la métrique M qui est une matrice de taille p symétrique et définie positive et on définit la distance entre deux individus par :

$$\begin{aligned} d^2(e_i, e_j) &= (e_i - e_j)^t M (e_i - e_j) \\ &= \langle e_i - e_j, e_i - e_j \rangle_M \end{aligned}$$

En théorie cette métrique dépend de l'utilisateur mais en pratique $M = I$
ce qui revient à utiliser le produit scalaire canonique or dans tous les logiciels la métrique prise par défaut est : $D_{\frac{1}{s^2}}$

$$M = D_{\frac{1}{s^2}} = \begin{pmatrix} \frac{1}{s_1^2} & 0 & . & 0 \\ 0 & \frac{1}{s_2^2} & . & . \\ . & . & . & . \\ 0 & . & . & \frac{1}{s_p^2} \end{pmatrix}$$

Inertie-Inertie totale :

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

$$\begin{aligned} I_g &= \sum_{i=1}^{n} p_i d^2(e_i, g) \\ &= \sum_{i=1}^{n} p_i (e_i - g)^t M (e_i - g) \quad [3] \\ &= \sum_{i=1}^{n} p_i \|e_i - g\|^2 \end{aligned}$$

avec p_i sont les poids des individus.

Remarque :

L'inertie dans un point a est définie par :

$$\begin{aligned} I_a &= \sum_{i=1}^{n} p_i d^2(e_i, a) \\ &= \sum_{i=1}^{n} p_i (e_i - a)^t M (e_i - a) \end{aligned} \quad [3]$$

Si $g = 0$

$$I_g = \sum_{i=1}^{n} p_i e_i' M e_i = trace(MV)$$

En effet : $p_i e_i' M e_i$ étant un scalaire et grâce à la commutativité :

$$\begin{aligned} I_g &= \sum_{i=1}^{n} p_i e_i' M e_i \\ &= \sum_{i=1}^{n} M e_i p_i e_i' \\ &= trace(M X^t D X) \\ &= trace(MV) \end{aligned}$$

Remarque :

- Si $M = I$,on a : $I_g = trace(V)$ avec :

$$V = \begin{pmatrix} S_1^2 & S_{12} & . & . & S_{1p} \\ . & . & . & . & . \\ . & . & . & . & . \\ S_{p1} & . & . & . & S_p^2 \end{pmatrix}$$

donc on a bien : $I_g = \sum_{i=1}^{p} S_i^2$
Si les données sont centrées réduites $I_g = 1 + 1... + 1 = p$
- Si $M = D_{\frac{1}{S^2}}$:

$$\begin{aligned} trace(MV) &= trace(D_{\frac{1}{S^2}} V) \\ &= trace(D_{\frac{1}{S}} V D_{\frac{1}{S}}) \\ &= trace(R) = p \end{aligned}$$

L'inertie est donc égale au nombre de variables et ne dépend pas de leurs valeurs pour des variables centrées réduites.

2.1.3 Espace des variables :

Chaque variable X^j est en fait une liste de n valeurs numériques. Pour étudier l' approximité des variables entre elles il faut munir cet espace d'une métrique c'est-à-dire trouver une matrice d'ordre n définie positive
et symétrique.
Sans hésitation le choix se porte sur la matrice diagonale des poids .
Par conséquent on a :

$$\begin{aligned} \langle X^j, X^k \rangle &= \sum_{i=1}^{n} p_i x_i^j x_i^k \\ &= (X^j)^t D(X^k) \\ &= COV(X^j, X^k) \\ &= S_{jk} \end{aligned}$$

Donc le produit scalaire représente la covariance des variables centrées.
De plus :

$$\|X^j\|_D^2 = S_j^2 \Rightarrow \|X^j\|_D = S_j$$

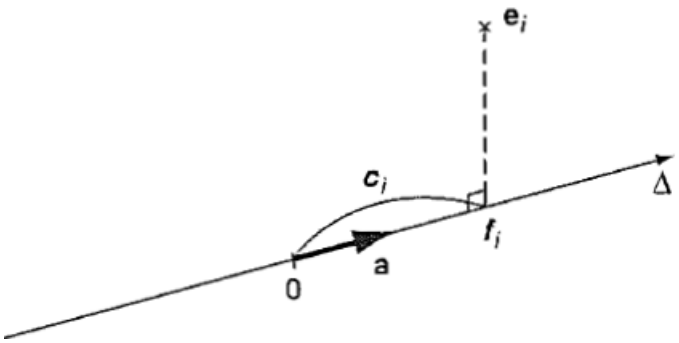
Donc la longueur d'une variable est donnée par son écart type.
L'angle entre deux variables est donné par :

$$\begin{aligned} cos(X^j, X^k) &= cos\theta_{jk} = \frac{\langle X^j, X^k \rangle}{\|X^j\| \|X^k\|} \\ &= \frac{S_{jk}}{S_j S_k} \\ &= \rho_{jk} \end{aligned}$$

On remarque que le coefficient de corrélation linéaire n'est autre que le cosinus d'angle entre les deux variables.

2.2 ACP et éléments principaux :

Considérons un axe ($\triangle$) dans l'espace des individus engendré par un vecteur unitaire $\vec{a}$ et projetons les individus sur cet axe .



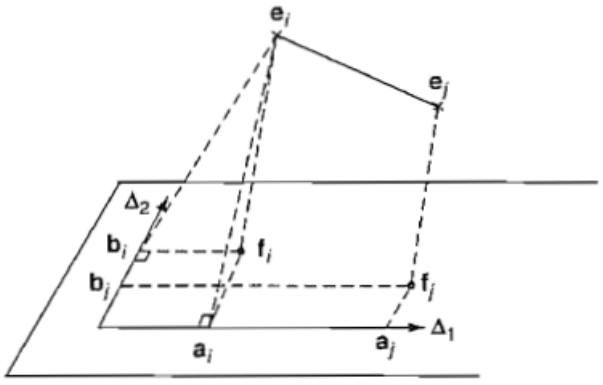
La liste des coordonnées C_i des individus sur ($\triangle$) forment une nouvelle variable.
$C_i = \langle e_i, a \rangle_M = e_i^t Ma = a^t M e_i$, la liste c_1,c_2,.. définit par :

$$XMa = Xu = \sum_{j=1}^{p} X^j u_j$$

En posant $u = Ma$

2.2.1 Méthode de l'ACP :

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



Le critère de choix de l'espace de projection s'effectue tel que l'inertie totale de nuage projeté sur le sous espace F_k soit maximale.
Soit P un projecteur m-orthogonale sur le sous espace F_k, donc le nuage projeté associé au tableau sera donnée par : XP^t
La matrice de variance de tableau XP^t pour les variables centrées sera donnée par :

$$\begin{aligned} (XP^t)^t D(XP^t) &= PX^t DXP^t \\ &= PVP^t \end{aligned}$$

Le problème est de trouver un projecteur qui va maximiser l'inertie totale du nuage projeté donc la trace de la variance.
En effet :

$$\begin{aligned} \text{inertie totale} &= trace(PVP^tM) \\ &= trace(PVMP) \\ &= trace(VMP^2) \\ &= trace(VMP) \end{aligned}$$

Donc on cherche un projecteur P M-orthogonale de rang k maximisant la trace(VMP) ce qui détermine notre sous espace F_k.

2.2.2 Eléments principaux :

Axes principaux :

On cherche la droite de R^p passant par g maximisant l'inertie de nuage projeté sur elle.
Soit a un vecteur, le projecteur M-orthogonale sur la droite portant a est définie par :

$$P = a(a^tMa)^{-1}a^tM$$

En remplaçant le projecteur P par sa formule dans la définition de l'inertie totale, on obtient :

$$\begin{aligned} trace(VMP) &= trace(VMa(a^tMa)^{-1}a^tM) \\ &= \tfrac{1}{a^tMa}trace(VMaa^tM) \\ &= \tfrac{1}{a^tMa}trace(a^tMVMa) \\ &= \tfrac{a^tMVMa}{a^tMa} \end{aligned}$$

La matrice MVM est appelée matrice d'inertie du nuage, elle définit la forme quadratique de l'inertie qui à tous vecteur a M-1 normé(i.e a est M-normé de 1) associé l'inertie projeté sur l'axe définie par a.
Cette matrice se confond avec la matrice de variance covariance
si et seulement si $M = I$
Pour obtenir le maximum il suffit donc que la dérivée de la trace par rapport à a s'annule

$$\begin{aligned} \tfrac{d}{da}(trace(VMP)) &= \tfrac{d}{da}(\tfrac{a^tMVMa}{a^tMa}) \\ &= \tfrac{(2MVMa)(a^tMa)-(2Ma)(a^tMVMa)}{(a^tMa)^2} \end{aligned}$$

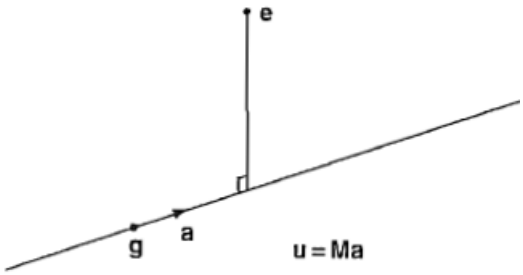
donc

$$MVMa = (\frac{a^tMVMa}{a^tMa})Ma$$

On pose $\lambda = \frac{a^t MVMa}{a^t Ma}$
En prend $VMa = \lambda a$ a est un vecteur propre de VM avec M matrice régulière et λ est la plus grande valeur propre associé à la matrice VM.
- Le premier axe est celui qui aura la plus grande valeur propre λ_1
- Le deuxième axe sera celui de la deuxième valeur propre et ainsi de suite .

Facteurs principaux :

Â l'axe a est associée la forme linéaire $u = Ma$ avec $u \in R^{p*}$(dual deR^p)



u est dit **facteur principale**
Puisque a est vecteur propre de VM :

$$VMa = \lambda a \Rightarrow MVMa = \lambda Ma$$

donc :

$$\boxed{\text{MVu}=\lambda u}$$

Donc on remarque que les facteurs principaux sont les vecteurs propres de la matrice MV. De plus ils sont M^{-1}-normé et de norme 1 ($u^t M^{-1} u = 1$)

Composantes principales :

Les composantes principales sont les variables C_i définies par les facteurs principaux

$$C_l = Xu_l$$

$$C_1 = u_1^1 X^1 + u_2^1 X^2 + u_3^1 X^3 + ... + u_p^1 X^p$$

C_1 est le vecteur renfermant les cordonnées des projections des individus
sur l'axe 1.
De plus : $VAR(C_l) = \lambda_l$

En effet :

$$\begin{aligned} VAR(C) &= C^t DC \\ &= u^t X^t DXu \\ &= u^t Vu \\ &= u^t \lambda M^{-1} u \\ &= \lambda \underbrace{u^t M^{-1} u}_{norme1} \\ &= \lambda.1 = \lambda \end{aligned}$$

Donc la variance de la composante principal est égale à l'inertie apportée par l'axe principale qui lui associé.
Résumé :

1. Axes principaux a : $VMa = \lambda a$ (M-orthogonale)
2. Facteur principaux u : $MVu = \lambda u$ (M^{-1}-orthogonale)
3. Composantes principales C : $C = Xu$ ou $u = Ma$

Remarque :
Les composantes principales C sont les vecteurs propres de la matrice XMX^tD.
En effet :
$MVu = \lambda u \Rightarrow MX^tDXu = \lambda u$
En multipliant les deux cotés par X on obtient :

$$XMX^tDXu = \lambda Xu \Rightarrow XMX^tDC = \lambda C$$

En pratique ,on calcule u en diagonalisant la matrice MV puis on détermine les composantes principales C avec $C = Xu$
Le choix de la métrique est toujours délicat en générale : $M = D_{\frac{1}{S^2}}$
or en pratique on va travailler avec un tableau centré réduit Z ce qui implique que $M = I$ et donc la matrice de variance covariance ne sera d'autre que la matrice de corrélation R.
Les facteurs principaux seront tous simplement rangés selon l'ordre décroissant des valeurs propres.
Ainsi la première composante principale $C = Zu$ sera une combinaison linéaire des centrés réduites ayant une variance maximale.

2.3 Interprétation et qualité de représentation :

Les axes factoriels fournissent des images approchées d'un nuage de points. Il est donc nécessaire de mesurer la qualité de l'approximation, tant pour chacun des points que pour l'ensemble du nuage.

L' ACP construit de nouvelles variables dites artificielles, et des représentations graphiques permettant de visualiser les relations entre les variables, ainsi que l'existence d'éventuels groupes d'individus et ceux de variables.

2.3.1 Qualité des représentations sur les plans principaux :

Dans une ACP on cherche une représentation des individus dans un espace de dimension réduite et qui nous fera perdre le moindre d'information possible.
le critère du pourcentage d'inertie totale expliquée permet de déterminer nombre d'axes retenus. En effet, on calcule :

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

Si par exemple $\frac{\lambda_1+\lambda_2}{I_g} = 0.9$, le nuage de point sera représentés dans le plan des deux premiers axes principaux.

2.3.2 Contribution apportée par les individus :

Si C_1 est trés corrélée avec X^j cela veut dire que les individus ayant une forte coordonnée positive sur le premier axe sont caractérisés par une valeur de X^j nettement supérieur à la moyenne donc il est très utile de calculer pour chaque axe la contribution apportée par les divers individus .

Considérons la l^{eme} composante pour le i^{eme} individu notée c_{li} on aura :

$$\sum_{i=1}^{n} P_i c_{li}^2 = \lambda_l$$

La contribution de chaque individu i à la composante c_l
sera définie par : $CTR(i) = \frac{P_i c_{li}^2}{\lambda_l}$ [3]
.
L'individu qui contribue le plus à la formation d'un axe sera l'individu ou les individus qui auront les plus fortes coordonnées Ainsi on définit la variable $cos^2\theta$ qui est le cosinus de l'angle formé entre l'axe principale et le vecteur reliant le centre de gravité au point, il nous donne la qualité de représentation du point par rapport à l'axe .
On note par QLT la qualité de représentation de l'individu sur le plan

$$QLT = cos^2\theta_1 + cos^2\theta_2 \quad [3]$$

Alors sur un plan un individu sera bien représenté si QLT sera forte .

2.3.3 Choix de la dimension :

Le principal intérêt de l'ACP consiste à réduire la dimension de l'espace des individus, on cherche alors la dimension de notre nouvel espace d'individu et pour cela on a les deux critères suivants :

1-Critère théorique :
Ils consiste à faire un test de significativité sur les valeurs propres c'est à dire les valeurs propres sont elles significativement différentes entre elles à partir d'un certain rang ?

On fait l'hypothèse que les n individus proviennent d'un tirage aléatoire dans une population gaussienne. On a alors :
$H_0 : \lambda_{k+1} = \lambda_{k+2} = ... = \lambda_p$
Sous H_0, on admet que :

$$c = (n - \frac{2p+11}{6})(p-k)ln(\frac{a}{g}) \quad [3]$$

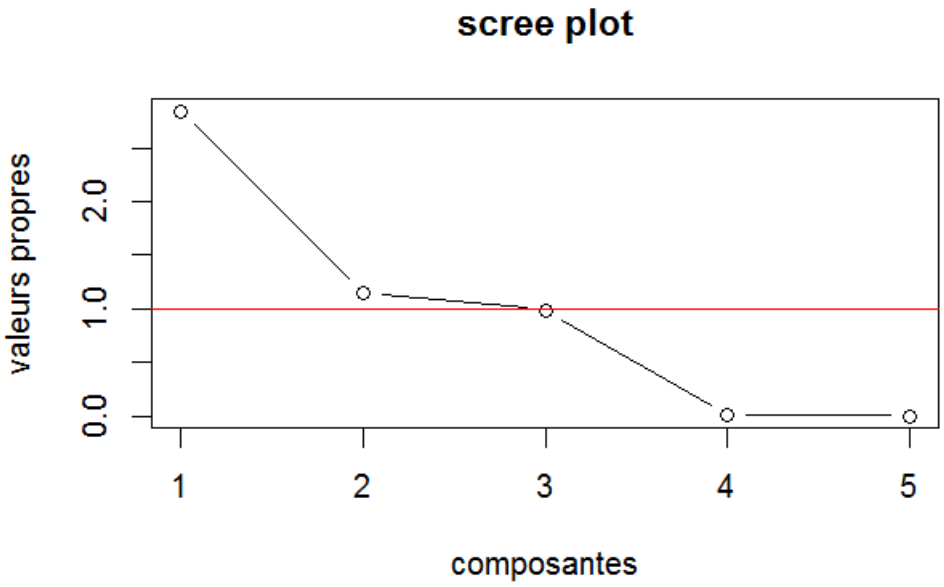
suit un loi de χ^2 de degré de liberté $\frac{(p-k+2)(p-k-1)}{2}$.
où a représente la moyenne arithmétique des k-p valeurs propres et g leur moyenne géométrique.

Si c est trés grand on rejetera l'hypothèse d'égalité des k-p valeurs propres et si H_0 est vrai on garde les premières valeurs propres.

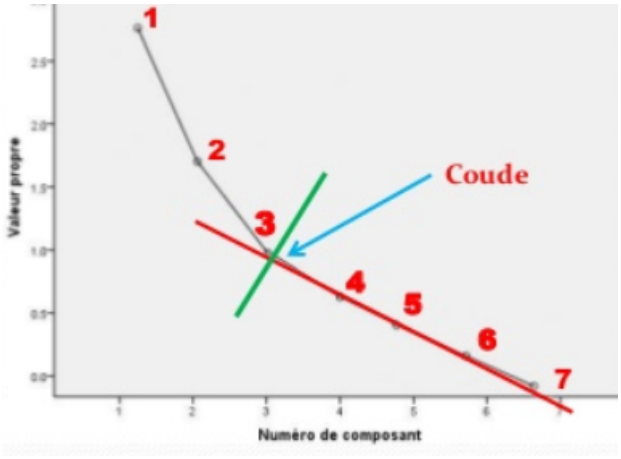
Remarque :
Ce critère n'est applicable que sur des matrices de corrélation dans le cas gaussien, en pratique il ne pourrait être utilisé qu'à titre indicatif.

2-Critères empiriques :
Ce sont les critères qu'on applique en réalité et les plus connus sont ceux de Kaiser et de Cattel.

Critère de Kaiser :
on ne retient que les axes dont l'inertie(la valeur propre) est supérieure à l'inertie moyenne $\frac{I_g}{p}$. Dans le cas d'une ACP normée on ne retiendra que les axes associés à des valeurs propres supérieures à 1($\frac{I_g}{p} = 1$).



Critère de Cattel(critère du coude) :
Il est basé sur le graphique scree plot qui représente en abscisse les composantes et en ordonnée les valeurs propres. Sur ce praphique des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement.



En pratique, on calcule : $\epsilon_1 = \lambda_1 - \lambda_2$,$\epsilon_2 = \lambda_2 - \lambda_3$....
Ensuite on détermine les différences secondes $\delta_1 = \epsilon_1 - \epsilon_2$,$\delta_2 = \epsilon_2 - \epsilon_3$....
et on retient les valeurs propres $\lambda_1, ...\lambda_K$ telles que les différences secondes soient positives.

2.3.4 Interprétation interne :

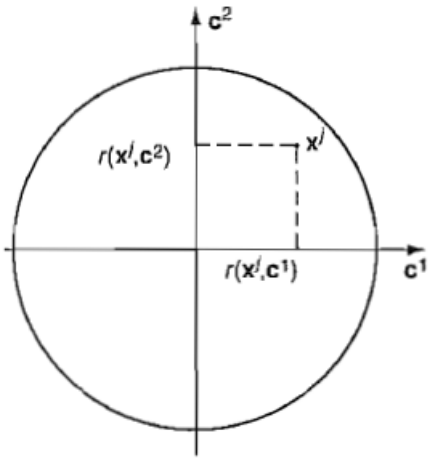
Pour donner une signification à la composante principale il faut la relier aux variables initiales X^j ,en calculant le coefficient de corrélation $r(C, X^j)$ est on s'intéresse au plus fort coefficient en valeur absolue.
Or $r(C, X^j) = r(C, Z^j) = \frac{COV(C,Z^j)}{S_C S_{Z^j}}$ avec Z^j sont centrés réduites.
Or $Var(C) = \lambda \Rightarrow S_C = \sqrt{\lambda}$
Donc on a bien :

$$r(C, X^j) = r(C, Z^j) = \frac{C^t D Z^j}{\sqrt{\lambda}}$$

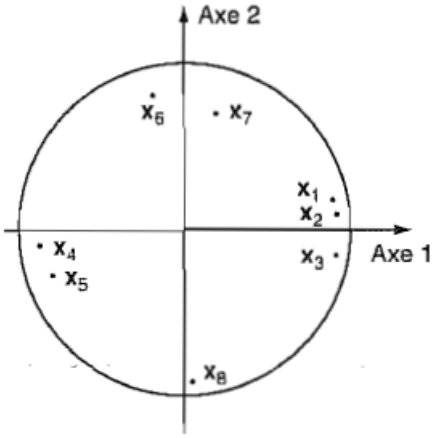
Or $C = Zu$ avec u le facteur principal associé à C et vecteur propre de R matrice de corrélation associé à la valeur propre λ
Donc :

$$\begin{aligned} r(C, Z^j) &= \tfrac{1}{\sqrt{\lambda}} C^t D Z^j \\ &= \tfrac{1}{\sqrt{\lambda}} u^t Z^t D Z^j \\ &= \tfrac{1}{\sqrt{\lambda}} (Z^j)^t D Z^j u \\ &= \tfrac{1}{\sqrt{\lambda}} R u \\ &= \tfrac{1}{\sqrt{\lambda}} \lambda u_j \\ &= \sqrt{\lambda} u_j \end{aligned}$$

avec $(Z^j)^t D Z^j u$ est la j^{eme} composante de Ru.
Ces calculs s'effectuent pour chaque composante principale et pour un couple de composantes principales C_1 et C_2 on représente ces corrélations sur un cercle appelé **cercle de corrélation**.



Chaque variable X^j est représenté par l'abscisse $r(X^j, C^1)$
et l'ordonné $r(X^j, C^2)$



On remarque que X_1, X_2, X_3est corrélé positivement avec C^1 , X_4, X_5 anticorrélé cet axe et X_6, X_7, X_8non corrélé avec C^1.

Remarque :

Si toute les variables sont corrélées positivement entre elles alors la première composante principale définit un facteur de taille.

2.4 Exemple :

Ci-dessous, un tableau de notes attribuées à 9 sujets dans 5 matières.

Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

On applique une ACP sur les donnée du tableau à l'aide du logiciel R .

1- On fait entrer les données da la façon suivante :

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

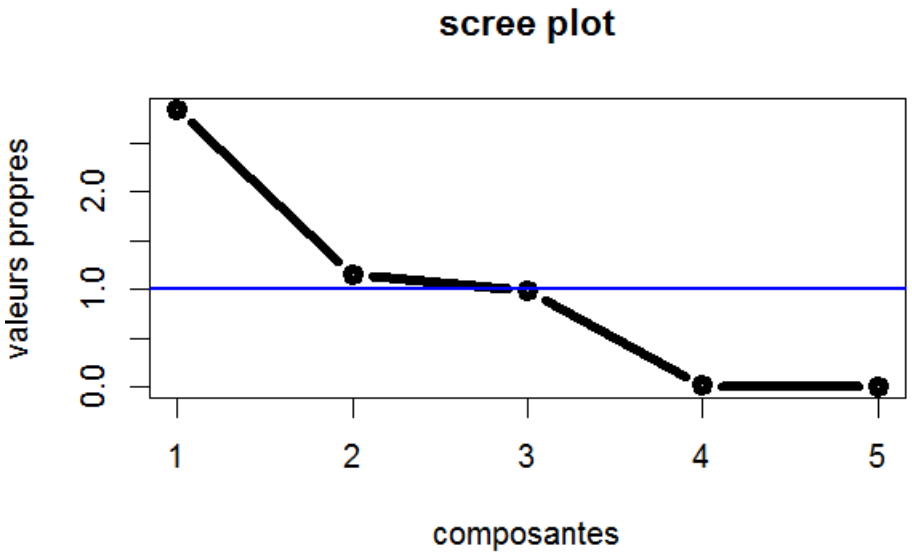
> latin<-c(5.5,9,9.5,15,12,7,11.5,9.5,12)
> science<-c(6,8,7,14.5,14,10,7,12.5,9.5)
> français<-c(5,8,11,15.5,12,5.5,14,8.5,12.5)
> musique<-c(8,9,11,8,10,13,10,12,18)
> donnée<-data.frame(math,science,français,latin,musique)
> donnée

2- On applique notre ACP en utilisant le package FactoMineR :
library(FactoMineR) # pour charger le package
donnée.acp<-PCA(donnée) # les variables sont centré réduite par défaut et les données graphiques sont affichées directement.

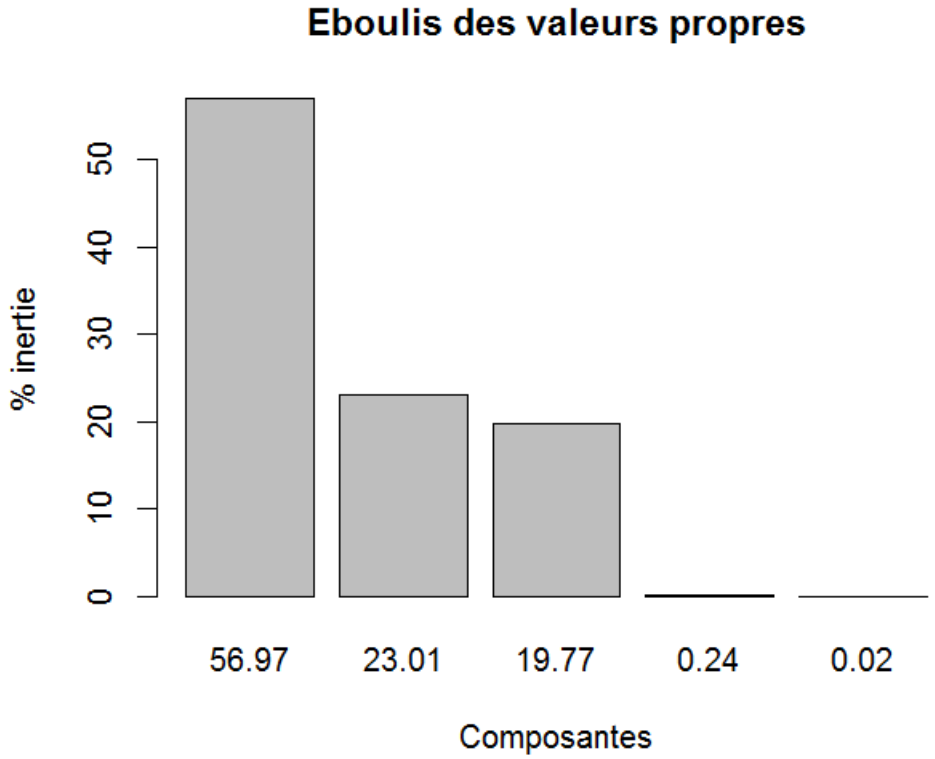
3-On calcule les valeurs propres par la commande **donnée.acp$eig**

	Valeurs propres	Pourcentage %	Pourcentage cumulé
1	2 ,8485	56,97	56,97
2	1 ,1503	23,00	79,97
3	0,9886	19,77	99,74
4	0,0118	0,24	99,98
5	0,0008	0,02	100

4-On fait un "scree plot" qui est le graphique des éboulis des valeurs propres de la façon suivates :
plot(donnée.acp$eig[,1],type="b",ylab="valeurs propres",xlab="composantes"
,lwd=5,main="scree plot")
abline(h=1,col="blue",lwd=2)



barplot(donnée.acp$eig[,2],ylab="% inertie",xlab="Composantes",
names.arg=round(donnée.acp$eig[,2],2))
title("Eboulis des valeurs propres")



L'inertie totale est répartie selon 5 valeurs propres, on ne va considérer que deux composantes car l'inertie cumulé sera 79.97
Avec le FactoMineR, on obtient tous les tableaux de résultats sur les variables en tapant la commande **donnée.acp$var** et qui se résume dans le tableau suivant :

var	QLT	F1	CO2	CTR	F2	CO2	CTR
math	0.98	0.81	0.65	22.78	-0.58	0.33	29.09
science	0.99	0.9	0.8	28.20	-O.43	0.19	16.45
français	0.96	0.76	0.57	20.16	0.62	0.39	33.08
latin	0.98	0.91	0.81	28.73	0.41	0.17	14.39
musique	0.08	0.05	0.01	0.10	0.27	0.07	6.18

De même pour les individus en utilisant la commande
donnée.acp$ind on a le tableau suivant :

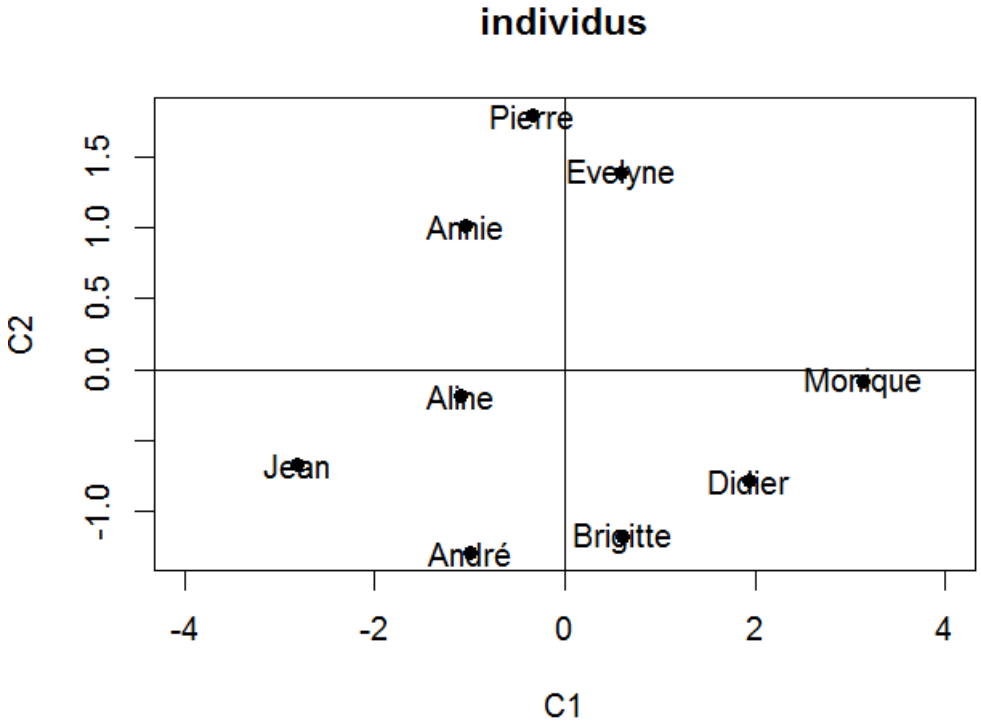
ind	QLT	F1	CO2	CTR	F2	CO2	CTR
Jean	0.93	-2.82	0.88	31.04	-0.67	0.05	4.32
Aline	0.74	-1.09	0.72	4.64	-0.18	0.02	0.32
Annie	0.95	-1.04	0.5	4.23	1.02	0.47	9.95
Monique	0.88	3.13	0.88	38.23	-0.03	0.00	0.08
Didier	0.99	1 .94	0.85	14.77	-0.78	0.14	5.87
André	0.70	-0.99	0.26	3.8	-1.29	0.44	16.18
Pierre	0.84	-0.34	0.03	0.45	1.79	0.81	31.07
Brigitte	0.81	0.61	0.17	1.46	-1.18	0.64	13.55
Evelyne	0.34	0.59	0.05	1.37	1.39	0.29	18.65

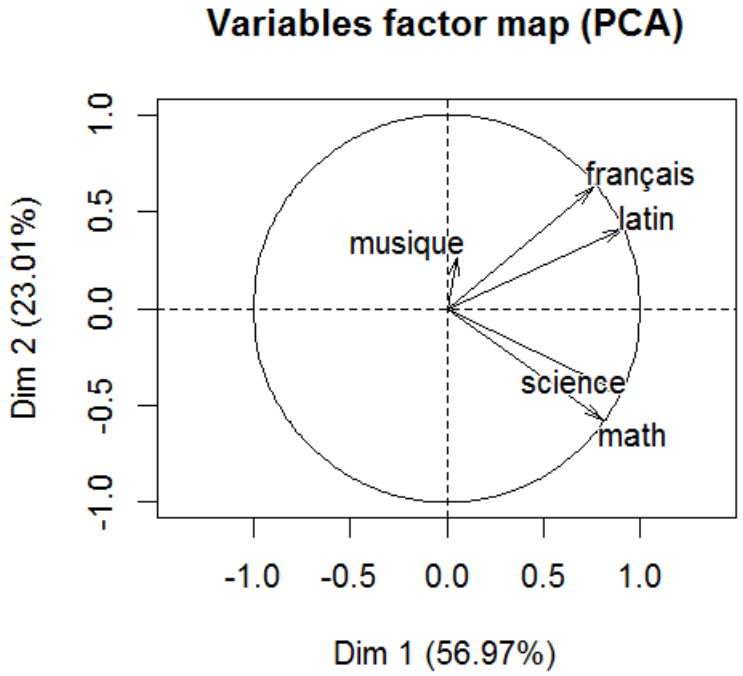
On remarque que "Didier" est bien représenté sur le plan avec une qualité de représentation égale à :

$$QLT(Didier) = 0.85 + 0.14 = 0.99$$

par contre "Evelyne" est très mal représenté ($QLT(Evelyne) = 0.34$).

Le package FactoMineR nous fournit directement la représentation des variables et des individus sur le plan factoriel en tapant **donnée.acp<-PCA(donnée)**.





Interprétation des résultats :

Individus :

axe2 :

On remarque que Monique et Didier sont en opposition avec Aline et Jean. En effet, si on observe dans le tableau des données les notes de ces élèves on trouve que Monique et Didier ont eu les meilleurs notes tandis que Jean et Aline ont eu des résultats très faibles.

axe1 :

André et Brigitte sont en opposition avec Evelyne et Pierre. En effet les deux étudiants André et Brigitte ont presque le même niveau, ils maitrisent les matières scientifique plus que les autre matières ce qui n'est pas vrai pour Evelyne et Pierre qui travaillent beaucoup plus dans les matières de littératures.

Pour les deux axes, Annie est toujours en opposition avec les bons élèves et d'après le graphe on remarque aussi qu'il est littéraire plus que scientifique (il est proche de Evelyne et Pière).

Variables :

axe1 :

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

axe2 :

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

Chapitre 3

Analyse factorielle des correspondances

À l'origine, l'analyse factorielle des correspondances (AFC) a été conçue pour étudier des tableaux appelés couramment tableaux de contingence.
Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus et il permet d'exprimer la liaison entre ces deux variables .

3.1 Tableau de contingence et nuages associés :

Soient deux variables qualitatives X et Y, comportant respectivement p et q modalités. On observe les valeurs de ces variables sur une population et on dispose d'un tableau de contingence N à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison de la modalité i de X et de la modalité j de Y.

y / x	y_1	....	y_j	....	y_q	
x_1	n_{11}	....	n_{1j}	....	n_{1q}	$n_{1.}$
....	....	....	....	....	....	....
x_i	n_{i1}	....	n_{ij}	....	n_{iq}	$n_{i.}$
....	....	....	....	....	....	....
x_p	n_{p1}	....	n_{pj}	....	n_{pq}	$n_{p.}$
	$n_{.1}$	....	$n_{.j}$	....	$n_{.q}$	n

Soit N la matrice des effectifs :

$$N = \begin{pmatrix} n_{11} & n_{12} & . & . & n_{1q} \\ n_{21} & . & . & . & n_{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ n_{p1} & n_{p2} & . & . & n_{pq} \end{pmatrix}$$

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement marges en lignes et marges en colonnes et ils sont calculées comme suit :

$$n_{i.} = \sum_j n_{ij}$$

$$n_{.j} = \sum_i n_{ij}$$

3.1.1 Représentation des profiles associés à un tableau de contingence :

Tableau des profils-lignes :

On appelle tableau des profils-lignes le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{i.}}$.
Ainsi ce tableau est définie par : $D_1^{-1}N$ [3]
avec :

$$N = \begin{pmatrix} n_{11} & n_{12} & . & . & n_{1q} \\ n_{21} & . & . & . & n_{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ n_{p1} & n_{p2} & . & . & n_{pq} \end{pmatrix} \quad et \quad D_1 = \begin{pmatrix} n_{1.} & . & . & . & 0 \\ 0 & n_{2.} & . & . & 0 \\ . & . & n_{3.} & . & 0 \\ . & . & . & . & . \\ 0 & . & . & . & n_{p.} \end{pmatrix}$$

Tableau des profils-colonnes :

On appelle tableau des profils-colonnes le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{.j}}$.
Ainsi ce tableau est définie par : ND_2^{-1} [3]
avec :

$$D_2 = \begin{pmatrix} n_{.1} & . & . & . & 0 \\ 0 & n_{.2} & . & . & 0 \\ . & . & n_{.3} & . & 0 \\ . & . & . & . & . \\ 0 & . & . & . & n_{.p} \end{pmatrix}$$

On appelle profil marginale ligne (resp profil marginale colonne) les quantités $\frac{n_{i.}}{n}$ (resp $\frac{n_{.j}}{n}$).
l'écriture matricielle sera alors : $\frac{D_1}{n}$ (resp $\frac{D_2}{n}$).

Centre de gravité des deux profils :

les profils lignes forment un nuage de **p** points dans R^q. Le centre de gravité de ce nuage est donné par :

$$g_l = \frac{1}{n}(D_1^{-1}N)'D_1 1 = \begin{pmatrix} \frac{n_{.1}}{n} \\ \frac{n_{.2}}{n} \\ . \\ . \\ \frac{n_{.q}}{n} \end{pmatrix} = \begin{pmatrix} p_{.1} \\ p_{.2} \\ . \\ . \\ p_{.q} \end{pmatrix}$$

Ainsi le centre de gravité de nuage des profils colonnes est définie par :

$$g_c = \frac{1}{n}(D_2^{-1}N')'D_2 1 = \begin{pmatrix} \frac{n_{1.}}{n} \\ \frac{n_{2.}}{n} \\ . \\ . \\ \frac{n_{p.}}{n} \end{pmatrix} = \begin{pmatrix} p_{1.} \\ p_{2.} \\ . \\ . \\ p_{p.} \end{pmatrix}$$

3.1.2 La métrique de χ^2 :

Comment mesurer la dispersion de ces nuages de profils ?
Autrement dit, quelle métrique choisir dans chacun des espaces pour obtenir une bonne analyse ?

La distance entre deux profils lignes :

Pour calculer la distance entre deux profils lignes i et i' on utilise
la formule suivante :

$$\boxed{\mathrm{d}^2_{\chi^2}(i,i') = \sum_{j=1}^{q} \frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2} \quad [3]$$

Il s'agit donc de la métrique diagonale nD_2^{-1}

La distance entre deux profils colonnes :

Par analogie, on définit la distance entre deux profils colonnes j et j' par :

$$\boxed{\mathrm{d}^2_{\chi^2}(j,j') = \sum_{i=1}^{p} \frac{n}{n_{i.}}\left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2} \quad [3]$$

Ici on a utilisé la matrice nD_1^{-1}

L'inertie totale :

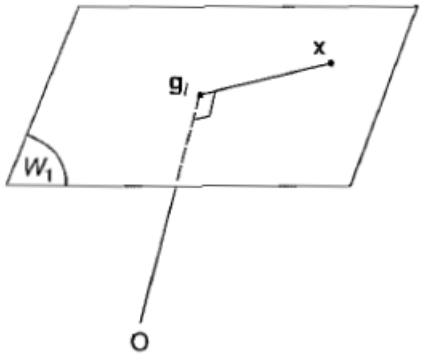
L'inertie totale du nuage de point est donnée par la formule suivante :

$$\boxed{\phi^2 = \frac{1}{n}\sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}} \quad [3]$$

on a aussi :

$$\phi^2 = \sum_i \frac{n_{i.}}{n} d^2_{\chi^2}(i, g_l) = \sum_i \frac{n_{i.}}{n} \sum_j \frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right)^2 \quad [3]$$

Remarque :
le nuage des points profils lignes est dans un sous espace appelé W_1 .
Le vecteur Og_l est orthogonal à ce sous espace au sens de la métrique du χ^2



En effet soit x un élément de W_1 :

$$(x - g_l)' n D_2^{-1} g_l = \langle Og_l, g_l x\rangle_{\chi^2} = 0$$

car

$$nD_2^{-1} g_l = (1,1,1,1,..,1)'$$

Et pour tout élément de W_1 on a : $x'1 = 1$ donc $g_l'1 = 1$
Alors $||g_l||^2_{\chi^2} = g_l'1 = 1$

3.2 La liaison entre deux variables qualitatives :

Lorsque tous les profils-lignes sont identiques c'est à dire $\forall j$ on a :

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = ... = \frac{n_{pj}}{n_{p.}}$$

on peut parler d'indépendance entre les deux variables X et Y.
On remarque aussi que :

$$\sum_{i=1}^{p} \frac{n_{ij}}{n_{i.}}(\frac{n_{i.}}{n}) = \frac{n_{j.}}{n}$$

Ce qui entraine que dans le cas d'indépendance on a :

$$n_{ij} = \frac{n_{i.}n_{.j}}{n}$$

On appelle $(\frac{n_{i.}n_{.j}}{n})_{1\leq i\leq p, 1\leq j\leq q}$ le tableau d'indépendance.

3.2.1 Caractère significatif de l'écart à l'indépendance :

Lorsque on étudie un tableau de contingence, c'est-à-dire une population de n individus à travers de deux variables qualitatives, il est classique de mesurer le caractère significatif de la liaison entre ces deux variables à l'aide de la statistique χ^2. Appliquée à un tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques.

Le test de χ^2 permet de s'assurer du caractère significatif de cette liaison.La démarche du test est la suivante :
Soient X et Y deux variables qualitatives :
H_0 **:** X et Y sont indépendantes
H_1 **:** X et Y ne sont pas indépendantes
Sous H_0 la statistique du test est définie par :

$$\chi^2_{calcul} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \quad [3]$$

Ce test a comme région critique au risque $\alpha = 0.05$:

$$D = \{|\chi^2_{calcul}| > \chi^2_{(p-1)(q-1)}\}$$

3.3 Analyse en composantes principales des nuages de points :

L'analyse factorielle des correspondances est définie comme étant le résultat d'une double ACP des deux profils qui sont en dualité.
-Profils lignes :
-Le tableau de données : $X = D_1^{-1}N$.
- La métrique : $m = nD_2^{-1}$
- La matrice de poids : $D = \frac{D_1}{n}$

-Profils colonnes :
-Le tableau de données : $X = D_2^{-1}N^t$.
- La métrique : $m = nD_1^{-1}$

- La matrice de poids : $D = \frac{D_2}{n}$

On a aussi besoin de centre de gravité $g = X^t D1$ et de la matrice de variance covariance :

$$V = X^t DX - gg^t$$

3.3.1 ACP non centrées et facteur trivial :

Le vecteur Og est orthogonal au support de ce nuage donc g est dit
facteur principal ou trivial autrement dit g est un vecteur propre associé à Vm avec $\lambda = 0$

Du point de vue technique, on peut montrer qu'il n'est pas nécessaire de centrer explicitement le nuage de point avant de l'analyser. En effet, mis à part le premier facteur, l'analyse du nuage par rapport à O sans centrage conduit aux mêmes facteurs que l'analyse du nuage centré c'est le but de ce paragraphe.

Les vecteurs propres de Vm sont les même que $X^t DXm$ et avec les même valeurs propres sauf pour g qui aura pour valeur propre $\lambda = 1$
En effet
pour g :

$$\begin{aligned} X^t DXmg &= Vmg + gg^t mg \\ &= 0 + g||g||^2_{\chi^2} \\ &= 0 + g \\ &= g \end{aligned}$$

C e qui implique que $\lambda = 1$
D'autre part on a pour tout vecteur u tel que $u \perp g$

$$\begin{aligned} X^t DXmu &= Vmu + gg^t mu \\ &= \lambda u + g\langle g, u\rangle_{\chi^2} \\ &= \lambda u \end{aligned}$$

On remarque qu'on a les mêmes valeurs propres donc il est inutile de centrer les tableaux des profils, on effectuera donc des ACP non centrées et on éliminera la valeur propre $\lambda = 1$ associée au facteur principale g car elle est maximale.

3.3.2 Calcul de l'ACP non centrées des nuages de points :

L'ACP pour profils lignes :

Facteurs principaux :
Ce sont les vecteurs propres de la matrice mX^tDX et on a :

$$\begin{aligned} mX^tDX &= (nD_2^{-1})(D_1^{-1}N)^t\tfrac{D_1}{n}(D_1^{-1}N) \\ &= D_2^{-1}N^tD_1^{-1}N \end{aligned}$$

Alors pour tout facteur principal u_k on a :

$$D_2^{-1}N^tD_1^{-1}Nu_k = \lambda u_k$$

Composantes principales :
Ce sont les vecteurs propres de la matrice XmX^tD
avec : $XmX^tD = D_1^{-1}ND_2^{-1}N^t$
En effet, on associe au facteur u_k la composante a_k tel que :

$$a_k = Xu_k = D_1^{-1}Nu_k$$

Donc :

$$\begin{aligned} XmX^tDa_k &= D_1^{-1}ND_2^{-1}N^ta_k \\ &= D_1^{-1}N\underbrace{D_2^{-1}N^tD_1^{-1}Nu_k}_{\lambda u_k} \\ &= \lambda_k D_1^{-1}Nu_k \\ &= \lambda_k a_k \end{aligned}$$

Les composantes principales a_k sont normalisées par :

$$\lambda = a^tD_1a = \frac{1}{n}\sum_{i=1}^{p} n_{i.}(a_i)^2$$

L'ACP pour profils colonnes :

Facteurs principaux :
Ce sont les vecteurs propres de la matrice $\mathrm{D}_1^{-1}ND_2^{-1}N^t$
Composantes principales :
Les composantes principales b_k sont les vecteurs propres de la matrice $D_2^{-1}N^tD_1^{-1}N$ et ils sont normalisées par :

$$\lambda = b^tD_2b = \frac{1}{n}\sum_{j=1}^{q} n_{.j}(b_j)^2 \quad [3]$$

.
Remarques :
1-Les composantes principales représentent les cordonnées factorielles des points représentatifs

des profiles lignes ou colonnes donc ces cordonnées
s'obtiennent en cherchant les vecteurs propres des produits des deux tableaux $D_1^{-1}N$ et $D_2^{-1}N^t$ dans un ordre ou l'autre.
2-Le cercle de corrélation n'ayant aucun intérêt ici car les variables sont qualitatives et par conséquent l'interprétation repose sur les contributions c'est le but du paragraphe qui va suivre.

La contribution :
La contribution totale est donnée par :

$$\lambda = \frac{1}{n}\sum_{i=1}^{p} n_{i.}(a_i)^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}(b_j)^2 \quad [3]$$

Ainsi on définit la contribution du profil ligne i à l'inertie par :

$$CTR(i) = \frac{n_{i.}}{n}.\frac{(a_i)^2}{\lambda} \quad [3]$$

La contribution du profil colonnes j est :

$$CTR(j) = \frac{n_{.j}}{n}.\frac{(b_j)^2}{\lambda} \quad [1]$$

En pratique on considère les catégories ayant la plus forte contribution pour cela on a le critère qui consiste à retenir $CTR(i) > \frac{n_{i.}}{n}$. [3]

3.3.3 Les formules de transition :

On cherche une relation entre les vecteurs a_k et b_k afin d'éviter la diagonalisation de deux matrices .
évidement on va choisir la plus petite matrice.
Par exemple on suppose que $p < q$ alors on va diagonaliser
la matrice $D_1^{-1}ND_2^{-1}N^t$ qui possède a comme vecteur propre donc on a :

$$D_1^{-1}ND_2^{-1}N^t a = \lambda a \ (*)$$

On multiplie les deux côtés par $D_2^{-1}N^t$

$$D_2^{-1}N^tD_1^{-1}ND_2^{-1}N^t a = \lambda D_2^{-1}N^t a$$

On pose $b = D_2^{-1}N^t a$ (b vecteur propre de la matrice $D_2^{-1}N^tD_1^{-1}N$)

$$D_2^{-1}N^tD_1^{-1}Nb = \lambda b$$

autrement dit $b = kD_2^{-1}N^t a$. On cherche maintenant à déterminer la valeur de k.
On sait que : $\lambda = b^t\frac{D_2}{n}b$ (condition de normalisation) donc on a bien :

$$\begin{aligned}\lambda &= ka^tND_2^{-1}\tfrac{D_2}{n}kD_2^{-1}N^t a \\ &= \tfrac{k^2}{n}a^tND_2^{-1}N^t a\end{aligned}$$

Or d'aprés (*) on a : $ND_2^{-1}N^ta = \lambda D_1$.
Par conséquent on obtient $\lambda^2 k^2 a^t \frac{D_1}{n} a = \lambda$ avec $\lambda = a^t \frac{D_1}{n} a$.
Ce qui implique que :

$$k^2\lambda = 1 \Rightarrow k = \frac{1}{\sqrt{\lambda}}$$

Donc on a bien :

$$b = \frac{1}{\sqrt{\lambda}} D_2^{-1} N^t a$$

De même pour a :

$$a = \frac{1}{\sqrt{\lambda}} D_1^{-1} N b$$

3.3.4 La décomposition de l'inertie :

On définit l'inertie totale par :

$$\phi^2 = \sum_{k=1}^{n} \lambda_k \quad [3]$$

Avec $n = min(p-1, q-1)$ car il y a au plus $min(p-1, q-1)$ valeurs propres.
Ainsi les pourcentages d'inertie sont donnés par : $\frac{\lambda_k}{\phi^2}$ [3]
.

3.3.5 Formule de reconstitution :

Soient X le tableau des profils lignes.
c_k le vecteur des coordonnées des lignes sur l' axe n$^o k$.
u_k facteur principal identique au vecteur des coordonnées des colonnes sur l'axe k divisé par $\sqrt{\lambda_k}$.
La matrice :$M = nD_2^{-1}$.
On a :

$$Xu_k = c_k$$

En multipliant les deux membres par $u_k' M^{-1}$ et en sommant sur k :

$$X \sum_{k=1} u_k u_k' M^{-1} = \sum_{k=1} c_k u_k' M^{-1}$$

Or $\sum\limits_{k=1} u_k u_k' M^{-1} = I$ car les u_k sont M^{-1}-normés.
Ce qui implique que :

$$X = \sum_{k=1} c_k u_k' M^{-1}$$

Alors on a :

$$\frac{n_{ij}}{n_{i.}} = \sum_{k=1} \frac{a_i^k b_j^k}{\sqrt{\lambda_k}} \frac{n_{.j}}{n}$$

On fait sortir la valeur propre trivial $\lambda = 1$ de la somme :

$$n_{ij} = \frac{n_{i.} n_{.j}}{n} (1 + \sum_{k=1} \frac{a_i^k b_j^k}{\sqrt{\lambda_k}})$$

3.3.6 Le choix du nombre de valeurs propres :

Le choix de la dimension pose les mêmes problèmes qu'en ACP .De nombreuses techniques empiriques ont été proposées (critère de Kaiser,critère du coude).
Il existe également une autre approche qui peut donner des indications intéressantes. Nous la détaillons ci-dessous.
***Le test de Malinvaud :** [3]
Ce test est basé sur la comparaison entre effectifs observés n_{ij} et effectifs calculés à l'aide de la formule de reconstitution .
Il est appliqué pour des effectifs théoriques au moins égaux à 5.
On fait l'hypothèse que les données forment un échantillon tiré aléatoirement et avec équiprobabilité dans une population tel que :

$$p_{ij} = p_{i.}p_{.j}(1 + \sum_{k=1}^{K} \alpha_{ik}\beta_{jk})$$

On définit l'effectif $\hat{n_{ij}}$ par :

$$\hat{n_{ij}}^K = \frac{n_{i.}n_{.j}}{n}(1 + \sum_{k=1}^{K} \frac{a_i^k b_j^k}{\sqrt{\lambda_k}})$$

c'est la reconstitution de la case ij à l'aide des K premiers axes.
On les compare aux n_{ij} avec un test de khi-deux.
Ainsi Malinvaud a suggéré d'utiliser la quantité :

$$Q_K = \sum_{i,j} \frac{(n_{ij} - \hat{n_{ij}}^K)^2}{\frac{n_{i.}n_{.j}}{n}}$$

et après un calcul élémentaire on trouve que cette quantité égale à :

$$n(I - \lambda_1 - ... - \lambda_K) = n(\lambda_{K+1} + ... + \lambda_{K+r})$$

avec $I = \phi^2 + 1$ et $r = min(p-1, q-1)$.
Q_K suit une loi de χ^2 à (p - K -1) (q - K -1) degré de liberté. Il s'agit donc d'une généralisation du test d'écart à l'indépendance qui correspond
au cas $K = 0$.
On fait des tests successifs pour $K = 0$ (test d'indépendance), pour $K = 1$,...etc jusqu'à arriver à une valeur de K' qui est la plus petit valeur pour laquelle Q_K est inférieure à la valeur limite de cette loi c'est à dire :

$$K' = min\{K \geq 1, Q_K < \chi^2_{(p-K-1)(q-K-1)}\}$$

ainsi K' c'est le nombre de valeurs propres à retenir.

Chapitre 4

Application

A fin d'illustrer les méthodes statistiques de l'analyse multivariés, nous allons appliquer ces méthodes sur des données réelles sur un groupe de 150 personnes portant sur l'indemnisation des assurances suites à des accidents corporels. Ces données nous ont été fournies par le médecin **Mr Hamidou Mohammed** que nous remercions fortement.
Dans un premier temps, nous avons recueilli sur léchantillon des 150 cas d'accident les variables suivantes :

*** Type de blessures :**
TCV : traumatisme de crâne et de visage.
FB : fracture de bassin.
FV : fracture de vertèbre.
AT : autres traumatisme.
TMS : traumatisme de membre supérieur.
TMI :traumatisme de membre inférieur.
PM : plaies multiples.
BR : brûlures.

***Age :** enfant, A1830(de 18 à 30 ans),...etc

***Type d'accident :** passager(Pa),piéton(Pt),conducteur(C).

***Taux d'indemnisation :** T020(de 0 à 20 %),T2040(de 20 à 40 %),....etc

***Durée pour guérir :** D0002(de 0 à 2 mois),D0204(de 2 à 4 mois), ...,D1012,plus (plus de 12 mois)

*** Genre :**M (masculin),F (féminin).

*** Période d'accident :** H (hiver),P(printemps),E (été),A (automne).

4.1 L'ACP :

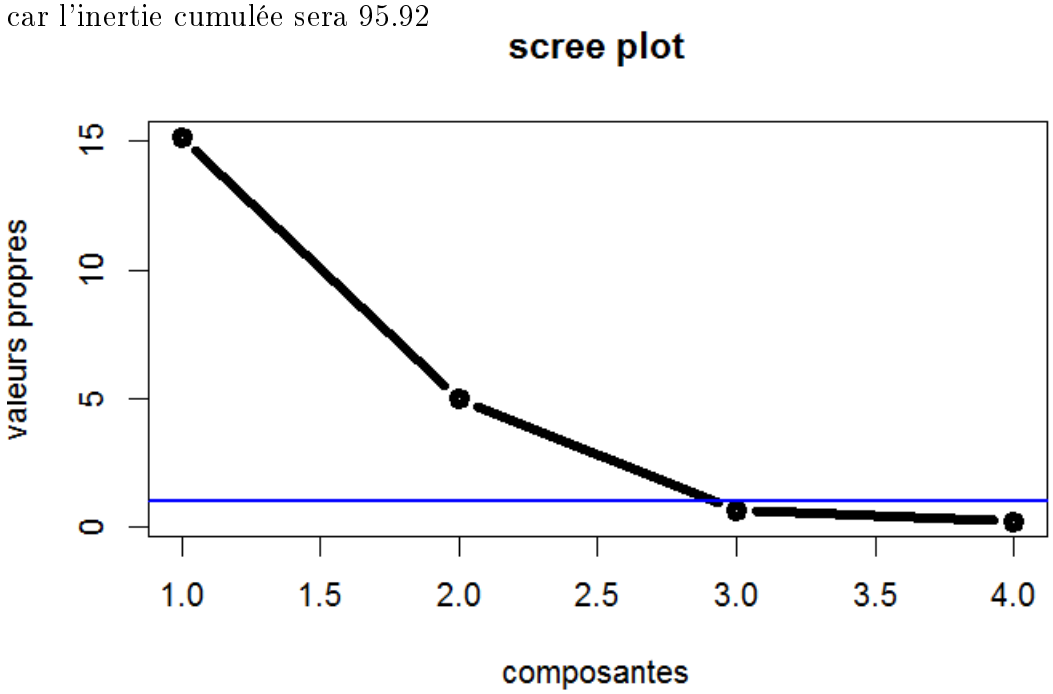
On va considérer un tableau comprenant en lignes les différententes tranches d'âge et en colonnes le reste des variables (types d'accident, durée de guérison, période d'accident, taux d'indemnisation,...)

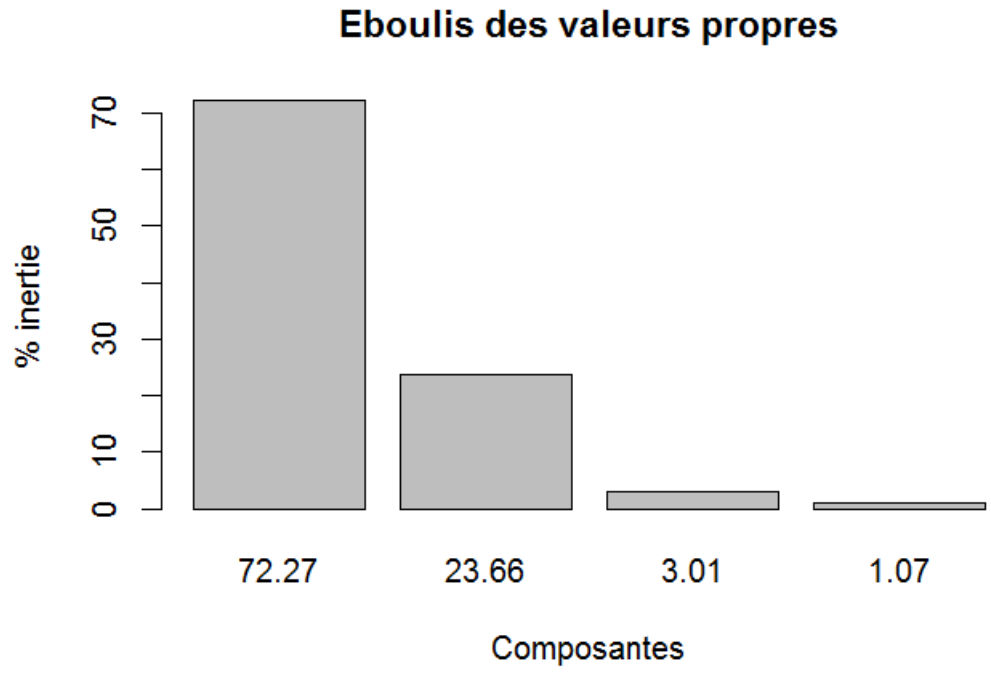
	M	F	T020	T2040	T4060	T6080	T80100	C	Pt	Pa	A	H	P	E	D0002	D0204	D0406	D0608	D0810	D1012	Plus
Enfant	20	16	25	11	1	0	0	0	24	13	2	12	12	11	24	12	1	0	0	0	0
A1830	20	3	11	6	4	1	0	14	1	7	2	6	5	9	13	4	2	0	2	1	0
A3050	43	15	18	16	10	2	1	26	10	21	7	12	16	23	30	9	7	3	4	3	2
A5070	19	8	13	10	2	1	0	11	7	8	3	7	8	9	14	9	2	1	0	0	1
A7090	6	0	3	1	1	1	0	4	0	2	1	3	1	1	3	2	0	0	1	0	0

On obtient les valeurs propres suivantes :

	Valeur propre	Pourcentage	Pourcentage cumulé
1	15.1763829	72.268490	72.26849
2	4.9678345	23.656355	95.92484
3	0.6317538	3.008352	98.93320
4	0.2240288	1.066804	100.00000

L'inertie totale est répartie selon 4 valeurs propres,on ne va prendre que deux composantes car l'inertie cumulée sera 95.92





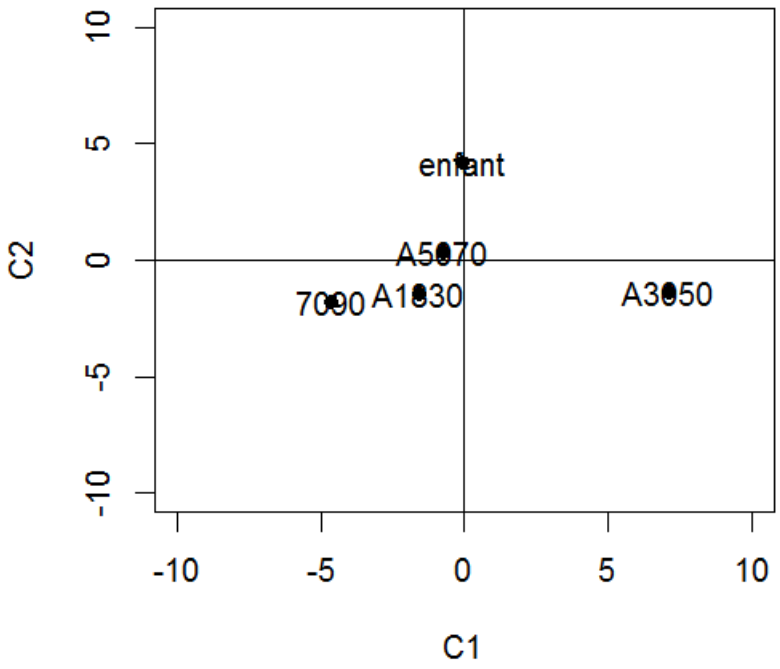
Les données sur les variables :

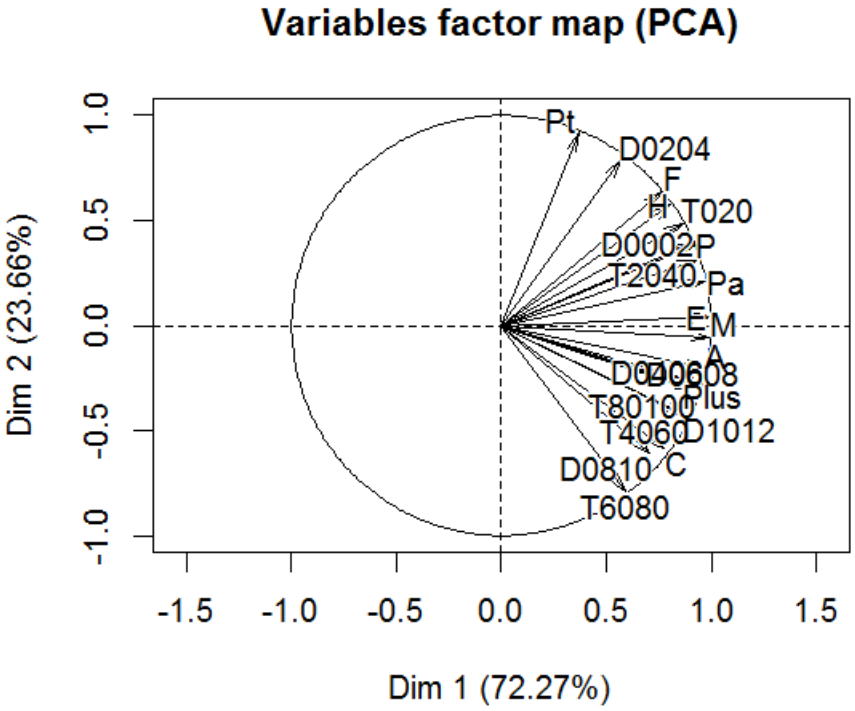
variables	QLT	F1	COS2	CTR	F2	COS2	CTR
M	0 .98	0.9896	0.979	6.453	-0.0548	0.003	0.060
F	0.99	0.7649	0.585	3.855	0.6396	0.409	8.236
T020	0.98	0.8639	0.746	4.918	0.4876	0.237	4.786
T2040	0.96	0.9293	0.863	5.691	0.3125	0.097	1.966
T4060	0.97	0.8900	0.792	5.219	-0.4324	0.187	3.765
T6080	0 .97	0.5875	0.345	2.274	-0.7917	0.626	12.618
T80100	0.93	0.9158	0.838	5.526	-0.3093	0.095	1.926
C	0.94	0.7768	0.603	3.976	-0.5866	0.344	6.927
Pt	0.97	0.3683	0.135	0.894	0.9160	0.839	16.892
Pa	0.98	0.9728	0.946	6.236	0.2109	0.044	0.896
A	0.97	0.9690	0.938	6.187	-0.2035	0.041	0.833
H	0.98	0.8083	0.653	4.305	0.5762	0.332	6.684
P	0.99	0.9230	0.851	5.613	0.3841	0.147	2.970
E	0.98	0.9925	0.985	6.491	0.0412	0.001	0.034
D0002	0.97	0.9102	0.828	5.459	0.3912	0.153	3.081
D0204	0.94	0.5676	0.322	2.123	0.7879	0.620	12.497
D0406	0.98	0.9584	0.918	6.053	-0.2805	0.078	1.584
D0608	0.90	0.9091	0.826	5.446	-0.2900	0.084	1.693
D0810	0.86	0.7085	0.501	3.307	-0.6070	0.368	7.418
D1012	0.93	0.8715	0.759	5.004	-0.4275	0.182	3.679
Plus	0.82	0.8674	0.752	4.957	-0.2681	0.071	1.447

Les données sur les individus :

individus	QLT	F1	COS2	CTR	F2	COS2	CTR
enfant	0 .98	-0.1029	0.0006	0.0139	4.2015	0.986	71.067
A1830	0.77	-1.6095	0.438	3.414	-1.4195	0.341	8.112
A3050	0 .99	7.1356	0.962	67.10	-1.3790	0.035	7.656
A5070	0.23	-0.7546	0.192	0.750	0.3675	0.045	0.543
A7090	0.97	-4.6684	0.859	28.72	-1.7704	0.123	12.619

Le plan factoriel :





Interprétation des résultats :

Les individus :

Les catégories "enfant" et "A5070" caractérisent le deuxième axe on peut dire que cet axe donne le classement des individus selon les durées de guérison ainsi on voit que "A7090" et "A3050" sont en opposition par rapport au premier axe.

Les variables :

Puisque les variables sont toutes proches du cercle de corrélation alors on peut déduire que la qualité de représentation de ces variables par rapport à ces axes est très satisfaisantes ce qui nous montre le tableau des données sur les variables.

On constate que les taux d'indemnisation qui sont plus importants sont en opposition avec ceux qui sont faible et on a la même remarque pour les durées de guérison.

On observe que les conducteurs sont en opposition avec les piétons et les passagers de plus cette catégorie est très proche des taux d'indemnisation très élevés et des durées de guérison importantes .

"Les masculin" sont en opposition avec "les féminin" par rapport au deuxième axe.

On voit que "A" est au voisinage des taux d'indemnisation élevés et des durées de guérison importantes ce qui montre que les accidents les plus graves se sont été passées dans cette saison.

4.2 L'AFC :

les données nous conduisent à appliquer les AFC suivantes :

4.2.1 AFC (taux d'indemnisation vs durée pour guérir) :

On a le tableau de contingence suivant :

Durée / Taux	D0002	D0204	D0406	D0608	D0810	D1012	Plus
T020	68	8	3	0	1	0	0
T2040	14	20	7	1	1	1	1
T4060	2	8	1	1	3	2	1
T6080	0	0	1	0	2	1	1
T80100	0	0	0	1	0	0	0

On fait entrer les données de ce tableau dans le logiciel R de la manière suivante :
taux<-c("T020","T2040","T4060","T6080","T80100")
durée<-c("D0002","D0204","D0406","D0608","D0810","D1012","plus")
données<-matrix(c(68,8,3,0,1,0,0,14,20,7,1,2,1,1,2,8,1,1,3,2,1,0,0,
1,0,2,1,1,0,0,0,1,0,0,0),nrow=5,byrow=TRUE)
dimnames(données)<-list(taux,durée)
données

L'AFC n'a pas d'intérêt que s'il y a une dépendance entre les deux variables donc on va faire un test d'indépendance.

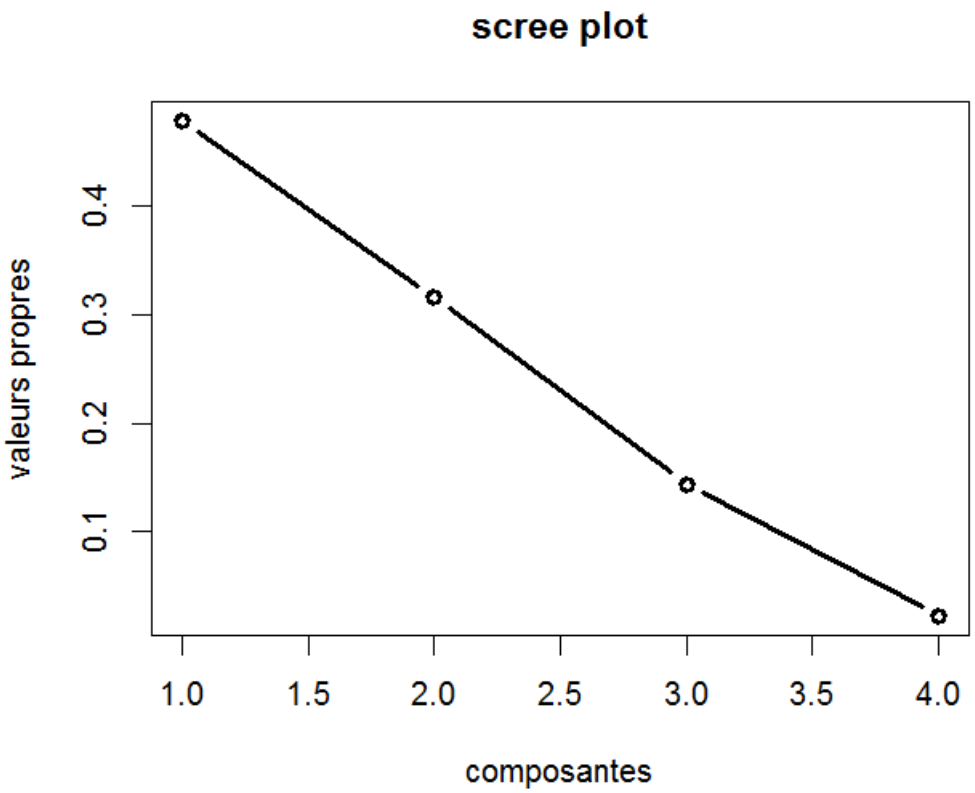
$$\chi^2_{calcul} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Grâce à la commande **chisq.test(données)** du logiciel R ,on peut calculer χ^2_{calcul} où on trouve 144.13.
Or pour un risque $\alpha = 0.05$ on a $\chi^2_{24} = 13.84$.
Donc $\chi^2_{calcul} > \chi^2_{24}$ ce qui implique la dépendance.

Pour appliquer l'AFC on va utiliser la commandes suivante :
library(FactoMineR)
x<-CA(données)
1-Valeurs propres :

	Valeur propre	Pourcentage	Pourcentage cumulé
1	0.478283	49.77505	49.77505
2	0.316385	32.92625	82.70130
3	0.143823	14.96768	97.66899
4	0.022398	2.331011	100.00

Le tableau des valeurs propres montre clairement que deux axes suffisent à décrire la liaison entre la durée pour guérir et le taux d'indemnisation .
Remarque :
Ici on ne peut pas appliquer le test de Malinvaud car n est grand et cela conduit à conserver un grand nombre d'axes.



2-Les résultats pour les profils lignes :

Les cordonnés :

```
$coord
            Dim 1      Dim 2
T020    -0.5634518  0.1027848
T2040    0.3424736 -0.1078476
T4060    1.0092595 -0.2085589
T6080    1.6094379 -1.1461553
T80100   3.1084950  6.2230437
```

La contribution :

```
$contrib
            Dim 1     Dim 2
T020    35.401900  1.780903
T2040    7.520299  1.127384
T4060   25.556495  1.649766
T6080   18.052676 13.840427
T80100  13.468631 81.601520
```

Le cosinus carré :

```
$cos2
            Dim 1      Dim 2
T020    0.9010221 0.02998332
T2040   0.3420052 0.03391562
T4060   0.8660797 0.03698363
T6080   0.4317151 0.21894532
T80100  0.1971988 0.79033209
```

3-Les résultats pour les profils colonnes :

Les cordonnés :

```
$coord
           Dim 1      Dim 2
D0002 -0.5422632  0.1071439
D0204  0.4183626 -0.1483084
D0406  0.4007314 -0.2668671
D0608  2.1497749  3.5003457
D0810  1.1510136 -0.6735552
D1012  1.4352739 -0.7427450
plus   1.4272479 -0.8667321
```

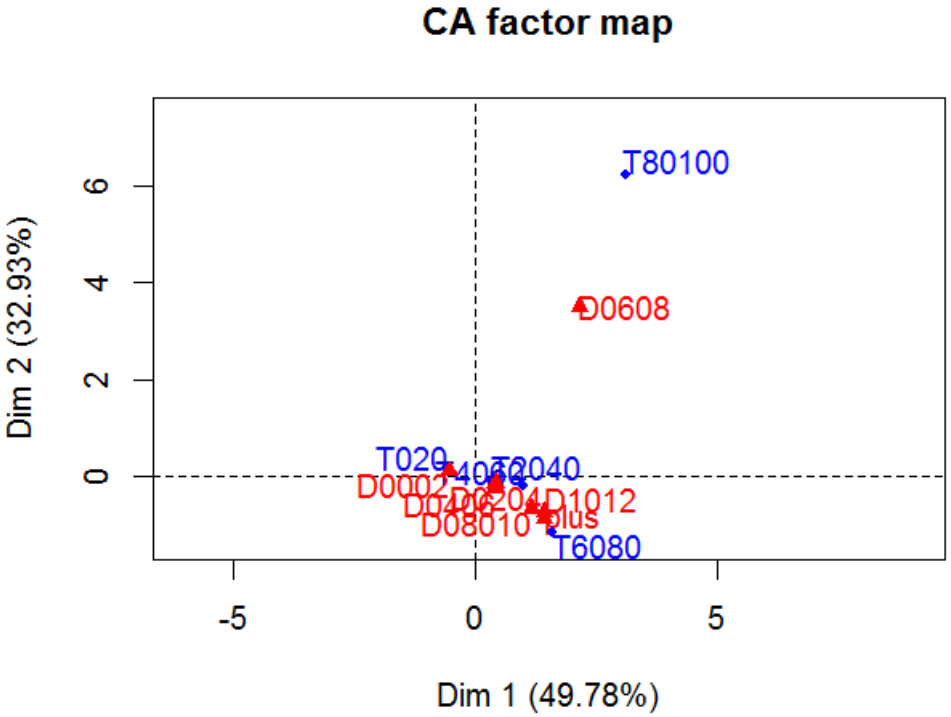
La contribution :

```
$contrib
           Dim 1     Dim 2
D0002 34.428855  2.031921
D0204  8.782767  1.668501
D0406  2.686032  1.800794
D0608 19.325482 77.452545
D0810 14.773181  7.647666
D1012 11.485573  4.649776
plus   8.518109  4.748797
```

Le cosinus carré :

```
$cos2
           Dim 1      Dim 2
D0002 0.9074327 0.03542658
D0204 0.3428167 0.04308120
D0406 0.3257365 0.14446070
D0608 0.2725778 0.72264726
D0810 0.5810729 0.19898304
D1012 0.6514616 0.17446113
plus  0.5624721 0.20743016
```

4-Le plan factoriel :



5-Discussion des résultats :

La variable taux d'indemnisation :

On considère les modalités qui vérifient la règle suivante :$CTR(i) > \frac{n_{i.}}{n}$

Par exemple pour T020 on a $CTR(1) = 35\%$ or $\frac{n_{1.}}{n} = \frac{80}{150}.100 = 53\%$ donc on ne prend pas cette modalité.
Remarque : On a :

$$CTR(i) = \frac{n_{i.}}{n}\frac{a_i^2}{\lambda} > \frac{n_{i.}}{n}$$

Ce qui implique que :

$$\frac{a_i^2}{\lambda} > 1$$

Donc il suffit de prendre les modalités qui vérifient :

$$|a_i| > \sqrt{\lambda}$$

Axe 1 :

-	+
	T4060 T6080 T80100

Axe 2 :

-	+
T6080	T80100

La qualité de représentation de la variable "taux d'indemnisation" sur chaque modalité est mesurée par la formule suivante :

$$cos^2F1 + cos^2F2$$

où F1 c'est le premier axe factoriel et F2 le deuxième axe factoriel.
Si cette quantité est proche de 1 on dit que la modalité est bien représentée .
On conclut que la modalité "T80100" a une très bonne qualité de représentation (0.98) et même pour "T020" et "T4060".
La variable durée pour guérir :
Axe 1 :

-	+
	D0608 D0810 D1012 plus

Axe 2 :

-	+
D0406 D0810 D1012 plus	D0608

On remarque que la modalité "D0608" est bien représenté avec une qualité de représentation 0.99 et même pour D0002.

L'interprétation du plan factoriel :

On voit que les modalités "T6080" et "T80100" définissent le premier axes c'est ce qu'on a vérifié par les calculs .

Ainsi on observe que le taux d'indemnisation qui est entre 80% et 100% est très proche d'une durée entre 6 et 8 mois ce qui montre que pour un taux élevé on a une durée pour guérir très importante.

On remarque aussi que les modalités "D0810","D1012" et "plus" qui sont des durées très longues sont situées au voisinage "T6080" et "T4060" ce qui confirme le résultat précédent .

En conclusion ,on constate qu'il y a presque une relation affine entre le taux d'indemnisation et la durée pour guérir.

Remarque :

On voit que la modalité "T020" malgré qu'elle soit bien représentée elle ne caractérise aucun axe et elle est plus proche des petites durées comme "D0002" et "D0204".

4.2.2 AFC(taux d'indemnisation vs type de blessures) :

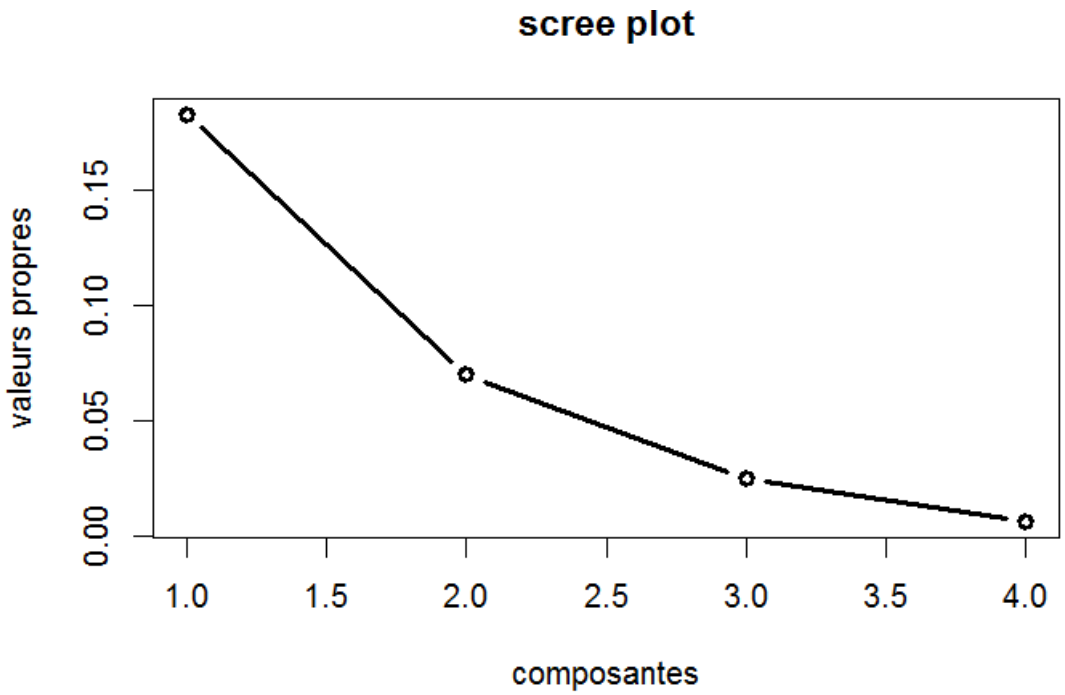
On a le tableau de contingence suivant :

	TMI	TMS	TCR	AT	FB	PM	BR	FV
T020	18	25	25	14	0	12	3	0
T2040	16	10	13	17	2	0	0	2
T4060	9	4	3	6	1	0	0	3
T6080	4	0	0	4	2	0	0	1
T80100	0	0	1	1	0	0	0	0

Le test de chi deux nous donne $\chi^2_{calcul} = 55.658$ or $\chi^2_{28} = 16.928$ ce qui montre qu'il y a une liaison entre les types de blessures et le taux d'indemnisation.

1- Valeurs propres :

	Valeur propre	pourcentage	Pourcentage cumulé
1	0.182375295	64.551545	64.55154
2	0.069567343	24.623288	89.17483
3	0.024632967	8.718813	97.89365
4	0.005951011	2.106354	100.0000



2-Les résultats pour profils lignes :

Les cordonnés :

```
$coord
             Dim 1        Dim 2
T020    -0.4066385 -0.067476670
T2040    0.3152388  0.008437056
T4060    0.3485962  0.528568234
T6080    0.9813158 -0.717071922
T80100   0.1615140 -0.172268754
```

La contribution :

```
$contrib
             Dim 1       Dim 2
T020    44.6433097  3.22260723
T2040   16.5958098  0.03116459
T4060    9.1322295 55.04200402
T6080   29.4834336 41.27114092
T80100   0.1452173  0.43308324
```

Le cosinus carré :

```
$cos2
             Dim 1        Dim 2
T020    0.96552667 0.0265861029
T2040   0.78046422 0.0005590561
T4060   0.28281263 0.6502129740
T6080   0.61512106 0.3284493643
T80100  0.02019199 0.0229705614
```

3-Les résultats pour profils colonnes :

Les cordonnés :

```
$coord
         Dim 1       Dim 2
TMI  0.1847472 -0.11506694
TMS -0.2300674  0.23153865
TCR -0.1878962 -0.03035713
AT   0.3258467 -0.06051678
FB   1.3776731 -0.67388098
PM  -0.9521940 -0.25582968
BR  -0.9521940 -0.25582968
FV   0.7850368  1.21519698
```

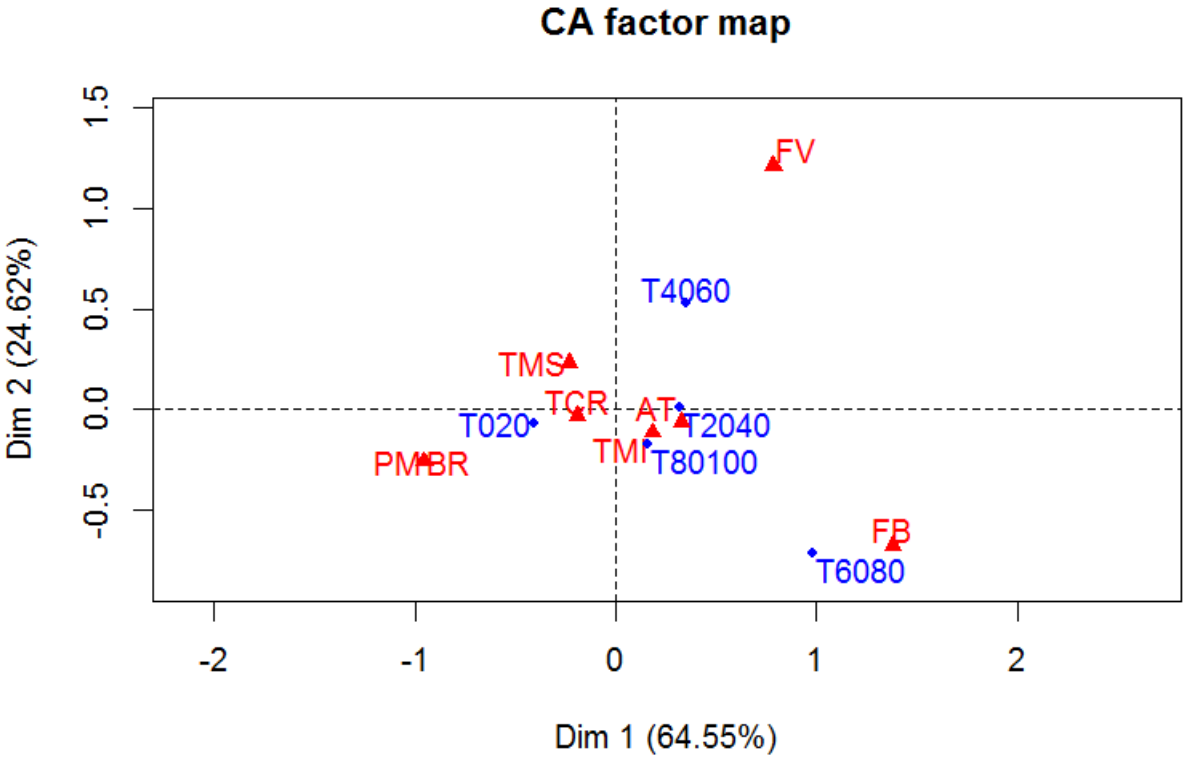
La contribution :

```
$contrib
        Dim 1     Dim 2
TMI  4.084998  4.154301
TMS  6.334995 16.820690
TCR  4.323708  0.295871
AT  12.412055  1.122353
FB  26.413755 16.567797
PM  30.283069  5.730750
BR   7.570767  1.432688
FV   8.576652 53.875551
```

Le cosinus carré :

```
$cos2
         Dim 1      Dim 2
TMI 0.5325363 0.20658306
TMS 0.4538846 0.45970836
TCR 0.5631898 0.01470077
AT  0.8120037 0.02800806
FB  0.7075058 0.16927937
PM  0.8794732 0.06348536
BR  0.8794732 0.06348536
FV  0.2863767 0.68620061
```

4-Le plan factoriel :



5-Discussion des résultats :

La variable taux d'indemnisation :

Axe 1 :

On a $\sqrt{\lambda_1} = 0.427$ donc on va prendre les modalités qui vérifient :

$$|a_i| > 0.427$$

-	+
	T6080 (29.48%)

On constate que "T6080" définit le premier axe.

Axe 2 :

$\sqrt{\lambda_2} = 0.264$ donc on prend les $|a_i| > 0.427$.

-	+
T4060 (55.04%)	T6080 (41.27%)

Toutes les modalités de cette variable ont une très bonne qualité de représentation sauf pour "T80100" elle est mal représentée (0.040).

La variable type de blessures :

Axe 1 :

-	+
PM (30.28%) BR (7.57%)	FV (8.58%)

La modalité "FV" est en opposition avec "PM" et "BR" ce qui montre que les fracture des vertèbres sont plus grave que les brûlures et les différentes plaies.

Axe 2 :

-	+
FB (16.57%)	FV (53.88%)

Les traumatisme du crâne et du visage ont une qualité de représentation plus faible que les autres types de blessure de plus elles ne définissent aucun axe.

L'interprétation du plan factoriel :

On remarque que "FV" et "FB" sont respectivement proche de "T4060" et "T6080" ce qui montre qu'il sont fortement indemnisées car elles sont des graves blessures.

D'autre part on voit que les plaies et les brûlures sont au voisinage d'un taux de 0% à 20% donc il sont faiblement indemnisées.

Remarques :

On constate que les traumatismes des membres inférieurs sont des blessures graves puisqu'ils sont proches d'un taux d'indemnisation très élevé ("T80100").

Ainsi on remarque que les autres types de traumatismes ont un taux d'indemnisation moyen entre 20 % et 40 %..

4.2.3 L'AFC globale :

Dans cette partie on va essayer de faire une AFC globale sur toutes les variables qu'on a. On va considérer les deux variables accident et victime avec :
Victime : contient les modalités genre,âge,type d'accident et le taux d'indemnisation.
Accident : contient les modalités type de blessures ,période et durée de guérison.

On obtient le tableau de contingence suivant :

Victime	Accident	Type de blessure								période				Durée de guérison						
		TMI	TMS	TCR	AT	FB	PM	BR	FV	A	H	P	E	D0002	D0204	D0406	D0608	D0810	D1012	PLUS
genre	M	34	30	28	25	4	10	3	5	12	33	28	35	56	27	9	4	6	3	3
	F	13	9	14	18	1	2	0	1	3	7	14	18	28	9	3	0	1	1	0
Age	Enf	11	12	17	4	1	0	1	0	2	12	12	11	24	12	1	0	0	0	0
	A1830	9	5	6	4	0	2	1	2	2	6	5	9	13	4	2	0	2	1	0
	A3050	15	15	12	23	3	6	0	3	7	12	16	23	30	9	7	3	4	3	2
	A5070	11	6	4	9	0	4	1	1	3	7	8	9	14	9	2	1	0	0	1
	A7090	1	1	3	3	1	0	0	0	1	3	1	1	3	2	0	0	1	0	0
Type d'accident	Pa	12	7	14	19	2	5	0	3	2	11	15	24	27	14	15	1	2	2	1
	C	13	17	14	17	2	4	3	3	9	13	15	19	28	13	4	3	4	2	2
	Pt	12	15	14	7	1	3	0	0	4	16	12	10	29	9	3	0	1	0	0
Taux	T020	18	25	25	14	0	12	3	0	6	21	22	31	68	8	3	0	1	0	0
	T2040	16	10	13	17	2	0	0	2	3	16	10	16	14	20	7	1	1	1	1
	T4060	9	4	3	6	1	0	0	3	3	2	9	5	2	8	1	2	3	2	1
	T6080	4	0	0	4	2	0	0	1	2	1	1	1	0	0	1	0	2	1	1
	T80100	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0

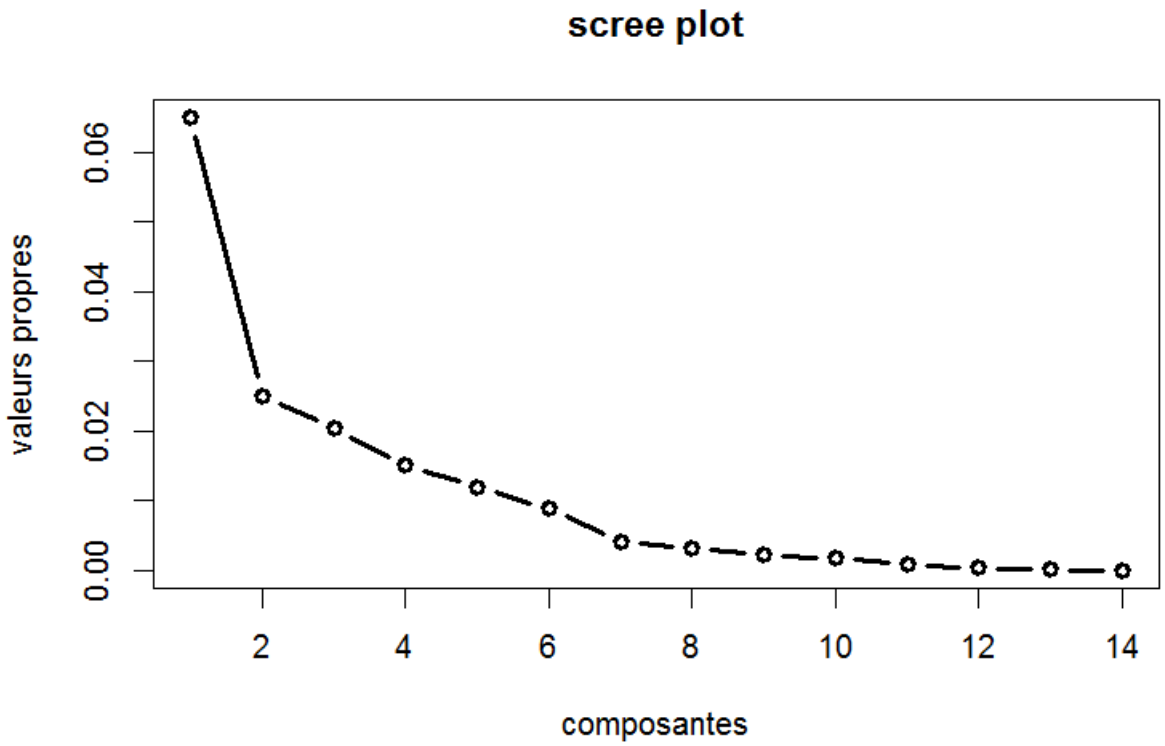
Le test de chi deux nous donne $\chi^2_{calcul} = 316.31$ or on ne peut pas trouver la valeur de χ^2_{252} dans la table de χ^2 ,pour cela on va admettre le résultat suivant :

$$\sqrt{2\chi^2_n} - \sqrt{2n-1} \approx N(0,1)$$

on trouve $\chi^2_{252} = 289.75$ où la dépendance.

1- Valeur propre :

	Valeur propre	pourcentage	Pourcentage cumulé
1	0.06489553	40.827973	40.82797
2	0.02501416	15.737255	56.56523
3	0.02042068	12.847340	69.41257
4	0.01502564	9.453142	78.86571
5	0.01200654	7.553724	86.41944
6	0.00888270	5.588409	92.00785
7	0.00407310	2.562527	94.57037
8	0.00324746	2.043089	96.61346
9	0.00222541	1.400086	98.01355
10	0.00178183	1.121014	99.13456
11	0.00092042	0.579070	99.71363
12	0.00032847	0.206653	99.92029
13	0.00010695	0.067287	99.98757
14	0.00001975	0.012426	100.00000



2-Les résultats pour les profils lignes :

Les cordonnés :

```
$coord
             Dim 1        Dim 2
M        0.03873201  0.001819409
F       -0.06798463 -0.015787663
enfant -0.24639497 -0.235707471
A1830   -0.04249815  0.015456321
A3050    0.15317573  0.147974199
A5070   -0.01144233 -0.019923278
A7090    0.24590266 -0.062740269
Pa       0.03428507 -0.073534260
C        0.10958330  0.131871936
Pt      -0.20152219 -0.041197645
T020    -0.39313809  0.142240304
T2040    0.19173587 -0.285429630
T4060    0.59203816 -0.134661705
T6080    1.28965098  0.187464831
T80100   1.01938229  1.910897620
```

La contribution :

```
$contrib
               Dim 1       Dim 2
M        0.412382695  0.00236075
F        0.508208537  0.07110252
enfant   5.782303612 13.72817307
A1830    0.106288362  0.03647429
A3050    3.488295535  8.44567103
A5070    0.008921621  0.07017222
A7090    0.983281512  0.16606304
Pa       0.150184816  1.79235625
C        1.803942172  6.77746735
Pt       4.308225407  0.46711766
T020    30.757814297 10.44574468
T2040    4.326955459 24.87726508
T4060   17.099053143  2.29503709
T6080   27.045548309  1.48258583
T80100   3.218594522 29.34240915
```

Le cosinus carré :

```
$cos2
              Dim 1        Dim 2
M       0.119348715 0.0002633532
F       0.050380522 0.0027169202
enfant  0.319243183 0.2921491717
A1830   0.024142297 0.0031933819
A3050   0.350391415 0.3269983177
A5070   0.001052947 0.0031922652
A7090   0.127292427 0.0082864563
Pa      0.021279352 0.0978876443
C       0.221108348 0.3201998580
Pt      0.408331376 0.0170652160
T020    0.816045400 0.1068241460
T2040   0.229499645 0.5085965375
T4060   0.581798013 0.0300996228
T6080   0.650413085 0.0137430933
T80100  0.109346598 0.3842429854
```

3-Les résultats pour les profils colonnes :

Les cordonnés :

```
$coord
            Dim 1        Dim 2
TMI     0.07723403 -0.070927617
TMS    -0.10251769 -0.005552246
TCR    -0.11840876 -0.026406449
AT      0.26197551  0.064530970
FB      0.82179983 -0.037351341
PM     -0.30036661  0.338372049
BR     -0.33850832  0.309615515
FV      0.52566513 -0.096658576
A       0.32057193  0.345199754
H      -0.10848407 -0.106338816
P      -0.01314287 -0.063666223
E      -0.06474800  0.028932867
D0002 -0.29946827  0.082069169
D0204  0.11143926 -0.324813689
D0406  0.22237450 -0.139928078
D0608  0.57459564  0.825575798
D0810  0.80284216  0.067774620
D1012  0.83868836  0.078416753
plus   0.89565738  0.125963409
```

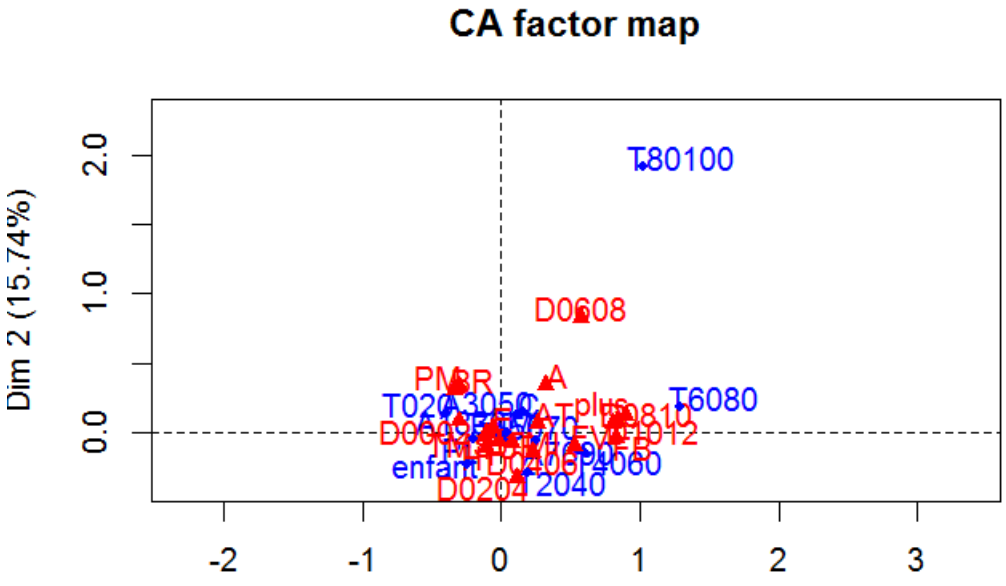
La contribution :

```
$contrib
             Dim 1        Dim 2
TMI     0.86837510  1.899982162
TMS     1.31025423  0.009970667
TCR     1.84564731  0.238138642
AT      8.76874488  1.380324709
FB     10.45909714  0.056053576
PM      3.35333272 11.040560931
BR      1.06476082  2.310934389
FV      4.92128239  0.431687587
A       4.69499476 14.123876275
H       1.43986282  3.589233140
P       0.02300601  1.400578836
E       0.69470201  0.359879961
D0002 23.12479876  4.505730805
D0204  1.42321426 31.368322981
D0406  1.76140774  1.809374682
D0608  4.60181623 24.646051373
D0810 13.47585146  0.249148850
D1012  8.71471773  0.197650826
plus   7.45413365  0.382499607
```

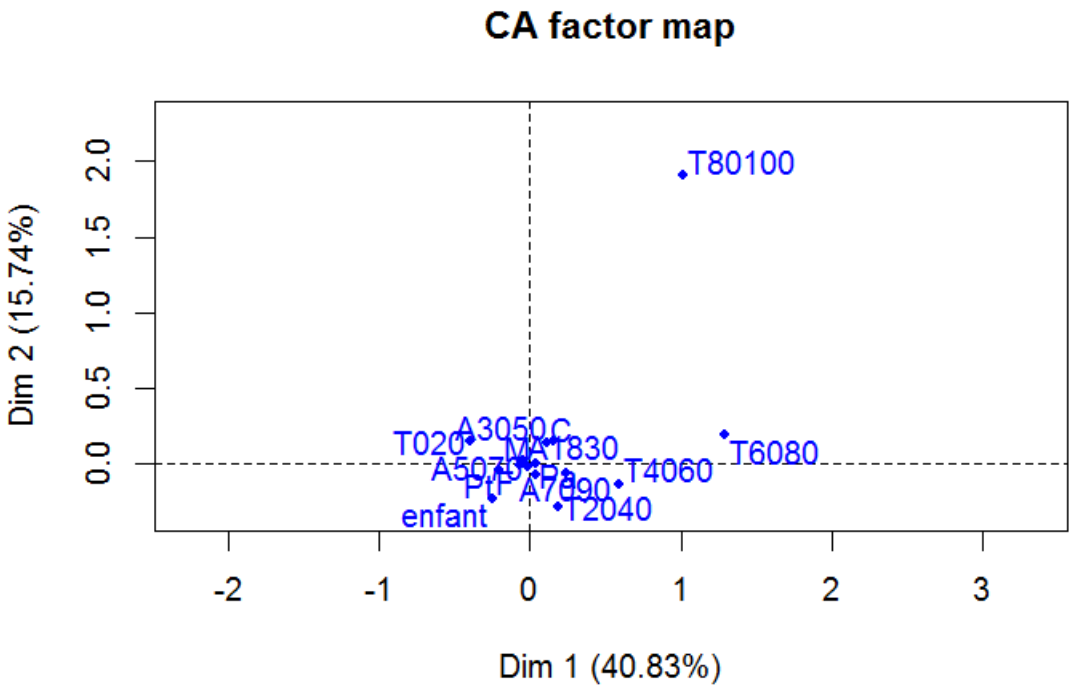
Le cosinus carré :

```
$cos2
            Dim 1        Dim 2
TMI    0.151707874 0.1279444515
TMS    0.164175622 0.0004815578
TCR    0.216488756 0.0107668255
AT     0.506344065 0.0307227951
FB     0.569045236 0.0011755121
PM     0.240673843 0.3054320161
BR     0.123659962 0.1034512711
FV     0.381354241 0.0128940998
A      0.363614460 0.4216296181
H      0.176969914 0.1700400264
P      0.006171475 0.1448193077
E      0.120163158 0.0239939426
D0002 0.858185293 0.0644524631
D0204 0.075326942 0.6399448154
D0406 0.283886449 0.1124046933
D0608 0.166971776 0.3446928837
D0810 0.720533956 0.0051348563
D1012 0.716249194 0.0062615323
plus   0.708452613 0.0140125055
```

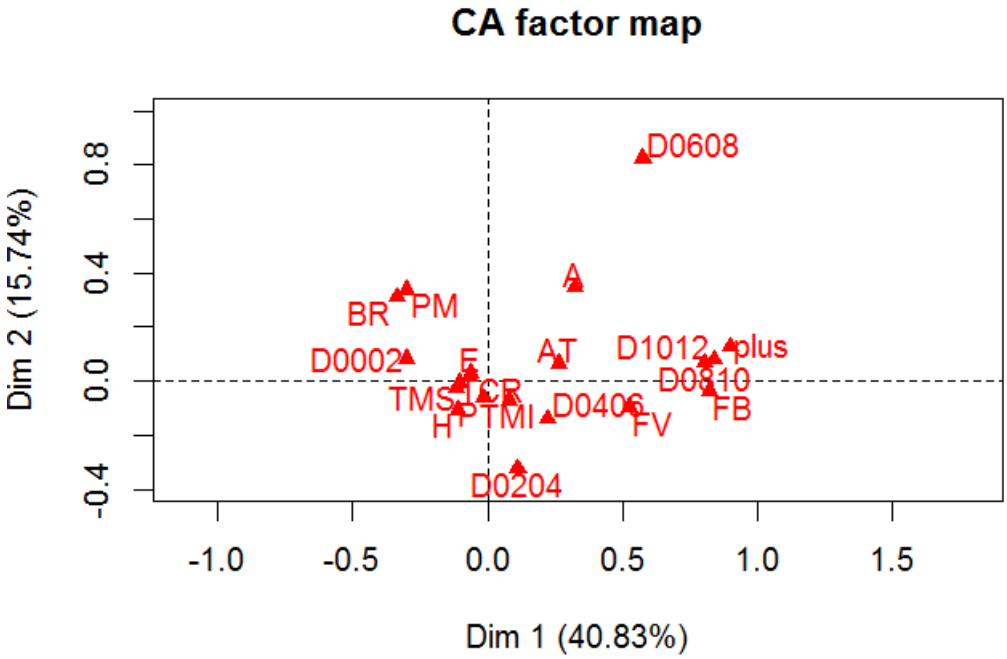
4-Le plan factoriel :



On voit que le graphe est surchargé donc on va utiliser la commande **invisible**. **plot(x , invisible=c("col","col.sup"))** pour afficher que les profils lignes et on obtient le graphe suivant :



plot(x , invisible=c("row","row.sup")) pour afficher que les profils colonnes et on obtient le graphe suivant :



5-Discussion des résultats :

La variable victime :

Axe1 :

On prend les modalités qui ont des cordonnées supérieures à $\sqrt{\lambda_1} = 0.25474$

-	+
T020	T4060
	T6080
	T80100

Axe2 :

On prend les modalités qui ont des cordonnées supérieures à $\sqrt{\lambda_2} = 0.1581$

-	+
Enfant T4060	T6080 T80100

On remarque que T020 est bien représentée (0.91) et même pour T2040 avec une qualité de représentation égale à 0.73 par contre on constate que A5070 et très mal représentée.

La variable accident :

Axe1 :

-	+
PM BR D0002	AT FB FV A D0608 D0810 D1012 plus

Axe2 :

-	+
D0204	PM BR A D0608

L'interprétation du plan factoriel :

En général , on constate que les blessures de type "FB" et "FV" ont des durées de guérison trés longues avec un taux d'indemnisation plus élevé car il sont des blessures très compliquées par contre on voit que "PM" et "BR" sont faiblement indemnisées avec des durées de guérison petites.

On remarque que les conducteurs sont plus proches de la modalité "M" ce qui est logique car la plupart des conducteurs sont des hommes ainsi on observe que cette catégorie a presque un âge entre 18 ans et 50 ans .

les deux modalitées "masculin" et "féminin" sont situées au centre du plan factoriel et elles sont tous proches des autres modalitées.

On observe que la plupart des piétons sont des enfants et des personnes qui ont un âge entre 50 et 70 ans (ces modalités sont proches entre eux) ainsi on remarque que les passagers ont un âge entre 50 et 90 ans.

On voit que les traumatismes crâniens et les traumatismes des membres supèrieurs sont les plus fréquents chez les piétons tandis que on trouve que les traumatismes des membres infèrieurs sont beaucoup plus chez les passagers.

On remarque aussi que "automne" est proche des durées longues et des taux d'indemnisation importants ce qui signifie que les accidents les plus graves se sont passés dans cette saison.

Conclusion générale

L'analyse factorielle est une technique statistique utilisée pour dépouiller des enquêtes : elle permet, quand on dispose d'une population d'individus pour lesquelles on possède de nombreux renseignements concernant les opinions, les pratiques et le statut (sexe, âge, etc.), d'en donner une représentation géométrique c'est-à-dire en utilisant un graphique qui permet de voir les rapprochements et les oppositions entre les caractéristiques des individus.

Au cours de ce projet, on a procédé à une étude des deux méthodes ACP et AFC et on a illustré chaque méthode par une application sur des données réelles sur un groupe de cent cinquante personnes portant sur l'indemnisation des assurances.

Les méthodes factorielles sont adaptées chacune à un type de tableau particulier, en analyse en composantes principales les données sont des individus (en ligne) décrits par des variables quantitatives (en colonne) or l'analyse factorielle des correspondances est dédiée aux tableaux de contingence.

Notre application a montrée qu'il y'a presque une relation affine entre la durée de guérison et le taux d'indemnisation cela veut dire que pour une blessure grave on a un taux très elevé et une durée de guérison importante et vice versa.

Bibliographie

[1] Ben Abdeljelil Khaoula ,*Cours d'algèbre linéaire*

[2] Brigitte Escofier,Jerôme Pagès,*Analyses factorielles simples et multiples*,4^{eme} édition,Edition Dunod,Paris 2008

[3] Gilbert Saporta ,*probabilités et analyse des données statistiques*, 3^{eme} édition ,Edition Technip,2006.

[4] Martin Bilodeau,David Brenner,*Theory of multivariate statistics*,Edition Springer,1999

[5] **lien :**http ://www.real-statistics.com/correlation/multiple-correlation

[6] Revue *Mathematics and Social Sciences*,2006,n^o 173.
lien : https ://msh.revues.org/2963 ?file=1

[7] RiccoRakotomalala,*Analyse de corrélation*
lien : eric.univ-lyon2.fr/ ricco/cours/cours/Analyse_*de*_Correlation.pdf

[8] Tutoriel FactoMineR pour l'analyse des correspondances multiples
lien : http ://www.quanti.ihmc.ens.fr/IMG/pdf/Tutoriel_FactomineR_ACM.pdf

Résumé

Les méthodes d'analyse factorielle sont largement utilisées dans tous les domaines. Elles consistent à rechercher des facteurs en nombre restreint et résumer le mieux possible les données considérées.

Dans ce travail on va traiter les principales caractéristiques des méthodes les plus courantes de l'analyse factorielle et on va donner pour chacune d'elle une application sur un groupe de 150 malades portant sur l'indemnisation des assurances suites à des accidents corporels.

Abstract

Factor analysis methods are widely used in all areas. They consist of searching for factors in limited numbers and summarizing as closely as possible the data considered.

In this work we will deal with the main characteristics of the most common methods of factor analysis and we will give for each of them an application on a group of 150 patients relating to the compensation of insurances following bodily injury.

ملخص

تستخدم أساليب تحليل العوامل على نطاق واسع في جميع المجالات و المتمثلة في البحث عن عوامل محدودة وتلخيص البيانات بأكبر قدر ممكن. في هذا العمل سوف نتطرق للخصائص الرئيسية للطرق الأكثر شيوعا لتحليل العوامل و سنقدم لكل واحد منهم تطبيق على مجموعة تتكون من 150 مريض تتعلق بتعويض التأمينات المتعلقة بالإصابات الجسدية .