Université ABOU-BEKR BELKAID Tlemcen
Faculté des sciences
Département de mathématiques

Mémoire en vue de l'obtention du diplôme du Master

Option : *Probabilités et statistique*

Sous le thème

Méthodes ACP et AFC en statistiques et leurs applications

Présenté par

Mme. MERAD Menel nedjla

Soutenu publiquement le 22 Octobre 2015 devant le jury composé de:

Mme	M. Dali Youcef	Président.
Mme.	W. Benyelles	Examinateur.
Mr.	M. Abbes	Examinateur.
Mr	T. Mourid	Encadreur.

Année Universitaire : 2014/2015

Remerciements

Ce travail s'inscrit dans le cadre d'un projet de fin d'étude, en vue de l'obtention du diplôme de Master, mené au niveau du Département de mathématiques de la Faculté des Sciences de l'Université Abou-Bekr Belkaïd de Tlemcen, sous la direction de Monsieur T. MOURID, Professeur à l'Université de Tlemcen.

Au terme de ce projet, je tiens à remercier Monsieur T. MOURID Professeur à l'Université de Tlemcen, pour l'honneur qu'il m'a fait de bien vouloir m'encadrer et pour les conseils qu'il m'a donnés lors de la réalisation de ce manuscrit.

Aussi, je souhaite adresser mes remerciements les plus sincères à madame W. BENYELLES qui s'est toujours montrée à l'écoute ainsi que l'aide et le temps qu'elle a bien voulu me consacrer et j'adresse aussi mes remerciements au membre du jury madame Dali Youcef et monsieur Abass.

Je tiens à remercier Monsieur B.MABKHOUT le chef de département de mathématique à l'université des Science Tlemcen pour son soutien.

J'adresse mes plus sincères remerciements à tous mes proches, mes collègues qui m'ont toujours soutenus et encouragé.

Enfin, Je tiens à exprimer ma profonde gratitude à toutes celles et ceux qui m'ont apporté leur soutien, leur amitié ou leur expérience tout au long de ce travail.

Example of a Thesis

Formatted with LaTeX

Using the University of Georgia

Style Macro Package, Version 2.0

by

Michael A. Covington

(Under the direction of Abraham Baldwin)

Abstract

This is the abstract, a brief summary of the contents of the thesis. It is limited to 150 words in length for a master's thesis or 350 words for a doctoral dissertation.

The abstract page(s) are not numbered and are not necessarily included in the bound copies. Likewise, the signature page is not counted in page numbering because not all copies contain it.

Throughout this sample thesis, **please note that the layout obtained with LaTeX is not meant to be a perfect duplicate of the Microsoft Word examples in the *Graduate School Style Manual.*** LaTeX has additional typographic tools at its disposal, such as small capitals and various subtle adjustments of spacing, which are used by the LaTeX UGa style sheet in accordance with the standard practices of the book-printing industry.

The index words at the bottom of the abstract should be chosen carefully, preferably with the help of one or two of your colleagues. They are the words by which people will find your thesis when searching the scientific literature. If you want to get credit for your ideas, be sure to choose a good set of index words so that people doing related work will know about yours.

Index words: word processing, computer typesetting, computer graphics, style sheets, typography, dissertations, theses (academic)

Example of a Thesis Formatted with LaTeX Using the University of Georgia Style Macro Package, Version 2.0

by

Michael A. Covington

B.A., The University of Georgia, 1977

M.Phil., Cambridge University, 1978

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2004

Example of a Thesis Formatted with LaTeX Using the University of Georgia Style Macro Package, Version 2.0

by

Michael A. Covington

Approved:

Major Professor: Abraham Baldwin

Committee: Committee Member's Name
Another Committee Member
Third Committee Member
Fourth Committee Member
Fifth (last) Committee Member

Electronic Version Approved:

Gordhan L. Patel
Dean of the Graduate School
The University of Georgia
May 2004

Table des matières

Page

Chapitre

Chapitre 1

Introduction

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. Les méthodes et les indices de dépendance varient selon la nature (qualitative, ordinale, numérique) des variables étudiées. On peut rendre compte de l'existence d'un lien entre deux variables numériques ou plus à l'aide du coefficient de corrélation linéaire, qui intervient dans les formules de différents indicateurs de liens statistiques. Lorsque les variables ne sont plus uniquement quantitatives, on dispose du rapport de corrélation qui est utilisé pour caractériser l'association entre une variable quantitative et un variable qualitative, comme il peut mesurer la liaison entre deux variables numériques lorsque la relation s'écarte de la linéarité. Lorsque les variables sont ordinales, on parle de corrélation des rangs, ou interviennent deux coefficients d'une grande importance qui sont : le coefficient de Spearman et le coefficient de Kendall. Et finalement lorsque les variables sont toutes qualitatives, on adopte une mesure différente de toutes celles qui l'a précèdent qui est l'écart à l'indépendance. C'est ce que nous allons détailler dans le premier chapitre.

La description des liaisons entre deux variables par des techniques statistiques bidimensionnelles conduit à se poser la question de la représentation simultanées de données en dimension plus grande que deux. Quelle graphique permettrait de "généraliser" le nuage de points tracé dans le

cas de deux variables permettant d'aborder la structure de corrélation présente entre plus de deux variables. L'outil utilisé est alors l'analyse en composantes principales (ACP). C'est ce que nous allons aborder dans le deuxième chapitre

L'analyse des corrélations canoniques ou encore analyse canonique simple est une méthode statistique, proposée en 1936 par Hotelling, surtout connue pour ses qualités théoriques, puisqu'elle englobe de nombreuses autres méthodes statistiques, parexemple, l'analyse factorielle des correspondances (AFC). Elle permet de décrire les relations linéaires qui existent entre deux ensembles de variables mesurées sur les mêmes individus,c'est ce que nous allons voir au troisième chapitre.

En fin, nous allons ajouter une application de la méthode statistique AFC sur des données réelles sur un groupe de cent personnes portant sur l'indemnisation des assurances suites à des accidents corporels.

Chapitre 2

Coefficient de corrélation

2.1 Le coefficient de corrélation linéaire de Bravais-Pearson

Soit X et Y deux variables numériques

Le coefficient de corrélation linéaire dit de ‹Bravais Pearson› est un indice statistique qui exprime l'intensité et le sens $(+ou-)$ de la relation linéaire entre deux variables quantitatives .Il est défini par la relation suivante :

Définition 1 *Le coefficient de corrélation est défini par*

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

où S_X et S_Y sont les écarts types de X et Y, le numérateur S_{XY} est la covariance.

Quand on observe $x_1, ..., x_n$ de X et $y_1, ..., y_n$ de Y , on peut estimer r_{XY} par :

$$\hat{r}_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\hat{S}_X \hat{S}_Y}$$

Où $\bar{x}$ et $\bar{y}$ sont les moyennes des X et Y respectivement, et $\hat{S}_X$,$\hat{S}_Y$ sont les estimateurs de S_X,S_Y respectivement.

Ce coefficient varie entre $-1et1$.

- si r est proche de 0, on peut dire qu'il n'y a pas de relation linéaire entre X et Y.

- *si r est proche de -1, il existe une forte relation linéaire négative entre X et Y .*

- *si r est proche de 1, il existe une forte relation linéaire positive entre X et Y.*

En présence de n couples, on a deux vecteurs de $\mathbb{R}^n$:

$$x = \begin{bmatrix} x_1 \\ . \\ . \\ x_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix}$$

Considérons les deux vecteurs des variables centrées de $\mathbb{R}^n$:

$$\acute{x} = \begin{bmatrix} x_1 - \bar{x} \\ . \\ . \\ x_n - \bar{x} \end{bmatrix} \qquad \acute{y} = \begin{bmatrix} y_1 - \bar{y} \\ . \\ . \\ y_n - \bar{y} \end{bmatrix}$$

r *est le cosinus formé par les vecteurs* $\acute{x}$ *et* $\acute{y}$.

En effet,

la covariance S_{XY} *de* x *et* y *est en fait le produit scalaire des variables centrées i.e de* $\acute{x}$ *et* $\acute{y}$.

$$S_{xy} = \acute{x} \cdot \acute{y} = \|\acute{x}\|\|\acute{y}\| \cos(\alpha)$$

Où α *est l'angle formé par* $\acute{x}$ *et* $\acute{y}$.

et les écarts type sont la norme des variables centrées i.e

$$S_x = \|\acute{x}\| \quad et \quad S_y = \|\acute{y}\|$$

alors :

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \cos(\alpha)$$

- *Si* $r = 1, \alpha = 0°$, *les vecteurs* $XetY$ *sont colinéaires (parallèles).*

- *Si* $r = 0, \alpha = 90°$*, les vecteurs* $XetY$ *sont orthogonaux.*
- *Si* $r = -1, \alpha = 180°$*, les vecteurs* $XetY$ *sont colinéaires de sens opposé.*

2.2 Matrice de corrélation de p variables

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.

$$X = [x_{ij}]_{i,j}$$

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

Le tableau des données centrées Y s'obtient en utilisant l'opérateur de centrage A.

On définit la matrice A par ;

$$A = I - \frac{1\acute{1}}{n} = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ & . & \\ & & . \\ 0 & & 1 \end{bmatrix} - \frac{1}{n} \begin{bmatrix} 1 \\ . \\ . \\ . \\ 1 \end{bmatrix} \begin{bmatrix} 1 & . & . & . & 1 \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ & . & \\ & & . \\ 0 & & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{n} & . & . & . & \frac{1}{n} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ \frac{1}{n} & . & . & . & \frac{1}{n} \end{bmatrix}$$

Matrice carrée de taille n, appelée opérateur de centrage de terme général $a_{ii} = 1 - \frac{1}{n}$ et $aij = -\frac{1}{n}$ $i \neq j$.

donc

$$Y = \begin{bmatrix} x_{11} - \bar{x}_1 & & x_{1p} - \bar{x}_p \\ & & \\ & & \\ x_{n1} - \bar{x}_1 & & x_{np} - \bar{x}_p \end{bmatrix}$$

La matrice de covariance des p variables est une matrice symétrique. Ses éléments diagonaux sont les variances et les éléments extra-diagonaux sont les covariances des couples de variables, elle est définie par la matrice V :

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

où $S_{kl} = \frac{1}{n} \sum_{i=1}^{n} x_{ik} x_{il} - \bar{x}_k \bar{x}_l$.

La matrice V s'obtient en utilisant Y : $V = \frac{1}{n} \acute{Y} Y$.

On définit la matrice symétrique positive C regroupant tous les coefficients de corrélation linéaire entre les p variables.

$$C = (r_{ij}) = \begin{bmatrix} 1 & r_{12} & ... & r_{1p} \\ & . & & \\ & & . & \\ r_{p1} & & & 1 \end{bmatrix}$$

En posant $D_{\frac{1}{S}}$ la matrice diagonale suivante :

$$D_{\frac{1}{S}} = \begin{bmatrix} \frac{1}{S_1} & & 0 \\ & & \\ & & \\ 0 & & \frac{1}{S_p} \end{bmatrix}$$

On a donc $C = D_{\frac{1}{S}} V D_{\frac{1}{S}}$

Exemple 2 *On se propose d'étudier la relation existant entre les variables suivantes : cylindrée, puissance, longueur, largeur, poids et vitesse de pointe pour 18véhicules*

	Nom	Cyl	$Puis$	Lon	Lar	$Poid$	$Vitesse$
1	$ALFASUD-TI-1350$	1350	79	393	161	870	165
2	$AUDI-100-L$	1588	85	468	177	1110	160
3	$SIMCA-1370-GLS$	1294	68	424	168	1050	152
4	$CITREN-GS-CLUB$	1222	59	412	161	930	151
5	$FIAT-132-1600GLS$	1585	98	439	164	1105	165
6	$LANCIA-BETA-1300$	1297	82	429	169	1080	160
7	$PEUGEOT-504$	1796	79	449	169	1160	154
8	$RENAULT-16-TL$	1565	55	424	163	1010	140
9	$RENAULT-30-TS$	2664	128	452	173	1320	180
10	$TOYOTA-COROLLA$	1166	55	399	157	815	140
11	$ALFETTA-$	1570	109	428	162	1060	175
12	$PRINCESS-1800-HL$	1798	82	445	172	1160	158
13	$DATSUN-200L$	1998	115	469	169	1370	160
14	$TAUNUS-2000-GL$	1993	98	438	170	1080	167
15	$RANCHO$	1442	80	431	166	1129	144
16	$MAZDA-9295$	1769	83	440	165	1095	165
17	$OPEL-REKORD-L$	1979	100	459	173	1120	173
18	$LADA-1300$	1294	68	404	161	950	140

La matrice V calculé avec $n-1$ en dénominateur :

	CYL	$PUIS$	LON	LAR	$POIDS$	$VITESSE$
CYL	139823.5294	6069.7451	5798.7059	1251.2941	40404.2941	3018.5686
$PUIS$	6069.7451	415.1928	288.9118	56.3922	2135.6961	208.8791
LON	5798.7059	288.9118	488.7353	99.7647	2628.3824	127.7353
LAR	1251.2941	56.3922	99.7647	28.2353	521.7059	30.5098
$POIDS$	40404.2941	2135.6961	2628.3824	521.7059	18757.4412	794.1078
$VITESSE$	3018.5686	208.8791	127.7353	30.5098	794.1078	147.3889

La matrice de corrélation C est la suivante :

	CYL	$PUIS$	LON	LAR	$POIDS$	$VITESSE$
CYL	1.00000	0.79663	0.70146	0.62976	0.78895	0.66493
$PUIS$	0.79663	1.00000	0.64136	0.52083	0.76529	0.84438
LON	0.70146	0.64136	1.00000	0.84927	0.86809	0.47593
LAR	0.62976	0.52083	0.84927	1.00000	0.71687	0.47295
$POIDS$	0.78895	0.76529	0.86809	0.71687	1.00000	0.47760
$VITESSE$	0.66493	0.84438	0.47593	0.47295	0.47760	1.00000

Conclusion :

Compte tenu des deux matrices, on constate que les variables sont fortement corrélées.

2.2.1 Caractère significatif d'un coefficient de corrélation

Le premier test qui vient à l'esprit est sur les valeurs significatives de la corrélation c'est-à-dire : le coefficient de corrélation r_{XY} est-il significativement différent de zéro ?.

Le test s'écrit :

$$H_0 \quad : \quad r_{XY} = 0$$

$$H_1 \quad : \quad r_{XY} \neq 0$$

2.2.2 Test exact

Le test étudié dans cette section est paramétrique. On suppose à priori que les observations proviennent d'un couple gaussien :

. Dans ce cas, la distribution sous H_0 de la statistique du test que nous présenterons plus bas est exacte. Le test de significativité est équivalent à un test d'indépendance..

2.2.3 Statistique du test

Sous H_0, la statistique :

$$T = \frac{R}{\sqrt{1-R^2}}\sqrt{n-2}$$

suit une loi de Student à $(n-2)$ degrés de liberté T_{n-2}.(Réf 1)

Où $R = \hat{r}_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\hat{S}_X \hat{S}_Y}$.

2.2.4 Région critique

La région critique (rejet de l'hypothèse nulle) du test au risque α s'écrit :

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

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.

Remarque

Autres hypothèses alternatives. On peut définir une hypothèse alternative différente ($H_1 : r < 0$ *ou* $H_1 : r > 0$). Les caractéristiques des distributions restent les mêmes. Pour un risque α donné, le seuil de rejet de H_0 est modifié puisque le test est unilatéral dans ce cas.

Exemple numérique

Toujours avec les caractéristiques numériques des véhicules :

Le coefficient de corrélation entre la variable poids et la vitesse vaut : $.r = 0.47760$.

On souhaite tester sa significativité au risque $\alpha = 0.05$.

Nous devons calculer les éléments suivants :

- La statistique du test : $t = \frac{0.47760.}{\sqrt{\frac{1-(0.47760)^2.}{18-2}}}.= 9.89$
- Le seuil théorique au risque α est $.t_{0.975} = 2.120$

Nous rejetons donc l'hypothèse nulle car $t >.t_{0.975}$ c'est-à-dire que les deux variables sont fortement liées.

2.2.5 Test asymptotique

Dans le cas générale lorsque $n > 100$, la loi de R suit approximativement une loi normale $N(0; \frac{1}{\sqrt{n-1}})$.

2.3 Corrélation partielle

Souvent, la dépendance entre deux variables est en fait la conséquence des variations d'une variable tierce.

Les coefficients de corrélation partielle constituent un moyen d'éliminer l'influence d'une ou plusieurs variables

Définition 3 *Dans le cas gaussien et pour un triplet* (X, Y, Z) *et un échantillon observé de n valeurs numériques de ce triplet* $((x_1, y_1, z_1), ..., (x_n, y_n, z_n))$. *Le coefficient de corrélation partielle de* X *et* Y *avec* Z *noté* $r_{XY.Z}$ *est défini à partir des corrélations brutes :*

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}}$$

L'idée est de retrancher de la relation direct entre (X, Y) *les relations respectives de* X *et* Y *avec* Z *puis un terme de normalisation de manière à ce que* -1 $\leq r_{XY.Z} \leq 1$

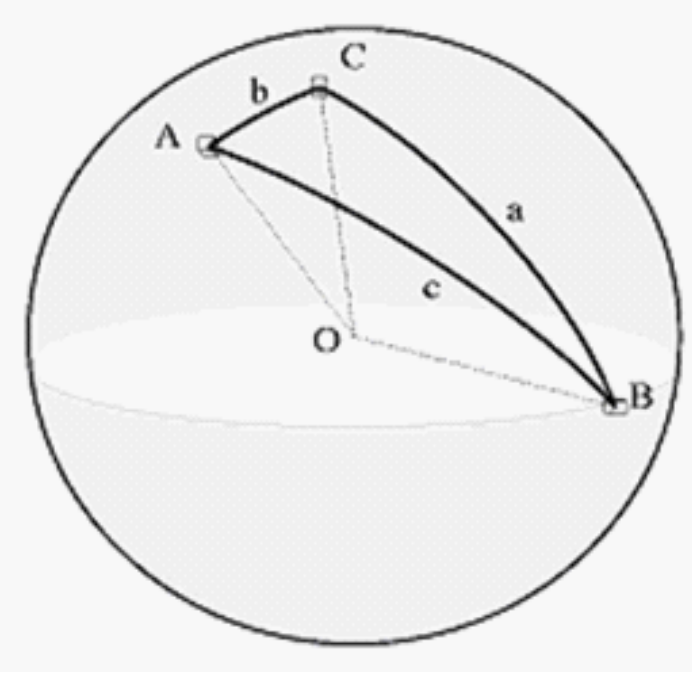
Preuve. La démonstration la plus rapide de la formule consiste à s'appuyer sur l'interprétation géométrique de la corrélation (cosinus).

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 grands cercles BC, AC et AB. Les coefficients de corrélations entre ces vecteurs sont

$$\begin{aligned} r_{BC} &= \cos(a) \\ r_{AC} &= \cos(b) \\ r_{AB} &= \cos(c) \end{aligned}$$

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


C
b
A
a
c
O
B

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

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, et

$$r_{ABC} = \cos(C)$$

est la « corrélation partielle » entre A et B quand C est fixé.

Autre démonstration :

La corrélation entre X et Y (étant donné) est la corrélation simple entre X et Y étant enlevé l'effet de linéaire de Z

Pour cela soit

$$\begin{aligned}X_{.Z} &= X - aZ \\ Y_{.Z} &= Y - bZ\end{aligned}$$

Où a et b sont les coefficients de la régression

$$\begin{aligned}a &= \frac{S_{XZ}}{S_Z^2} \\ b &= \frac{S_{YZ}}{S_Z^2}\end{aligned}$$

Le coefficient de corrélation entre $X_{.Z}$ et $Y_{.Z}$ est

$$r_{X_{.Z}Y_{.Z}} = \frac{S_{X_{.Z}Y_{.Z}}}{S_{X_{.Z}} S_{Y_{.Z}}}$$

On calcule

$$\begin{aligned} S_{X_{.Z}Y_{.Z}} &= S_{XY} - aS_{Y_{.Z}} - bS_{X_{.Z}} + abS_Z^2 \\ S^2_{X_{.Z}} &= S_X^2 - 2aS_{XZ} + a^2 S_Z^2 \\ S^2_{Y_{.Z}} &= S_Y^2 - 2bS_{YZ} + b^2 S_Z^2 \end{aligned}$$

En substituant a et b :

$$\begin{aligned} S_{X_{.Z}Y_{.Z}} &= S_{XY} - \frac{S_{XZ}S_{YZ}}{S_Z^2} \\ S^2_{X_{.Z}} &= S_X^2 - \frac{S_{XZ}^2}{S_Z^2} \\ S^2_{Y_{.Z}} &= S_Y^2 - \frac{S_{YZ}^2}{S_Z^2} \end{aligned}$$

Donc

$$r = \frac{S_{XY} - \frac{S_{XZ}S_{YZ}}{S_Z^2}}{\sqrt{(S_X^2 - \frac{S_{XZ}^2}{S_Z^2})(S_Y^2 - \frac{S_{YZ}^2}{S_Z^2})}} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

■

-Lorsque Z est indépendante de X et Y

$$r_{xy.z} = r_{xy}$$

ie : Z n'a aucun effet dans la relation entre X et Y.

-Lorsque Z est fortement lié positivement avec X et Y ,on peut aboutir au résultat :

$$r_{xy.z} \simeq 0$$

ie : dans la relation (XY) tout est expliqué par Z.

Cette formule se généralise et permet de calculer les divers coefficients de corrélation partielle. Il suffit de remplacer dans la formule précédente les corrélations simples par les corrélations partielles :

Par exemple, si on veut éliminer l'effet linéaire de la variable W en plus de Z , il suffit de remplacer dans la formules précédentes les corrélations simples par les corrélations partielles :

$$r_{XY.ZW} = \frac{r_{XY.Z} - r_{XW.Z} \times r_{YW.Z}}{\sqrt{1 - r^2_{XW.Z}}\sqrt{1 - r^2_{YW.Z}}}$$

Exemple numérique

Le T_iO_2 et le S_iO_2 sont des bons indices de la maturité magmatique des roches volcaniques. On pourrait vouloir éliminer l'effet de la différenciation magmatique sur les corrélations entre les autres variables. Lors de la différenciation magmatique, les minéraux Ferro-magmatiques cristallisent en premier. On observera donc typiquement une corrélation positive entre F_eO et M_gO

.Cependant, ces deux éléments se trouvent en compétition pour occuper les mêmes sites de cristallisation sur les minéraux. Ceci entraîne que pour des roches de maturité magmatique comparable, on devrait observer une corrélation négative entre F_eO et M_gO.

On a alors mesuré S_iO_2, M_gO, F_eO , et on a obtenu, avec 30 observations les corrélations simples suivantes entre ces trois éléments :

	S_iO_2	M_gO	F_eO
S_iO_2	1	-0.86	-0.75
M_gO	-0.86	1	0.50
F_eO	-0.75	0.50	1

Ainsi la corrélation partielle entre M_gO et F_eO (étant donné l'effet de S_iO_2 enlevé) est

$$r_{M_gOF_eO,S_iO_2} = \frac{0.50 - (-0.86)(-0.75)}{\sqrt{(1-(-0.86)^2)(1-(-0.75)^2)}} = -0.429$$

C'est loin d'être la même situation par rapport au coefficient de corrélation simple !

2.3.1 Signification d'un coefficient de corrélation partielle

Si l'hypothèse de normalité est vérifiée, nous adoptons la même démarche que pour la corrélation simple (brute).

Les hypothèses à tester sont :

$$\begin{aligned} H_0 &= r_{XY.Z} = 0 \\ H_1 &= r_{XY.Z} \neq 0 \end{aligned}$$

La statistique du test s'écrit :

$$T = \frac{R_{XY.Z}}{\sqrt{1 - R_{XY.Z}}}\sqrt{n-3}$$

suit une loi de Student à $(n-3)$ degrés de liberté T_{n-3}.Réf 1

La région critique du test est définie par :

$$D = |T| > t_{1-\frac{\alpha}{2}}(n-3)$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $(n-3)$ degrés de liberté.

Reprenons notre exemple : Au seuil $\alpha = 5\%$ la valeur critique est 2.052 : et vaut 2.468 , donc la liaison est significative.

2.4 Corrélation multiple

Définition 4 *Soit une variable numérique Y et un ensemble de p variables également numériques $X_1, X_2, ...X_p$.*

Le coefficient de corrélation multiple R_1 est la valeur maximale prise par le coefficient de corrélation linéaire entre Y et une combinaison linéaire des X_j :

$$R_1 = \sup_{a_1,a_2,...a_p} r(Y; \sum_{j=1}^{p} a_j X_j)$$

On a toujours $0 \leq R_1 \leq 1$.

En d'autres termes, si on pose $Y^\star = b_0 + b_1 X_1 + ... + b_p X_p$

on désire que $Y^\star$ soit le plus proche possible de Y.

Alors si l'espace des variables $\mathbb{R}^n$ est muni de la métrique D , on exigera que $\|Y - Y^\star\|^2$ soit minimal

Donc $R_1 = 1$ s'il existe une combinaison linéaire des X_j telle que

$$y = a_0 + \sum_{j=1}^{p} a_j X_j$$

Interprétation géométrique

On rappelle que le coefficient de corrélation est le cosinus de l'angle formé de $\mathbb{R}^n$ par des variables centrées.

Considérons le sous-espace W de $\mathbb{R}^n$ (de dimension au plus égale à $p+1$), engendré par les combinaisons linéaires des X_j et la constante 1

$Y^\star$ est alors la projection orthogonale sur le sous-espace W On a .

$$R_1 = \frac{cov(Y, Y^\star)}{S_Y S_{Y^\star}}$$

Et puisque :

$$\begin{aligned} cov(Y, Y^\star) &= \|Y - \bar{Y}\| \|Y^\star - \bar{Y}\| \cos(Y - \bar{Y}, Y^\star - \bar{Y}) \\ S_Y &= \|Y - \bar{Y}\| \qquad\qquad S_{Y^\star} = \|Y^\star - \bar{Y}\| \end{aligned}$$

alors

$$R_1 = \frac{cov(Y, Y^\star)}{S_Y S_{Y^\star}} = \cos(Y - \bar{Y}, Y^\star - \bar{Y})$$

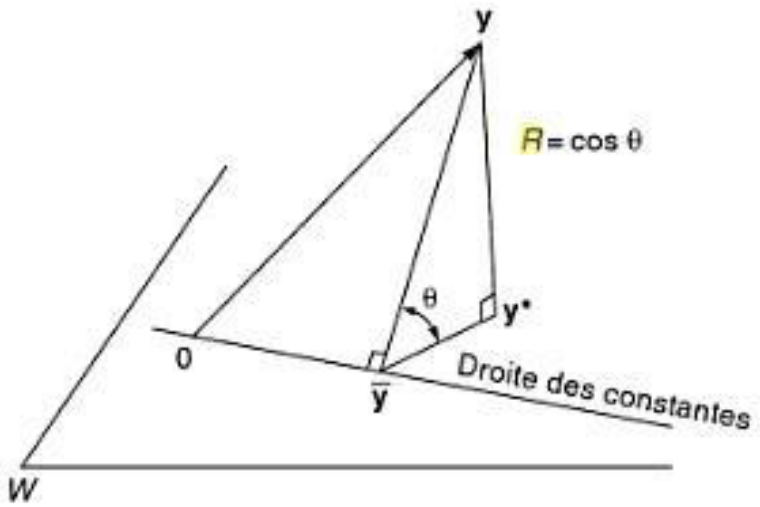
R est le cosinus de l'angle θ formé par la variable centrée $Y - \bar{Y}$ et W, c'est-à-dire l'angle formé par $Y - \bar{Y}$ et sa projection orthogonale $Y^\star - \bar{Y}$.

2.4.1 Calcul de R

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

$$R_1^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - y_i^\star)^2}{\sum (y_i - \bar{y})^2} = \frac{\|Y - \bar{Y}\|^2 - \|Y^\star - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} = \frac{\|Y^\star - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} = \frac{S_{Y^\star}^2}{S_Y} = \cos^2 \theta$$

Soit A la matrice de projection orthogonale sur W, alors :



$$R_1^2 = \frac{(Y-\bar{Y})^t A (Y-\bar{Y})}{\|Y-\bar{Y}\|^2}$$

En effet

$$\begin{aligned} \|Y^\star - \bar{Y}\|^2 &= (Y^\star - \bar{Y})^t (Y^\star - \bar{Y}) = (AY - \bar{Y})^t (AY - \bar{Y}) = Y^t A^t AY - Y^t A^t \bar{Y} - \bar{Y}^t AY + \bar{Y}^t \bar{Y} \\ &= Y^t AY - Y^t A\bar{Y} - \bar{Y}^t AY + \bar{Y}^t A\bar{Y} = (Y-\bar{Y})^t A (Y-\bar{Y}) \end{aligned}$$

car

$$\begin{aligned} Y^\star &= AY \\ A &= A^t (\textit{symétrique}) \\ A^2 &= A \end{aligned}$$

2.4.2 Signification d'un coefficient de corrélation multiple

Si Y est indépendante des X_j, sachant que les n observations proviennent d'un couple gaussien, alors :

$$\frac{\hat{R}_1^2}{1-\hat{R}_1^2}\frac{n-p-1}{p} = F(p, n-p-1)$$

est une loi de Fisher

On retrouve comme cas particulier la loi du coefficient de corrélation linéaire simple en faisant $p = 1$.

2.5 Corrélation des rangs

Il arrive souvent de ne disposer que d'un ordre sur un ensemble d'individus et non de valeurs numériques d'une variable mesurable : soit parce qu'on ne dispose que de données du type classement, ou bien parce que les valeurs numériques d'une variable n'ont que peu de sens et n'importent que par leur ordre.

2.5.1 Le coefficient de corrélation des rangs de Spearman

La corrélation de Spearman (nommée d'après Charles Spearman), ou rho de Spearman, est étudiée lorsque deux variables statistiques semblent corrélées sans que la relation entre les deux variables soit de type affine. Elle consiste à trouver un coefficient de corrélation, non pas entre les valeurs prises par les deux variables mais entre les rangs de ces valeurs.

Exemple 5 *Classer les sujets selon leur rang pour chacune des deux variables*

$Sujet$	$Rang : K$	$Rang : L$	$Différence\ des\ rangs$	$Différence\ au\ carré$
	$Var1$	$Var2$		
1	5	9	-4	16
2	7	4	3	9
n	12	12	0	0

Le coefficient de Spearman est défini par :

$$r_S = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)}$$

Où

$n =$ taille de l'échantillon.

$\sum_i d_i^2 =$ somme des différences au carré.

Le coefficient de Spearman n'est autre que le coefficient de corrélation de Pearson calculé sur les rangs :

$$\begin{aligned} r_{KL} &= \frac{cov(K,L)}{S_K S_L} \\ \hat{r}_{KL} &= \frac{\sum_i (k_i - \bar{k})(l_i - \bar{l})}{\sqrt{\sum_i (k_i - \bar{k})^2} \sqrt{\sum_i (l_i - \bar{l})^2}} \end{aligned}$$

Le fait que les rangs soient des permutations de $[1...n]$ simplifie les calculs :

$\bar{k} = \bar{l} = \frac{1}{n} \sum_{i=1}^{n} i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$

$S^2(K) = S^2(L) = \frac{n^2-1}{12}$.

on a

$$r_S = \frac{\frac{1}{n}\sum_i k_i l_i - \bar{k}\bar{l}}{S_K S_L}$$

d'où :

$$r_S = \frac{\frac{1}{n}\sum_i k_i l_i - (\frac{n+1}{2})^2}{\frac{n^2-1}{12}}$$

Si on pose $d_i = k_i - l_i$ différence des rangs d'un même objet selon les deux classements on a

$$\begin{aligned}
\sum_{i=1}^{n} k_i l_i &= -\frac{1}{2}\sum_i -2k_i l_i = -\frac{1}{2}\sum_i \left[(k_i - l_i)^2 - k_i^2 - l_i^2\right] \\
&= -\frac{1}{2}\sum_i (k_i - l_i)^2 + \frac{1}{2}\sum_i k_i^2 + \frac{1}{2}\sum_i l_i^2 \\
&= -\frac{1}{2}\sum_i (k_i - l_i)^2 + \sum_i k_i^2
\end{aligned}$$

mais :

$$\sum_{i=1}^{n} k_i^2 = \sum_i l_i^2 = \frac{n\,(n+1)\,(2n+1)}{2}$$

somme des carrés des nombres entiers, d'où

$$\begin{aligned}
r_{KL} &= \frac{-\frac{1}{2n}\sum_i d_i^2 + \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}} \\
r_{KL} &= -\frac{6\sum_i d_i^2}{n(n^2-1)} + \frac{2(n+1)(2n+1) - 3(n+1)^2}{n^2-1} \\
&= 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = r_S
\end{aligned}$$

La définition de r_s comme coefficient de corrélation linéaire sur des rangs nous indique que :

$r_s = 1 \Rightarrow \sum_i d_i^2 = 0 \Rightarrow k_i = l_i \ \forall i \Rightarrow$ 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.

2.5.2 Signification d'un coefficient de Pearson

En présence d'un échantillon de n couples de rangs (k_i, l_i) $i = 1...n$ obtenus, soit par observation directe d'un couple (K, L) de rangs, soit par transformation en rangs des valeurs d'un couple (X, Y) de valeurs réelles :

Lorsque $n < 100$, on se rapportera à la table du coefficient de corrélation de Spearman.

La région critique est $\mid R_s \mid > S$:

Si $\mid R_s \mid > S$: il y a concordance de classements ;

Si $\mid R_s \mid < -S$ il y a discordance de classements.

Exemple 6 *Mettons en relation la taille et le poids de 15 personnes qu'on ordonnera de manière croissante comme suit :*

$Numéro$	$Tailles$	$Poids$	$k_i(taille)$	$l_i(poids)$	d_i^2
1	1.697	77.564	15	14	1
2	1.539	55.000	3	1	4
3	1.629	76.657	10	12	4
4	1.633	62.596	11	6	25
5	1.500	58.068	2	3	1
6	1.679	72.575	11	11	9
7	1.643	82.000	15	15	4
8	1.626	76.667	13	13	16
9	1.543	58.060	2	2	9
10	1.542	71.668	4	10	36
11	1.621	68.039	8	8	0
12	1.577	70.060	7	9	4
13	1.557	61.689	6	1	1
14	1.496	67.585	1	7	36
15	1.637	59.874	12	4	64
				$somme$	214

A partir de la dernière formule, on a :

$$r_S = 0.6179$$

Au risque $\alpha = 0.05$*, la valeur critique, d'après la table du coefficient de corrélation de Pearson, est :*$S = 0.521$

On a $r_S > S$ *, alors il y a concordance des classements.*

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

Principe et méthode

Si l'on demande à deux enseignants de ranger, par exemple, quatre dissertations (a, b, c, d) en fonction de la qualité de leur style. Leur classement est le suivant :

$Dissertation$	a	b	c	d
$EnseignantA$	3	4	2	1
$EnseignantB$	3	1	4	2

Lorsque les dissertations sont réarrangées de telle sorte que celles de l'enseignant 1 apparaissent rangées dans l'ordre naturel, le tableau devient :

$Dissertation$	d	c	a	b
$EnseignantA$	1	2	3	4
$EnseignantB$	2	4	3	1

Il faut alors déterminer combien de paires de rangs de l'enseignant B sont dans un ordre naturel l'un par rapport à l'autre. Ainsi, les rangs de la première paire 2 et 4 sont dans l'ordre naturel, 2 précède 4. On affecte alors la valeur + 1 à cette paire. Les rangs de la seconde paire 2 et 3 sont dans un ordre correct et obtiennent + 1. La troisième paire (2 et 1) n'est pas dans un ordre correct et reçoit la valeur - 1. Il faut alors considérer toutes les paires qui incluent le rang 4, puis le rang 3 et cette démarche nous permet de calculer la somme de tous les scores obtenus :

(+1) + (+1) + (-1) + (-1) + (-1) + (-1) = -2

Maintenant, le total maximum possible qui peut être atteint par les scores affectés à l'ensemble des paires de jugements de l'enseignant B est obtenu lorsque tous les jugements des deux enseignants sont en parfait accord. Ce total maximum est le résultat de la combinaison de quatre choses prises deux à deux = 6.

Le degré de relation existant entre les deux séries de rangs est alors indiqué par le rapport du total des scores des rangements du juge B au total maximum possible :

$$\tau = \frac{\text{total des scores}}{\text{total maximum possible}} = \frac{-2}{6} = -0.33$$

Le total maximum de combinaisons de n objets pris deux à deux est $\frac{1}{2n(n-1)}$,

On définit le coefficient de Kendall τ par la formule suivante :

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

où n = le nombre d'objets ou d'individus rangés dans les deux séries.

Le calcul de S peut être simplifié de la façon suivante. Quand les rangs d'un des juges sont dans l'ordre naturel, et que les rangs correspondants de l'autre juge sont dans le même ordre, la valeur de S est déterminée en partant du premier nombre sur la gauche et en comptant le nombre de rangs sur sa droite qui lui sont supérieurs et en soustrayant de ce nombre, le nombre de rangs sur sa droite qui sont inférieurs. Ainsi, lorsque les rangs de l'enseignant B sont 2, 4, 3, 1, à la droite du rang 2 sont les rangs 3 et 4 qui sont supérieurs et le rang 1 qui est inférieur. Le rang 2 contribue donc (+2 -1) = +1 à S. Pour le rang 4, aucun rang à sa droite n'est supérieur, mais deux (les rangs 3 et 1) sont inférieurs. Le rang 4 contribue donc de (0 - 2) = - 2 à S. Pour le rang 3, aucun rang sur la droite n'est supérieur, mais un (le rang 1) est inférieur, et donc le rang 3 participe de (0 - 1) = - 1 à S. Leur participation totale à S est donc :

$S = (+1) + (-2) + (-1) = -2$

Connaissant la valeur de S, il est possible de calculer la valeur observée de :

$$\tau = \frac{2S}{n(n-1)} = \frac{-4}{4(4-1)} = -0.33$$

2.6 Coefficient de Daniels et de Guttman

Les trois coefficients de corrélations (Pearson, Spearman, Kendall) peuvent être présentés comme 3 cas particuliers d'une même formule, dite formule de Daniels.

On considère pour toute paire d'individus i,j deux indices a_{ij} et b_{ij} le premier associé à la variable X , le deuxième associé à la variable Y (par exemple $a_{ij} = x_i - x_j$) et on définit le coefficient suivant :

$$\frac{\sum\sum a_{ij}b_{ij}}{\sqrt{\left(\sum\sum a_{ij}^2\right)\left(\sum\sum b_{ij}^2\right)}}$$

qui varie entre -1 et +1 d'après l'inégalité de Schwarz.

En prenant :

$a_{ij} = x_i - x_j$ et $b_{ij} = y_i - y_j$ on trouve le coefficient r de Bravais-Pearson ($\sum\sum (x_i - x_j)^2 = 2n^2 S_x^2$) .

$a_{ij} = k_i - k_j$ et $b_{ij} = l_i - l_j$ où les k et les l sont les rangs de classement selon X et Y on obtient le coefficient de Spearman .

$$\begin{aligned} a_{ij} &= \textit{signe de } (x_i - x_j) = \frac{x_i - x_j}{\mid x_i - x_j \mid} \\ b_{ij} &= \textit{signe de } (y_i - y_j) \end{aligned}$$

on obtient le coefficient τ de Kendall.

Chapitre 3

Analyse en composantes principales

3.1 Tableau des donnée et espace associé

3.1.1 Introduction

L'objectif de l'analyse en composantes principales ACP est de fournir un outil de visualisation des données. Elle permet de faire une réduction de la dimension et aussi explorer les liaisons entre variables et ressemblance entre individus.

3.1.2 Les données et leurs caractéristiques

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

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

La variable $X^j = \begin{pmatrix} x_1^j \\ . \\ x_i^j \\ . \\ x_n^j \end{pmatrix}$

x_i^j est la valeur prise par la variable j sur l'individu i ;

X^j liste les n valeurs de la variable j qu'elle prend sur les n individus ;

L'individu $e_i = (x_i^1, ..., x_i^p)$ liste les p valeurs qu'il prend sur les p variables.

3.1.3 Principe de l'ACP

On cherche une représentation des n individus $e_1, e_2, ...e_n$ dans un espace F_k de $\mathbb{R}^p$ tel que k soit le plus petit possible, c'est-à-dire , on cherche à définir k nouvelles variables combinaison linéaire des p variables initiales contenant le plus d'informations possible.

Les k variables sont appelées composantes principales ;

les axes qu'elles déterminent sont appelés axes principaux ;

les formes linéaires associées sont appelées facteurs principaux.

3.1.4 Matrice des poids

Si les données ont été recueillies d'un tirage aléatoire alors les probabilités des n individus ont toutes la même importance égale à $\frac{1}{n}$. Or ceci n'est pas toujours le cas, donc il est utile de travailler avec les poids des différents individus et de les regrouper dans une matrice.D appelée matrice des poids :

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

3.1.5 Le point moyen ou centre de gravité

Le vecteur g ou centre de gravité est le vecteur des moyennes arithmétiques de chaque variable.

$$g^t = (\bar{x}^1, ..., \bar{x}^p)$$

avec $\bar{x}^j = \frac{1}{n} \sum_{i=1}^{n} x_i^j$

on a

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

où

1_n désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1, et D représente la matrice des poids.

Soit Y le tableau centré associé à la matrice X défini par :

$$y_i^j = x_i^j - \bar{x}^j$$

On a

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

3.1.6 Matrice de covariance et matrice de corrélation

La formule de la matrice de covariance $V = \frac{1}{n} Y^t Y$ établie au chapitre précédent avec des poids égaux à $1/n$ se généralise comme suit :

$$\begin{aligned} V &= X^t DX - gg^t = X^t DX - X^t D 1_n 1_n^t DX \\ &= Y^t DY \end{aligned}$$

sachant que $X^t DX = \sum_{i=1}^{n} p_i e_i e_i^t$.

Si on note $D_{\frac{1}{S}}$ la matrice diagonale des inverses des écarts types.

$$D_{\frac{1}{S}} = \begin{bmatrix} \frac{1}{S_1} & & 0 \\ & & \\ & & \\ 0 & & \frac{1}{S_p} \end{bmatrix}$$

et $D_{\frac{1}{S^2}}$ la matrice diagonale des inverses des variances, le tableau des données centrées et réduites Z telles que :

$$z_i^j = \frac{x_i^j - \bar{x}^j}{S_j}$$

va s'écrire :

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

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables est noté C.

Rappelons que $C = D_{\frac{1}{s}} V \ D_{\frac{1}{s}} = Z^t DZ$.

3.2 Espace des individus

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

3.2.1 rôle de la métrique

Problème ; quelle métrique choisir ?

Pour mesurer la distance entre deux individus A et B. telles que :

$A = (x_A, y_A)$ et $B = (x_B, y_B)$,on utilise la formule suivante :

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

Dans des espaces $\mathbb{R}^p$, on généralise cette notion . La distance euclidienne entre deux individus (e_i, e_j) s'écrit :

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

Reste le problème des unités. Pour le résoudre, on choisit de transformer les données en données centrées et réduites.

La formule de Pythagore n'est valable que si les axes sont perpendiculaires, mais en statistique ce n'est que par pure convention que l'on présente les caractères par des axes perpendiculaires, on utilise donc la formulation générale suivante :

$$d^2(e_i, e_j) = (e_i - e_j)^t M (e_i - e_j)$$

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

$$< e_i, e_j >= e_i^t M e_j$$

En théorie M dépend de l'utilisateur. En pratique $M = I$ revient à utiliser le produit scalaire canonique. La métrique la plus utilisée et qui est souvent l'option par défaut des logiciels est la métrique $M = D_{\frac{1}{S^2}}$ ce qui revient à diviser chaque caractère par son écart-type.Entre autres avantages, la distance entre deux individus ne dépend plus des unités de mesure, ce qui est trés utile lorsque les variables ne s'expriment pas avec les mêmes unités.

3.2.2 Inertie et inertie totale

On appelle inertie totale du nuage de points la moyenne pondérée des carrés des distances des points $(e_i)_{i=1..p}$ par rapport au centre de gravité g :

$$\begin{aligned} I_g &= \sum_{i=1}^{p} p_i d^2(e_i, g) \\ &= \sum_{i=1}^{p} p_i (e_i - g)^t M (e_i - g) \\ &= \sum_{i=1}^{p} p_i \parallel e_i - g \parallel^2 \end{aligned}$$

où $p_i > 0$,et ils sont appelés poids avec $\sum p_i = 1$.

L'inertie totale I_g est la trace de la matrice MV, où V est la matrice de variance.

En effet

$$I_g = \sum_{i=1}^{p} p_i \parallel e_i - g \parallel^2$$

Si g=0

$$I_g = \sum_{i=1}^{p} p_i e_i^t M e_i$$

Comme $.p_i e_i^t M e_i$ est un scalaire, et grâce à la commutativité sous la trace

$$\begin{aligned} I_g &= trace(\sum_{i=1}^{p} p_i e_i^t M e_i) \\ &= trace(\sum_{i=1}^{p} M e_i p_i e_i^t) \\ &= trace(M X^t D X) \\ &= trace(MV) \end{aligned}$$

Avec $X^t D X = \sum_{i=1}^{p} p_i e_i^t M e_i$

Si $M = I$ l'inertie est égale à la somme des variances des p variables

$$\begin{aligned} I_g &= \sum_{i=1}^{p} S_i^2 \\ &= trace(V) \end{aligned}$$

Si $M = D_{\frac{1}{S^2}}$ alors

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

.L'inertie est donc égale au nombre de variables alors :

$$I_g = p = nbr\ de\ variables$$

3.3 ACP-axes principaux, composantes principales,facteurs principaux

Comme c'était déjà précisé, le critère du choix de l'espace de projection s'effectue tel que la moyenne des carrées des distances entre les projections soit la plus grande possible. Ce qui implique qu'il faut que l'inertie du nuage projeté sur ce sous espace soit maximale.

Pour cela on définit P un projecteur M-orthogonal sur l'espace F_k tel que

$$P^2 = P \ \text{ et } P^t M = MP$$

Ce nuage projeté est associé au tableau XP^t(sur chaque individu e_i(ligne de X) on aura un vecteur projeté $e_i P^t$)

La matrice de covariance du tableau (XP^t) est $(XP^t)^t\, D\, (XP^t) = PVP^t$

L'inertie du nuage projeté vaut donc $trace(PVP^tM)$

On a

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

Le problème est donc de trouver P le projecteur M-orthogonal de rang k maximisant $trace(VMP)$ ce qui déterminera F_k.

3.3.1 Les axes principaux

Nous devons chercher une droite de $\mathbb{R}^p$ passant par le centre de gravité g, et maximisant l'inertie du nuage projeté sur cette droite.

Soit a un vecteur porté par cette droite.

Le projecteur M-orthogonal sur cette droite est :

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

L'inertie du nuage projeté sur cette droite vaut d'après ce qui précède :

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

La matrice MVM est dite matrice d'inertie du nuage, elle définit la forme quadratique d'inertie qui à tout vecteur $\vec{a}$ de M-norme=1, associe l'inertie projeté sur cet axes .

La matrice se confond avec la matrice de covariance ssi $M = I$.

Pour obtenir le maximum de $\frac{a^tMVMa}{a^tMa}$, on doit dériver et annuler par rapport à a :

$$\frac{d}{da}(\frac{a^tMVMa}{a^tMa}) = \frac{(a^tMa)2MVMa - (a^tMVMa)2Ma}{(a^tMa)^2}$$

on obtient

$$\begin{aligned}(MVMa)(a^tMa) &= (a^tMVMa)Ma \\ &\Rightarrow \quad MVMa = \frac{(a^tMVMa)Ma}{a^tMa}\end{aligned}$$

Puisque M est régulière

$$VMa = \lambda a$$

a est donc vecteur propre de matrice VM et λ sa valeur propre.

Le premier axe est celui qui va correspondre à la plus grande valeur propre.

Le deuxième axe est celui de la deuxième valeur propre.

On appelle axes principaux d'inertie les vecteurs propres de VM normé à 1,ils sont au nombre de p.

3.3.2 Facteurs principaux :

Â l'axe principal a est M-normé à 1 est associé le facteur principal $u = Ma$.

Puisque a est un vecteur propre de VM , on a :

$$\begin{aligned} VMa &= \lambda a \\ &\Leftrightarrow MVMa = \lambda Ma \\ &\Leftrightarrow MVu = \lambda u \end{aligned}$$

Donc les facteurs principaux u seront aussi les vecteurs propres de la matrice MV.

3.3.3 Composantes principales :

A chaque axe est associée une variable appelée composante principale.

La composante principale C_i est définie par les facteurs principaux

$$\begin{aligned} C_i &= Xu_i \\ C_1 &= X_1^1 u_1 + ... + X_p^1 u_p \end{aligned}$$

C_1 est le vecteur renfermant les coordonnées des projecteurs des individus sur l'axe 1

C_2 est le vecteur renfermant les coordonnées des projecteurs des individus sur l'axe 2.

Propriétés d'une composante principale :

$$V(C_i) = \lambda_i$$

En effet :

$$\begin{aligned} V(C) &= C^t DC \\ &= U^t X^t DXU \\ &= U^t VU \\ &= \lambda U^t M^{-1} U \\ &= \lambda \end{aligned}$$

Les C_i sont non corrélées deux à deux, car les axes associés sont orthogonaux.

Les composantes principales sont elles-mêmes vecteurs propres d'une matrice de taille n :

En effet :

$MVu = \lambda u$ s'écrit $MX^t DXu = \lambda u$

En multipliant à gauche par X et en remplaçant Xu par C on a :

$XMX^t DC = \lambda C$

La variance d'une composante principale est égale à l'inertie apportée par l'axe principal qu'il est associé.

Représentation des individus

$$C^j = \begin{bmatrix} C_1^j \\ . \\ . \\ C_n^j \end{bmatrix}$$

La j^{ème} composante principale fournit les coordonnés des n individus sur le j^{ème} axe principale.

-Si on désire une représentation plane des individus, la meilleure sera celle réalisée grâce aux deux premières composantes principales.

Dans le cas où, on travaille avec un tableau centré réduit Z associé à X, on utilisera la métrique $M = 1$, ce qui implique que la matrice de variance-covariance et les facteurs principaux sont tout simplement les vecteurs propres de la matrice de corrélation C rangé selon l'ordre décroissant des valeurs propres .

3.4 Interprétation et qualité des résultats d'une ACP

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

3.4.1 Interprétation

La méthode la plus naturelle pour donner une signification à la composante principale est de la relier aux variables initiales X^j, en calculant son coefficient de corrélation

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

comme $V(C) = \lambda$ alors :

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

On calcule pour chaque composante principale cette corrélation.

$$\begin{aligned} r(X^j, C) &= \frac{1}{\sqrt{\lambda}}(Z^j)^t DC \\ &= \frac{1}{\sqrt{\lambda}}(Z^j)^t DZU \\ &= \frac{1}{\sqrt{\lambda}}C^j U \\ &= \sqrt{\lambda}U_j. \end{aligned}$$

$(Z^j)^t DZ$ représente la $j^{ème}$ ligne de $Z^t DZ$,donc $(Z^j)^t DZU$ est la j^{ème} composante de CU.

3.4.2 Importance des individus

Dire que C_1 est très corrélée avec X^j signifie que les individus ayons une forte coordonnée positive sur l'axe 1 sont caractérisés par une valeur de la variable x^j nettement supérieure à la moyenne.

Il est très utile de calculer pour chaque axe la contribution apportée par les divers individus à cet axe .

La contribution de l'individu i à la composante C_k est définie par :

$$\frac{p_i c_{ki}^2}{\lambda_k}$$

avec c_{ki} : est la valeur de la composante c_k pour le ième individus et $\lambda_k = \sum_{i=1}^{n} p_i c_{ki}^2$.

3.4.3 Qualité de la représentation sur les plans principaux

Le but de l'ACP est d'obtenir la représentation des individus dans un espace de dimension plus faible que p, les questions qui se posent sont : quel degré de perte d'information doit on subir ainsi de savoir combien de facteurs va-t-on retenir

Le critère globale

$$\frac{\lambda_i}{\lambda_1 + ... + \lambda_p}$$

cette valeur mesure l'inertie expliqué par l'axe i

On mesure la qualité de F_k par :

$$\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p}$$

3.4.4 Nombre d'axe à retenir

Différentes procédures sont complémentaires :

Règle de Kaiser

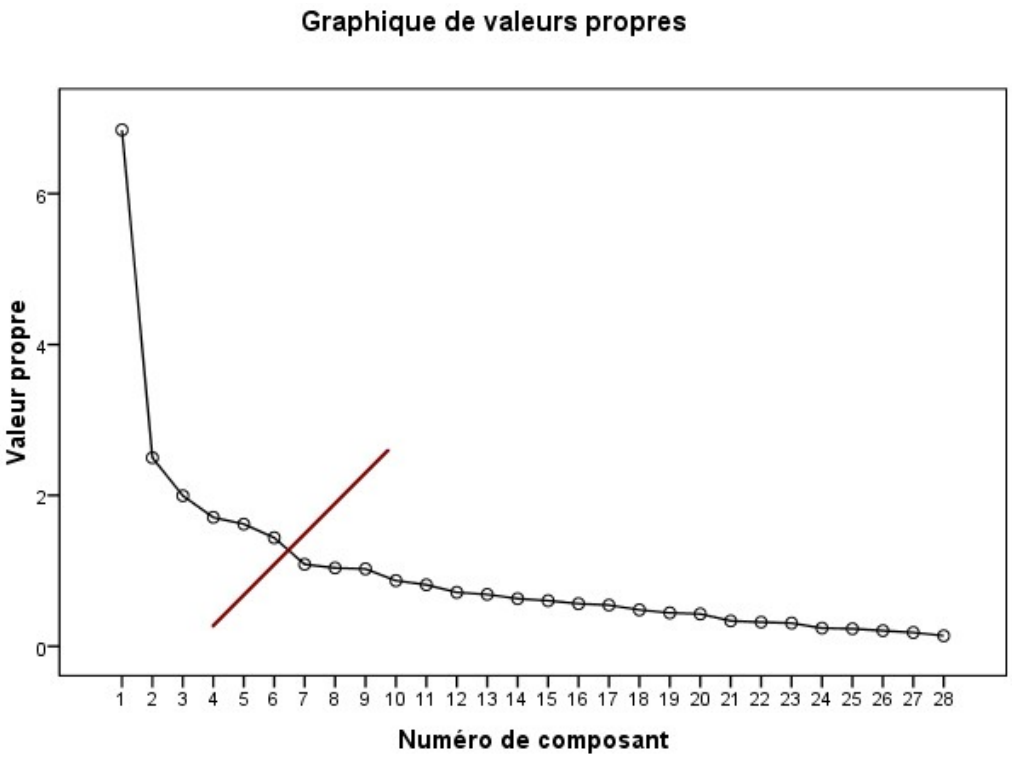
On considère que, si tous les éléments de Y sont indépendants, les composantes principales sont toutes de variances égales (égales à 1 dans le cas de l'ACP réduite). On ne conserve alors que les valeurs propres supérieures à leur moyenne car seules jugées plus "informatives" que les variables initiales ; dans le cas d'une ACP réduite, ne sont donc retenues que celles plus grandes que 1.

Part d'inertie

Diviser l'inertie totale par le nombre de variables initiales, ce qui donne l'inertie moyenne par variable notée $I.M.$

Il faut conserver tous les axes apportant une inertie supérieure à cette valeur $I.M.$

Si les variables sont centrées et réduites, on prend l'inertie supérieure à 1.



Critère de Cattel

Le critère de Cattel (la règle du coude) préconise de détecter sur un diagramme des valeurs propres, l'éxistence du coude

On doit conserver les axes associés aux valeurs propres situées avant le coude

Chapitre 4

Analyse canonique et la comparaison de deux groupes de variables

4.1 Introduction

Lorsque n individus sont décrits par deux ensembles de variables en nombre p et q respectivement, on cherche à examiner les liens existant entre ces deux ensembles afin de savoir s'ils mesurent les mêmes propriétés.

4.2 Les données

On observe sur n individus p variables quantitatives ($p \leq n$), et q autres variables quantitatives ($q \leq n$).

On appelle X_1 de $dim(n \times p)$, le tableau des variables explicatives, et X_2 de $dim(n \times q)$ celui des variables à expliquer par X_1 ou variables dépendantes.

Le tableau des données est donc de la forme suivante :

$$X = \left[\begin{array}{c|c|c} & | & \\ & | & \\ X_1 & | & X_2 \\ & | & \\ & | & \\ & | & \end{array}\right]$$

On suppose que les variables X_1 et X_2 sont centrées.

On note alors les deux sous-espaces de $\mathbb{R}^n$ engendrés par les colonnes de $X_1 et\ X_2$ respectivement :

$$W_1 = \{x \mid x = X_1 a\} \quad et \quad W_2 = \{y \mid y = X_2 b\}$$

$W_1 et W_2$ sont les deux ensembles de variables que l'on peut construire par combinaisons linéaires des deux groupes. Ces deux espaces sont souvent appelés << *potentiels* de prévision>> .

Si ces deux espaces sont confondus cela prouve que l'on peut se contenter d'un seul des deux ensembles de variables, car ils ont le même pouvoir de description. S'ils sont orthogonaux, c'est que les deux ensembles de variables appréhendent des phénomènes totalement différents. On étudiera les positions géométriques de W_1 et W_2 en cherchant les éléments les plus proches, ce qu'il permettra de connaître $\dim(W_1 \cap W_2)$

Les applications directes de l'analyse canonique sont peu nombreuses, elle n'en constitue pas moins une méthode fondamentale (rechercher des couples de variables en corrélation maximale).

4.2.1 But de l'analyse canonique

On cherche à expliquer le groupe de variables X_2 par le groupe de variables X_2,ou juste décrire les ressemblances entre ces deux groupes.

4.2.2 Recherche des variables canoniques

On supposera que $\mathbb{R}^n$ est muni de la métrique D. La technique est alors la suivante :

chercher le couple (ξ_1, η_1) de vecteurs normés où $\xi_1 \in W_1$ *et* $\eta_1 \in W_2$ telle que :

$$\begin{aligned} \xi_1 &= a_{11}X_{11} + \cdots + a_{1p}X_{1p} = X_1 a_1 \\ \eta_1 &= b_{11}X_{21} + \cdots + b_{1q}X_{2q} = X_2 b_1 \end{aligned}$$

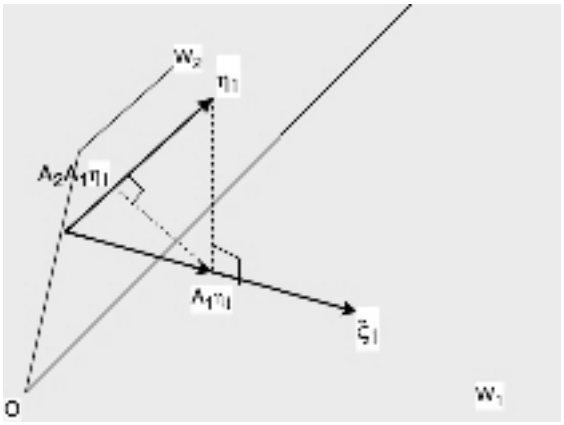
$\xi_1$et η_1 sont des combinaisons linéaires respectives de X_1et X_2 telle que .$\xi_1$et $\eta_1$soit les plus corrélés possible.

Remarque

Les vecteurs a et b appelés premiers facteurs ne sont pas uniques. Pour assurer leurs unicités, on impose à $\xi_1$et η_1 d'être de variance unité.

La corrélation r_1 entre $\xi_1$et η_1 est appelée première corrélation canonique, et $\xi_1$et η_1 sont appelés les premières variables canoniques.

En générale, $\xi_1$et η_1 n'expliquent pas l'ensemble des liaisons entre les X_1et les X_2,on cherche alors de nouvelle variables $\xi_2$et η_2 telle que ξ_2 D-orthogonale à ξ_1 et η_2 D-orthogonale à η_1 de corrélation maximale et de variance unité.



$$\begin{aligned}\xi_2 &= a_{21}X_{11} + \cdots + a_{2p}X_{1p} = X_1 a_2 \\ \eta_2 &= b_{21}X_{21} + \cdots + b_{2q}X_{2q} = X_2 b_2\end{aligned}$$

On obtient ainsi les p couples de variables canoniques et une suite de corrélation canonique décroissante $r_1 \geq r_2 \geq \cdots \geq r_p$ (on posera $p = \dim W_1$ et $q = \dim W_2$ avec $p \leq q$).

Notons A_1 et A_2 les opérateurs de projection D-orthogonale sur W_1 et W_2 resp

.

Rappel d'algèbre linéaire

Le projecteur orthogonal sur l'espace E engendré par les colonnes de X est l'application linéaire qui fait correspondre à u sa projection orthogonale sur E .Ce projecteur s'écrit

$$P = X(X^tX)^{-1}X^t$$

Les expressions matricielles de explicites de A_1 et A_2 sont

$$\begin{aligned} A_1 &= X_1(X_1^t D X_1)^{-1} X_1^t D \\ A_2 &= X_2(X_2^t D X_2)^{-1} X_2^t D \end{aligned}$$

ETUDE DE LA SOLUTION DANS $\mathbb{R}^n$

Il s'agit de chercher le premier couple de variable canoniques (ξ_1,η_1) tq $r(\eta_1, \xi_1)$ soit maximal. avec

$$r(\xi_1, \eta_1) = \cos(\xi_1, \eta_1)$$

En supposant pour l'instant que $\eta_1$et ξ_1 ne sont pas confondus, on voit géométriquement que $\eta_1$doit être tel que $A_1\eta_1$ sa projection sur W_1 soit colinéaire à ξ_1. En effet l'élément le plus proche de η_1 est la projection D-orthogonale de η_1 sur W_1.

Réciproquement, η_1 doit être l'élément de W_2 le plus proche de ξ_1, donc η_1 doit être colinéaire à $A_2A_1\eta_1$.

Notre problème revient donc à trouver les valeurs et les vecteurs propres de A_2A_1 puisque $A_2A_1\eta_1 = \lambda_1\eta_1$.

Inversement, ξ_1 est un vecteur propre de A_1A_2 associé à la même valeur propre.

λ_1 représente le carré du cosinus de l'angle formé par $\eta_1$et ξ_1,ce qui entraine $\lambda_1 \leq 1$.

Le cas $\lambda = 1$ nous donne $\xi_1 = \eta_1$,donc $\eta_1 \in W_1 \cap W_2$.

Les vecteurs propres de A_2A_1 appartiennent à W_2 :

En effet, en pré-multipliant $A_2A_1\eta_1 = \lambda_1\eta_1$ par A_2 on trouve puisque $A_2^2 = A_2$,

$$A_2A_1\eta_1 = \lambda_1 A_2\eta_1$$

donc

$$A_2\eta_1 = \eta_1$$

On trouve de même que les vecteurs propres de A_1A_2 appartiennent à W_1.

Montrons que A_2A_1 est diagonalisable.

Puisque les vecteurs propres de A_2A_1 appartiennent à W_2,il suffit d'étudier la restriction de A_2A_1 à W_2.

Théorème 7 *La restriction de A_2A_1 à W_2 est D-symétrique.*

Si nous notons $< x; y >$ le produit scalaire associé à la métrique D :

$$< x \ ; \ y >= x^t Dy$$

il faut monter que quel que soit $x, y \in W_2$:

$$< x \ ; \ A_2A_1y >=< A_2A_1x \ ; \ y >$$

on a :

$$\begin{aligned} < \ \ x \ ; \ A_2A_1y >=< A_2x \ ; A_1 \ y > \\ = \ \ < x \, ; \, A_1y > \\ = \ \ < A_1x \, ; \, y > \\ = \ \ < A_1x \, ; \, A_2y > \\ = \ \ < A_2A_1x \ ; \ y > \qquad c.q.f.d \end{aligned}$$

Ceci entraîne que la restriction de A_2A_1 est D-symétrique, et par suite A_2A_1 est diagonalisable, ses vecteurs propres sont D-orthogonaux et ses valeurs propres λ_i sont réelles et supérieures ou égales à zéro($\lambda_i \geq 0$).

A_2A_1 possède au plus $\min(p, q)$ valeurs propres non identiquement nulles. L'ordre de multiplicité de $\lambda_1 = 1$ est alors la dimension de $W_1 \cap W_2$; les vecteurs propres associés à des valeurs propres nulles de rang inférieur à q engendrent la partie de W_2 D-orthogonale à W_1.

Les vecteurs propres ξ_i et η_i D-normés de A_2A_1 et de A_1A_2 sont associés aux même valeurs propres et vérifient les relations suivantes :

$$\begin{aligned}
A_2A_1\eta_i &= \lambda_i\eta_i \\
A_1A_2\xi_i &= \lambda_i\xi_i \\
\sqrt{\lambda_i}\eta_i &= A_2\xi_i \\
\sqrt{\lambda_i}\xi_i &= A_1\eta_i \\
\eta_i^t D\eta_j &= 0 \quad \text{et} \quad \xi_i^t D\xi_j = 0 \quad \textit{pour} \quad i \neq j
\end{aligned}$$

qui entraîne de plus :

$$\eta_i^t D\xi_j = 0 \quad \text{pour } i \neq j$$

Solutions dans $\mathbb{R}^p et\ \mathbb{R}^q$

Les variables canoniques ξ_i et η_i s'expriment comme combinaisons linéaires des colonnes de X_1et X_2 respectivement :

$$\begin{aligned} \xi_i &= X_1 a_i \\ \eta_i &= X_2 b_i \end{aligned}$$

Les a_i et b_i sont les facteurs canoniques qui s'obtiennent directement de la manière suivante :

$$A_1 A_2 \xi_i = \lambda_i \xi_i \iff A_1 A_2 X_1 a_1 = \lambda_i X_1 a_i$$

En remplaçant les projecteurs par leur expression on a :

$$X_1 (X_1^t D X_1)^{-1} X_1^t D X_2 (X_2^t D X_2)^{-1} X_2^t D X_1 a_i = \lambda_i X_1 a_i$$

En multipliant par $(X_1^t X_1)^{-1} X_1^t$,on trouve :

$$(X_1^t D X_1)^{-1} X_1^t D X_2 (X_2^t D X_2)^{-1} X_2^t D X_1 a_i = \lambda_i a_i$$

et de même

$$(X_2^t D X_2)^{-1} X_2^t D X_1 (X_1^t D X_1)^{-1} X_1^t D X_2 b_i = \lambda_i b_i$$

Les matrices $X_i^t D X_j$ s'interprètent comme des matrices de covariance, on note :

$$\begin{aligned} V_{22} &= X_2^t D X_2 & V_{21} &= X_2^t D X_1 \\ V_{11} &= X_1^t D X_1 & V_{12} &= X_1^t D X_2 \end{aligned}$$

Les équations des facteurs canoniques s'écrivent alors :

$$\begin{aligned} V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}a_i &= \lambda_i a_i \\ V_{22}^{-1}V_{21}V_{11}^{-1}V_{12}b_i &= \lambda_i b_i \end{aligned}$$

Comme on a :$\xi_1 = X_1 a_i$ et $\eta_i = X_2 b_i$,si on désire que les variables canoniques soient de variance unité, on nommera les facteurs principaux de la manière suivante :

$$a_i^t V_{11} a_i = 1 \quad \text{et} \qquad b_i^t V_{22} b_i = 1$$

On en déduit :

$$b_i = \frac{1}{\sqrt{\lambda_i}} V_{22}^{-1} V_{21} a_i \qquad \text{et} \qquad a_i = \frac{1}{\sqrt{\lambda_i}} V_{11}^{-1} V_{12} b_i$$

4.3 Analyse canonique généralisée

Étendre l'analyse canonique à plus de deux groupes de variables mène à la difficulté suivante :

il n'existe pas de mesure simple de la liaison entre plus de deux variables. Il y aura donc autant de façon d'obtenir des variables canoniques que de manières de définir une ≪*corréllation*≫entre p variables.

On peut prendre par exemple comme mesure la somme des corrélations deux à deux, la somme des carrés de corrélations. Toute généralisation est donc plus ou moins arbitraire. Celle que nous présentons ici a l'avantage d'être simple et la plus riche d'interprétations, car elle se relie à toutes les autres méthodes d'analyse des données.

4.3.1 Propriété de l'analyse canonique ordinaire

Soit deux ensembles de variables centrées X_1 et X_2 , les variables canoniques ξ et η vecteurs propres de A_1A_2 et A_2A_1 respectivement, possèdent la propriété suivante :

$$\xi + \eta \text{ est vecteur propre de } A_1 + A_2$$

En effet, posons z tel que $(A_1 + A_2)\, z = \mu z$, en pré-multipliant par A_1 ou A_2 on trouve :

$$A_1 (A_1 + A_2)\, z = \mu A_1 z$$

Soit :

$$A_1A_2z = (\mu - 1)A_1z \qquad \text{et} \qquad A_2A_1z = (\mu - 1)A_2z$$

ce qui donne

$$\begin{aligned} A_1A_2A_1z &= (\mu - 1)^2 A_1z \\ A_2A_1A_2z &= (\mu - 1)^2 A_2z \end{aligned}$$

Donc au même coefficient multiplicateur prés,A_1z et A_2z ne sont autres que les variables canoniques ξet η , comme $(A_1 + A_2)\, z = \mu z$ on trouve $\mu z = \xi + \eta$ ce qui montre la propriété annoncée.

Le coefficient de corrélation multiple de z avec X_i vaut :

$$R_i^2 = \frac{z^t D A_i z}{z^t D z} = \frac{\parallel A_i z \parallel^2}{\parallel z \parallel^2}$$

Car les variables étant centrées, R_i est le cosinus de l'angle formé par z et W_i.

4.3.2 Généralisation

De la propriété précédente découle la généralisation suivante : plutôt que de chercher directement des variables canoniques dans chacun des sous-espaces W_i associés à des tableaux de données X_i , on chereche une variable auxiliaire z appartenant à la somme des W_i telle que $\sum_{i=1}^{p} R^2(z; X_i)$ soit maximal

z est alors vecteur propre de $A_1 + A_2 + ... + A_p$: $(A_1 + A_2 + ... + A_p)\, z = \mu z$.

on obtient en suite ,des variables canoniques ξ_i en projetant z sur les W_i : $\xi_i = A_i z$.

Si on pose $X = (X_1 \mid X_2 \mid ... \mid X_p)$, matrice à n lignes et $\sum_{i=1}^{p} m_i$ colonnes, la variable z se met sous la forme Xb et au lieu de chercher z comme vecteur propre d'une matrice n,n il vaut mieux chercher b qui possède $\sum_{i=1}^{p} m_i$ composantes.

Comme $A_i = X_i(X_i^t D X_i)^{-1} X_i^t D$,en posant $V_{ii} = X_i^t D X_i$ matrice de variance-covariance du i^{ème} groupe et

$$M = \begin{bmatrix} V_{11}^{-1} & & & \\ & V_{22}^{-1} & & \\ & & & \\ & & & V_{pp}^{-1} \end{bmatrix}$$

la matrice bloc-diagonale des V_{ii}^{-1} on trouve que :

$\sum_{i=1}^{p} A_i = \sum_{i=1}^{p} X_i V_{ii}^{-1} X_i^t D$ s'écrit en fait $\sum_{i=1}^{p} A_i = X M X^t D$

Donc z est vecteur propre de XMX^tD,et puisque $z = Xb$,si X est de plein rang,b est vecteur propre de MX^tDX :

$$XMX^tDz = \mu z$$
$$MX^tDXb = \mu b$$

On reconnaît alors les équations donnant les composantes principales et les facteurs principaux, dans l' ACP du tableau X avec la métrique M.

La généralisation de l'analyse canonique présentée ici est donc équivalente à une ACP, ce qui nous ramène à une optique de description des individus tenant compte des liaisons par bloc plutôt qu'à une optique de déscription des relations entre variables.

En particulier si chaque groupe est réduit à une seule variable ($m_i = 1, i = 1, ...p$),on retrouve l'ACP avec la métrique $D_{\frac{1}{S^2}}$puisque z rend alors maximal $\sum_{i=1}^{p} r^2(z; X^i)$.

4.3.3 Critiques de l'AC

• L'AC décrit les relatons linéaires existant entre 2 ensembles de variables : les premières étapes mettent en évidence les directions de l'espace des variables selon lesquelles les deux ensembles sont les plus proches.

• Mais il est possible que les variables canoniques soient faiblement corrélées aux variables des tableaux X et Y. Donc elles sont difficilement interprétables.

En effet, les variables d'origine n'interviennent pas dans les calculs de détermination des composantes canoniques, seuls interviennent les projecteurs sur les espaces engendrés par ces variables.

4.4 Exemple

On observe les résultats de 5 individus a un premier groupe d'épreuves (2 épreuves contenues dans X), puis a un second (3 épreuves contenues dans Y).

Calcule des corrélations :

	X1	X2	Y1	Y2	Y3
X1	1.0000000000	0.393863181	0.7454993	0.1153846	-0.0003556671
X2	0.3938631807	1.000000000	0.8981462	-0.5012804	-0.0003311080
Y1	0.7454993164	0.898146239	1.0000000	-0.2192645	-0.1062469927
Y2	0.1153846154	-0.501280412	-0.2192645	1.0000000	-0.7853129298
Y3	-0.0003556671	-0.000331108	-0.1062470	-0.7853129	1.0000000000

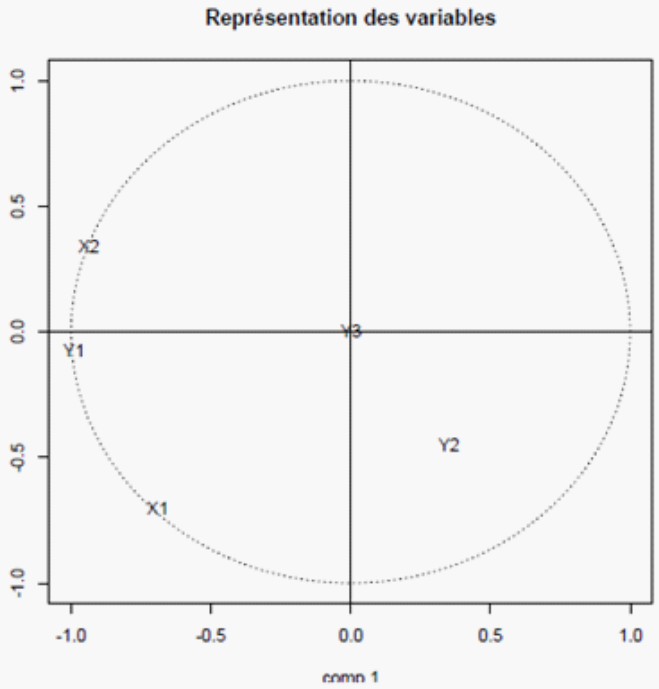
Il existe des corrélations fortes entre X et Y , donc on peut continuer l'analyse

Calcule des matrices R_X et R_Y :

```
Vxx=correl$Xcor
       X1          X2
X1 1.0000000 0.3938632
X2 0.3938632 1.0000000
Vyy=correl$Ycor
        Y1          Y2          Y3
Y1  1.0000000 -0.2192645 -0.1062470
Y2 -0.2192645  1.0000000 -0.7853129
Y3 -0.1062470 -0.7853129  1.0000000
Vxy=correl$XYcor[1:2,3:5]
         Y1          Y2          Y3
X1 0.7454993  0.1153846 -0.0003556671
X2 0.8981462 -0.5012804 -0.0003311080
 Vyx=correl$XYcor[3:5, 1:2]

         X1             X2
Y1  0.7454993164  0.898146239
Y2  0.1153846154 -0.501280412
Y3 -0.0003556671 -0.000331108
```

```
Rx=solve(Vxx)%*%Vxy%*%solve(Vyy)%*%Vyx

       X1          X2
X1 0.95447455 0.0219809
X2 0.03353452 0.9838086

Ry=solve(Vyy)%*%Vyx%*%solve(Vxx)%*%Vxy

           Y1           Y2           Y3
Y1  0.99647821 -0.02026026 -0.0004488666
Y2 -0.01005411  0.94196143 -0.0001387071
Y3  0.09757536  0.73766447 -0.0001564495
```

- Y3 a un comportement qui ne peut pas être prévu par aucune des vraibles du tableau X
- Y2 n'est pas assez proche du bord du cercle pour être interprétable (pas assez bien représenté).
- X1 et dans une moindre mesure X2 sont liés à l'axe 1, de même que Y1. En revanche Y2 et Y3 ne le sont pas. Ainsi, l'axe 1 montre une forte corrélation entre la réussite aux deux épreuves de X et celle de Y1. En terme de prévision, cela veut que le resultat à l'épreuve Y1 peut être prédit par les résultats aux épreuves de X.
- L'axe 2 isole la réussite à X1 du reste



Chapitre 5

Application

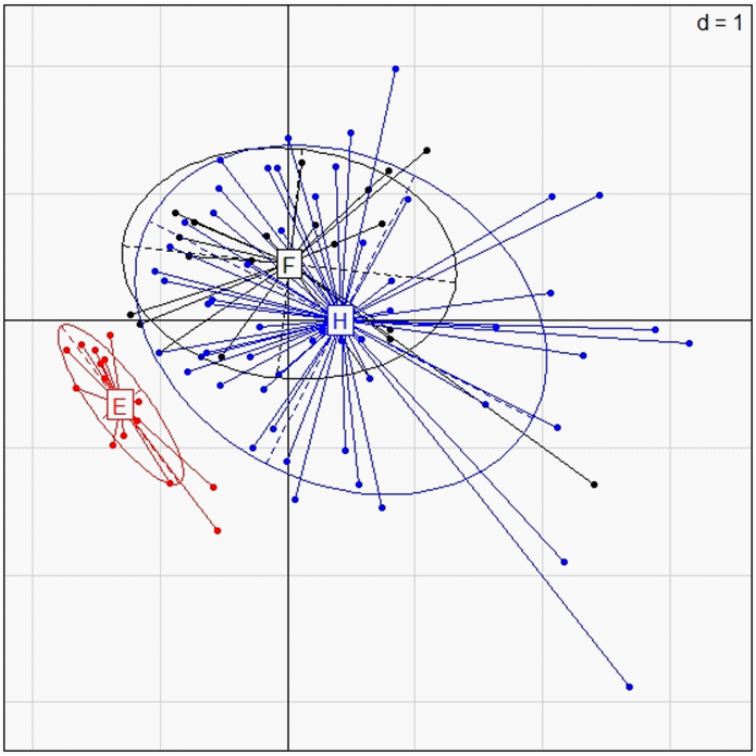
5.1 Introduction

Afin d'illustrer les méthodes statistiques de l'analyse multivariés, nous allons appliquer ces méthodes sur des données réelles sur un groupe de cent 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 l'échantillon des cent cas d'accident les variables suivantes : le type de l'accident, le genre de la personne ayant subi l'accident (homme, femme ou enfant) le taux d'indemnisation ainsi que l'âge de la personne.

Nous avons pensé à appliquer une ACP (Analyse des Composantes Principales) sur ces données pour évaluer la corrélation entre les variables considérées. Cependant, cette méthode nécessite des variables quantitatives ce qui n'est pas notre cas. Pour cela nous avons transformé les variables qualitatives en des valeurs numériques pour appliquer une ACP.

On applique notre ACP sur les données age, indemnité et à la fin nous allons introduire la variable qualitative genre sur le graphique suivant :



On voit que les enfants sont un peu isolé, ils n'ont pas d'intersection avec les hommes et les femmes, donc l' indemnités des enfants est un peu spéciale.

Il y a intersection entre la bulle des femme est celle des hommes,on conclut que les hommes et les femmes sont indemnisés de la même manière.

Ce que nous avons obtenu comme conclusion sur ces données s'avèrent très insuffisant. Ce qui nous a conduit à utiliser d'autres méthodes en particulier une AFC (Analyse Factorielle des Correspondance).

5.1.1 Introduction sur l'AFC

L'AFC est une forme particulière de l'ACP. Cette méthode s'applique à des tableaux de contingence croisant particulièrement deux variables qualitatives X et Y à m_1 et m_2 modalités . Les données sont les effectifs des individus ayant les deux modalités données.

L'objectif est de faire une synthèse sur le tableau pour répondre aux questions :

pour une variable donnée, certaines modalités sont-elles proches ou éloignées.

entre les deux variables, certaines modalités « s'attirent-elles » davantage ou au contraire « se repoussent ».

Remarque

L'AFC n'a d'intérêt que si il y a dépendance entre les deux variables, dans le cas contraire elle n'apporte pas d'information.

Le tableau se présente sous la forme :

		$variable qualitative 2$		
		$modalité 1 ...$	$modalité j ...$	$modalité\ m_2$
$variable qualitative 1$:	$modalité 1$	n_{11}	n_{1j}	n_{1m_2}
	$modalité\ i$	n_{i1}	n_{ij}	n_{im_2}
	$modalité\ m_1$	$n_{m_1 1}$	$n_{m_1 j}$	$n_{m_1 m_2}$

Pour mesurer les distances entre modalités, il est nécessaire de calculer au préalable la distribution de chaque modalité d'une variable en fonction de l'autre variable.

On définit ainsi les profils colonnes et les profils lignes qui sont les distributions respectives des modalités des deux variables.

Un profil ligne : $\frac{n_{ij}}{n_{i.}}$

Un profil colonne : $\frac{n.ij}{n_{.j}}$.

Chaque profil est alors assimilé à un point de coordonnées les proportions par rapport aux modalités de l'autre variable. Le nuage des profils ligne est alors projeté sur des axes factoriels en conservant le

maximum d'inertie et il en est de même pour le nuage des profils colonne. La distance entre deux profils ligne i,í est calculée à l'aide de la distance du χ^2

$$d^2_{\chi^2} = \sum_{j=1}^{m_2} \frac{n}{n_{.j}} (\frac{n_{ij}}{n_{i.}} - \frac{n_{íj}}{n_{i.}})^2$$

-Les valeurs propres (inertie projetée sur l'axe) sont inférieures à 1.

5.2 Application

Pour les données que nous disposons nous allons appliquer trois AFC pour pouvoir donner une interprétation sur la variable "indemnité" par rapport aux variables "type accident" "age" et "genre"..

5.2.1 Première AFC

La première variable est : indemnité (ayant les modalités : faible,moyenne et forte)

La deuxième variable est : accident (ayant les modalités tous les types d'accidents considérés)

Le tableau de contingence est le suivant :

	FAIBLE [1,20[	MOYENNE [20,40[	FORTE [40,100]
AR+FRCO	0	0	1
BRZ	1	0	0
CCPJCH	1	0	0
CVP	1	0	0
FRB	1	4	3
FRC	3	1	0
FRCH	1	0	0
FRCL	1	1	0
FRCO	2	1	0
FRCO+FRB	0	0	1
FRCO+FRCU	0	0	1
FRCOU	0	0	1
FRCU	1	3	1
FRCU+FRB	0	0	1
FRCU+FRI	1	1	0
FRE	2	4	0
FRG	1	4	0
FRG+FRBA	1	0	0
FRH	1	2	1
FRN	2	0	1
FROJTRC	0	0	1
FRP	5	2	0

FRV	0	1	2
FRV+FRH	0	0	1
FRVO	0	0	1
L+FRH+TRC	0	0	1
LE	1	0	0
PEM	0	1	0
PG	1	0	0
PM	1	0	0
RI	1	0	0
RUR+FRV	0	0	1
RV	0	1	0
TRB+TRP	1	0	0
TRC	9	3	2
TRC+FRC	0	0	2
TRC+LE	0	0	1
TRC+TRE	0	1	0
TRC+TRE+TRG	0	1	0
TRCO+TRB	2	0	0
TRG	1	0	0
TRH	1	0	0
TRO	1	0	0
TRTH	1	1	0
TRV	1	0	0

L'application de l' AFC à notre tableau de contingence nous donne les résultats suivants :

$coord

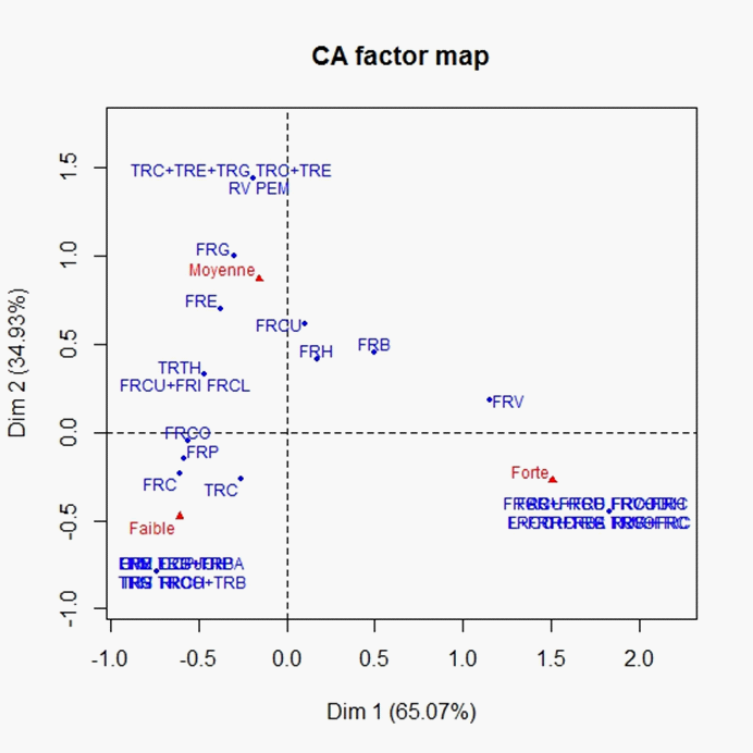
	Dim 1	Dim 2
AR+FRCO	1.8284727	-0.44960219
BRZ	-0.7412203	-0.79025660
CCPJCH	-0.7412203	-0.79025660
CVP	-0.7412203	-0.79025660
FRB	0.4972393	0.45516479
FRC	-0.6038079	-0.23141861
FRCH	-0.7412203	-0.79025660
FRCL	-0.4663955	0.32741938
FRCO	-0.5580038	-0.04513928
FRCO+FRBA	1.8284727	-0.44960219
FRCO+FRCU	1.8284727	-0.44960219
FRCOU	1.8284727	-0.44960219
FRCU	0.1025080	0.61908546
FRCU+FRB	1.8284727	-0.44960219
FRCU+FRI	-0.4663955	0.32741938
FRE	-0.3747873	0.69997805
FRG	-0.3015007	0.99802498
FRG+FRBA	-0.7412203	-0.79025660
FRH	0.1760277	0.41258299
FRN	-0.7412203	-0.79025660
FROJTRC	1.8284727	-0.44960219
FRP	-0.5841776	-0.15158461
FRV	1.1551249	0.18196366
FRV+FRH	1.8284727	-0.44960219
FRVO	1.8284727	-0.44960219
L+FRH+TRC	1.8284727	-0.44960219
LE	-0.7412203	-0.79025660
PEM	-0.1915708	1.44509537
PG	-0.7412203	-0.79025660
PM	-0.7412203	-0.79025660
RI	-0.7412203	-0.79025660
RUR+FRC	1.8284727	-0.44960219
RV	-0.1915708	1.44509537
TRB+TRP	-0.7412203	-0.79025660
TRC	-0.2563393	-0.26258769
TRC+FRC	1.8284727	-0.44960219
TRC+LE	1.8284727	-0.44960219
TRC+TRE	-0.1915708	1.44509537
TRC+TRE+TRG	-0.1915708	1.44509537
TRCO+TRB	-0.7412203	-0.79025660
TRG	-0.7412203	-0.79025660
TRH	-0.7412203	-0.79025660
TRO	-0.7412203	-0.79025660
TRTH	-0.4663955	0.32741938
TRV	-0.7412203	-0.79025660

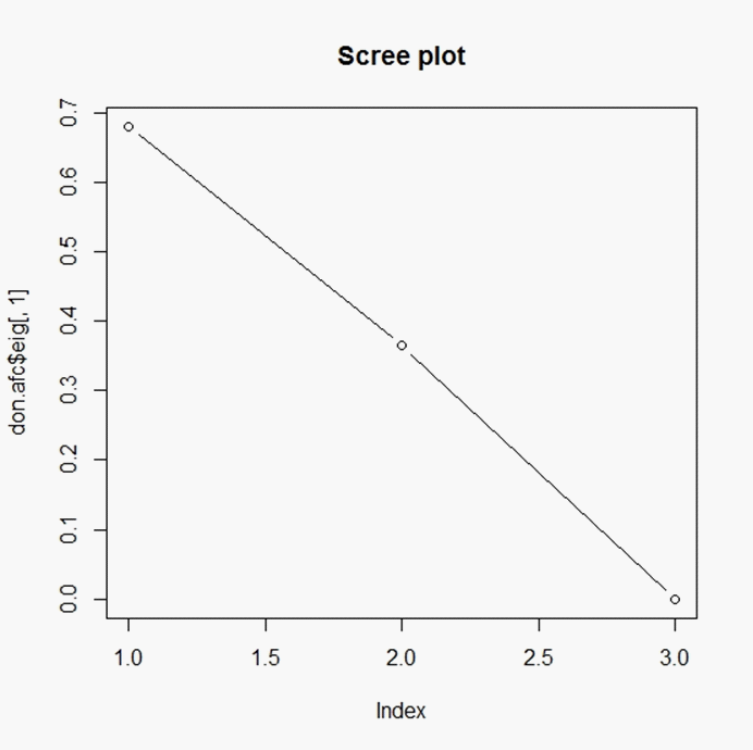
$contrib

	Dim 1	Dim 2
AR+FRCO	4.91951013	0.55402765
BRZ	0.80842460	1.71163384
CCPJCH	0.80842460	1.71163384
CVP	0.80842460	1.71163384
FRB	2.91048812	4.54257287
FRC	2.14586558	0.58712578
FRCH	0.80842460	1.71163384
FRCL	0.64015285	0.58764274
FRCO	1.37448647	0.01675348
FRCO+FRBA	4.91951013	0.55402765
FRCO+FRCU	4.91951013	0.55402765
FRCOU	4.91951013	0.55402765
FRCU	0.07730907	5.25225519
FRCU+FRB	4.91951013	0.55402765
FRCU+FRI	0.64015285	0.58764274
FRE	1.24012616	8.05739566
FRG	0.66879271	13.64983573
FRG+FRBA	0.80842460	1.71163384
FRH	0.18237569	1.86619587
FRN	1.61684920	3.42326769
FROJTRC	4.91951013	0.55402765
FRP	3.51505983	0.44084189
FRV	5.89011837	0.27224831
FRV+FRH	4.91951013	0.55402765
FRVO	4.91951013	0.55402765
L+FRH+TRC	4.91951013	0.55402765
LE	0.80842460	1.71163384
PEM	0.05400121	5.72357819
PG	0.80842460	1.71163384
PM	0.80842460	1.71163384
RI	0.80842460	1.71163384
RUR+FRC	4.91951013	0.55402765
RV	0.05400121	5.72357819
TRB+TRP	0.80842460	1.71163384
TRC	1.35364010	2.64576561
TRC+FRC	9.83902027	1.10805530
TRC+LE	4.91951013	0.55402765
TRC+TRE	0.05400121	5.72357819
TRC+TRE+TRG	0.05400121	5.72357819
TRCO+TRB	1.61684920	3.42326769
TRG	0.80842460	1.71163384
TRH	0.80842460	1.71163384
TRO	0.80842460	1.71163384
TRTH	0.64015285	0.58764274
TRV	0.80842460	1.71163384

$cos2

	Dim 1	Dim 2
AR+FRCO	0.94298555	0.05701445
BRZ	0.46801384	0.53198616
CCPJCH	0.46801384	0.53198616
CVP	0.46801384	0.53198616
FRB	0.54409114	0.45590886
FRC	0.87192148	0.12807852
FRCH	0.46801384	0.53198616
FRCL	0.66986719	0.33013281
FRCO	0.99349867	0.00650133
FRCO+FRBA	0.94298555	0.05701445
FRCO+FRCU	0.94298555	0.05701445
FRCOU	0.94298555	0.05701445
FRCU	0.02668503	0.97331497
FRCU+FRB	0.94298555	0.05701445
FRCU+FRI	0.66986719	0.33013281
FRE	0.22280737	0.77719263
FRG	0.08363045	0.91636955
FRG+FRBA	0.46801384	0.53198616
FRH	0.15399672	0.84600328
FRN	0.46801384	0.53198616
FROJTRC	0.94298555	0.05701445
FRP	0.93691575	0.06308425
FRV	0.97578603	0.02421397
FRV+FRH	0.94298555	0.05701445
FRVO	0.94298555	0.05701445
L+FRH+TRC	0.94298555	0.05701445
LE	0.46801384	0.53198616
PEM	0.01727029	0.98272971
PG	0.46801384	0.53198616
PM	0.46801384	0.53198616
RI	0.46801384	0.53198616
RUR+FRC	0.94298555	0.05701445
RV	0.01727029	0.98272971
TRB+TRP	0.46801384	0.53198616
TRC	0.48796070	0.51203930
TRC+FRC	0.94298555	0.05701445
TRC+LE	0.94298555	0.05701445
TRC+TRE	0.01727029	0.98272971
TRC+TRE+TRG	0.01727029	0.98272971
TRCO+TRB	0.46801384	0.53198616
TRG	0.46801384	0.53198616
TRH	0.46801384	0.53198616
TRO	0.46801384	0.53198616
TRTH	0.66986719	0.33013281
TRV	0.46801384	0.53198616


CA factor map
TRC+TRE+TRG
TRO+TRE
RV
PEM
FRG
Moyenne
FRE
FRCU
FRH
FRB
TRTH
FRCU+FRI
FRCL
FRV
FRCO
FRP
FRC
TRC
Forte
Faible
+TRB
Dim 2 (34.93%)
Dim 1 (65.07%)
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0



La qualité de la représentation de la variable "accident" sur chaque modalité (type d'accident) est mesurée par la formule suivante. Par exemple pour la modalité FRV on calcule

$$\cos^2 F_1 + \cos^2 F_2 = 0.99$$

où F_1 est le premier axe factoriel,et F_2 est le deuxième axe factoriel. La règle consiste à dire que si la somme des cosinus est très proche de 1 on conclut que la modalité FRV à une très bonne qualité de représentation.

On fait de même pour les autres modalités et on trouve la même qualité de représentation. Donc on conclut que toutes les modalités de la variable "accident" sont très bien représentées.

Nous allons partager les modalités " types d'accidents" de la façon suivante : ceux qui sont du coté positif et ceux qui sont du coté négatif pour les deux variables.

Première variable "accident"

La règle est : on choisit les types d'accidents qui ont une contribution supérieure ou égale à $\frac{100\%}{45} = 2.2\%$

Pour le premier axe :

Groupe 1 (+)	Groupe 2 (-)
FRV (5.89%)	FRP (3.5%)
FRB (2.91%)	FRC (2.2%)
FRCO+FRBA (4.9%)	TRC+FRC (4.9%)
FRCO+AR (4.9%)	
FRCO+FRCU (4.9%)	
FRCU+FRB(4.9%)	
FRV+FRH (4.9%)	
FRVO (4.9%)	
FRH+TRC+LE (4.9)	
FRC+RUR (4.9%)	
TRC+LE (4.9%)	
FROJTRC (4.9%)	

Les modalités du groupe 1 (corrélation positive) sont en opposition avec les modalités du groupe 2 (corrélation négative). On peut dire que les types d'accident du groupe1 sont plus graves et ceux du deuxième groupe et vice versa.

Pour le deuxième axe, avec les mêmes règles on a :

Groupe 3 (+)	Groupe 4 (-)
FRB (4.9%)	TRC(2.64%)
FRCU (5.2%)	FRN(3.42%)
FRE (8%)	TRCO+TRB(3.4%)
FRG (13.6%)	
PEM (5.7%)	
RV (5.7%)	
TRC+TRE (5.72%)	
TRC+TRE+TRG (5.72%)	

Les modalités du groupe 3 sont en opposition avec les modalités du groupe 4

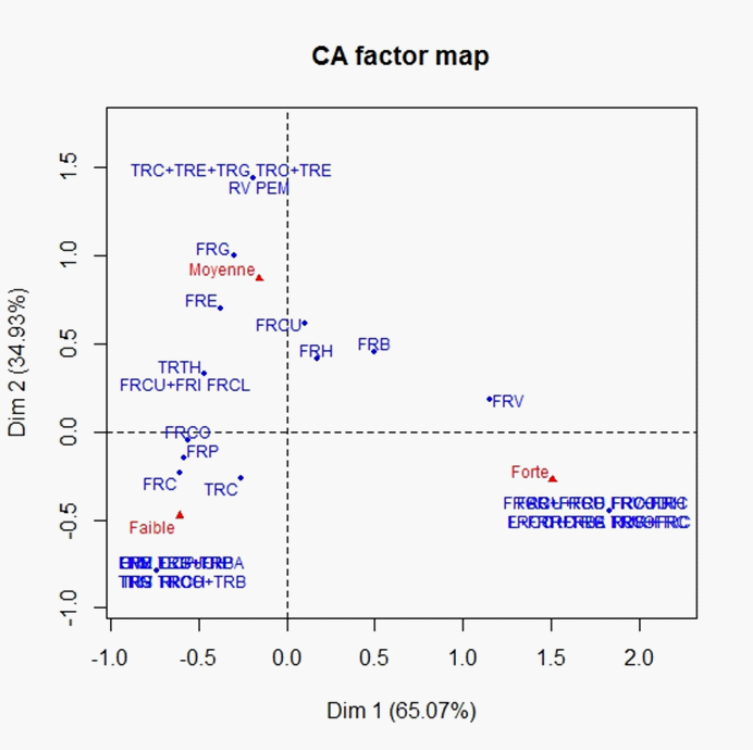
Deuxième variable " indemnité"

c'est la variable à expliquer et la règle d'interprétation consiste à comparer la modalité ayant pour coordonnée a_i de l'axe factoriel par rapport a sa valeur propre.

On a :

$$\begin{aligned} \lambda_1 &= 0.7 \Rightarrow \sqrt{\lambda_1} = 0.83 \\ \lambda_2 &= 0.4 \Rightarrow \sqrt{\lambda_2} = 0.63 \end{aligned}$$

FORTE : $\mid a_i \mid = 1.5 > 0.83 = \sqrt{\lambda_1}$



MOYENNE $\quad | a_i | = 0.8 > 0.63 = \sqrt{\lambda_2}$

FAIBLE $\quad | a_i | = 0.5 < 0.63$ et $| a_i | = 0.5 < 0.83$

Dans ce calcul la valeur de $| a_i |$ défint deux axes : 1er axe "modalité FORTE " (supérieur à $\sqrt{\lambda_1}$) , 2eme axe " modalité MOYENNE" (supérieur à $\sqrt{\lambda_2}$).

Pour la modalité FAIBLE ,on ne peut rien dire sauf qu'on remarque qu'elle est en opposition avec la modalité FORTE.

On voit que la modalité FORTE est très proche du premier axe et que la modalité MOYENNE est proche du deuxième axe. Donc on conclut que la modalité FORTE définit le premier axe et la modalité MOYENNE définit le deuxième axe c'est ce qu'on a vérifié par les calculs.

On voit aussi que les modalités du "Groupe 1" qui sont toutes des fractures sont très proches de la modalité FORTE ainsi elles sont fortement indemnisées. Par contre le "Groupe 2" est "proche" de la modalité FAIBLE , on peut dire qu'ils sont faiblement indemnisés. Mais aussi il peut être considéré proche de la modalité MOYENNE et donc moyennement indemnisés. (voir la qualité de représentation de la modalité FAIBLE sur le graphe précèdent) .

Pour les types d'accidents " très proches " de la modalité MOYENNE , on dit qu'ils sont moyennement indemnisés et on remarque que ce sont des traumatismes.

5.2.2 Deuxième AFC

Variable 1= indémnité (ayant les modalités FAIBLE,MOYENNE,FORTE)

Variable 2 = genre (ayant les modalités H homme, F femme, E enfant).

Le tableau de contingence est le suivant :

	H	F	E
FAIBLE	19	15	12
MOYENNE	18	9	5
FORTE	19	2	1

Les résultats de l' AFC sont

```
$coord
              Dim 1       Dim 2
Faible    0.30118173  0.03138422
Moyenne -0.01250962 -0.06873958
Forte    -0.61154780  0.03436328

$contrib
              Dim 1    Dim 2
Faible   33.63577092 20.36423
Moyenne   0.04036683 67.95963
Forte    66.32386225 11.67614

$cos2
             Dim 1       Dim 2
Faible   0.98925824 0.010741755
Moyenne 0.03205711 0.967942887
Forte    0.99685255 0.003147453

$inertia
     Faible     Moyenne       Forte
0.042179885 0.001562118 0.082537740
```

5.2.3 Première variable

On choisit les types d'accidents qui ont une contribution supérieur ou égale à $\frac{100\%}{3} = 33.33\%$.

Axe1

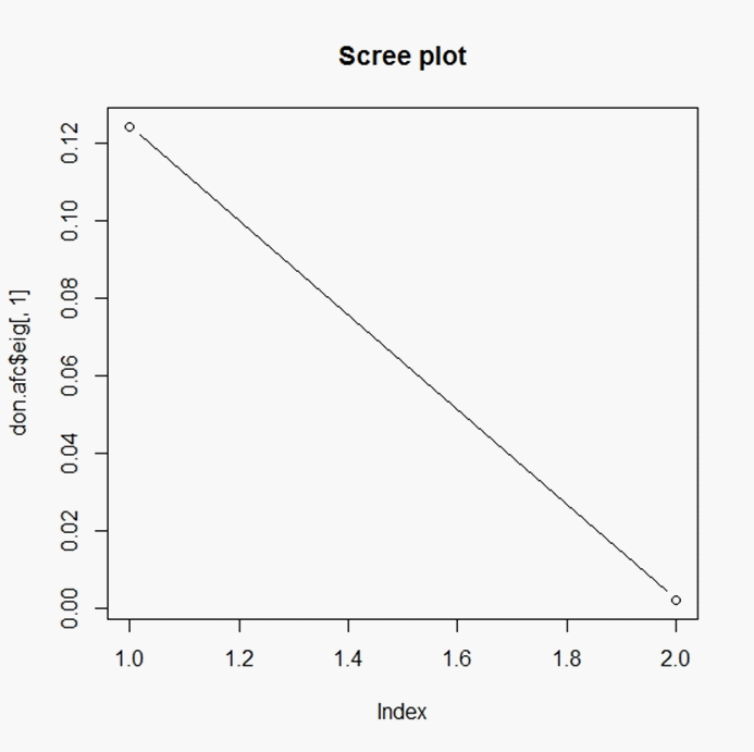
:

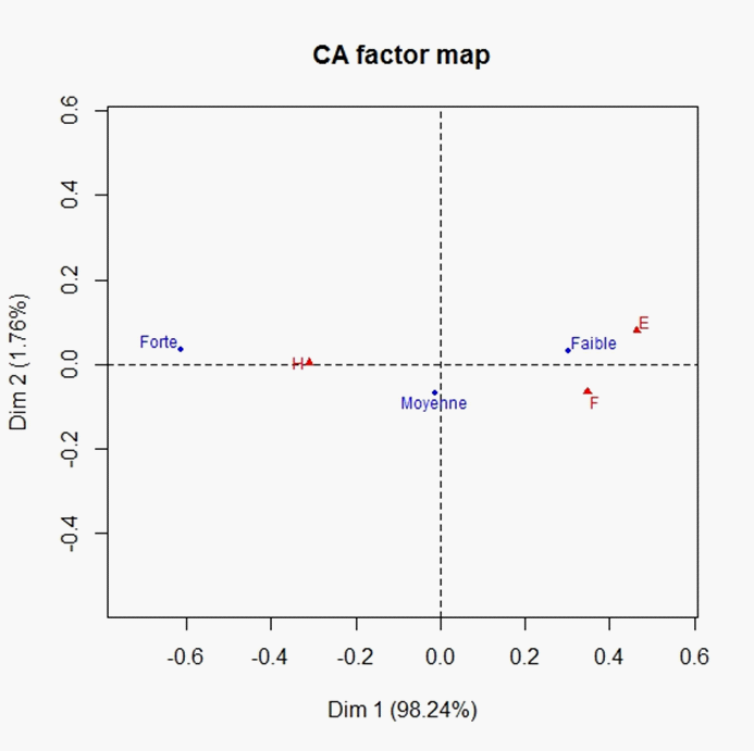
+	-
FAIBLE(33.36)	FORTE(66.3)

Axe2 :

+	-
	MOYENNE(67)

Deuxième variable





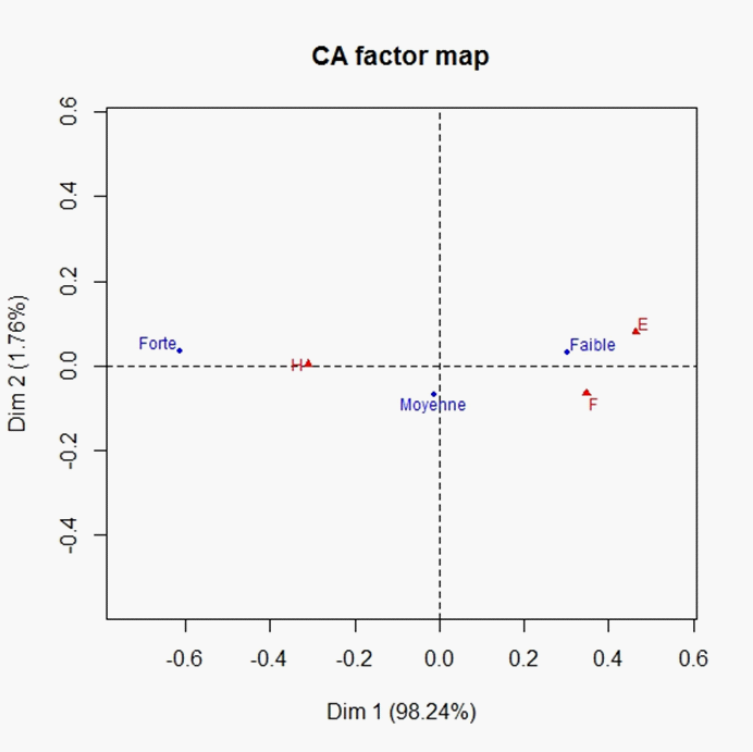
On voit qu'on a une seule valeur propre $\lambda = 0.12 \Rightarrow \sqrt{\lambda} = 0.34$

H : homme : $\mid a_i \mid = 0.35 > 0.34 = \sqrt{\lambda}$.

E : enfant : $\mid a_i \mid = 0.5 > 0.34$.

F : femme $\mid a_i \mid = 0.39 > 0.34$.

+	-
E	H
F	



Intèrprétation

Les modalités femme, enfant et homme définissent le même axe qui est le premier

On voit que la modalité FAIBLE est trés proche des modalités E et F. Elle a également une trés bonne qualité de représentation (99%) ,on peut donc conclure que les enfants et les femmes sont faiblement indemnisés

La modalité FORTE est en opposition avec la modalité FAIBLE, ce qui est logique et comme la modalité H a une très bonne qualité de représentation, on conclu que les hommes sont fortement indemnisés.

5.2.4 Troisième AFC

Variable 1 = âge (ayant les modalités : Enfant,Jeune,Adulte,Senior)

Variable 2 = indemnité (ayant les modalités :FAIBLE ,MOYENNE,FORTE)

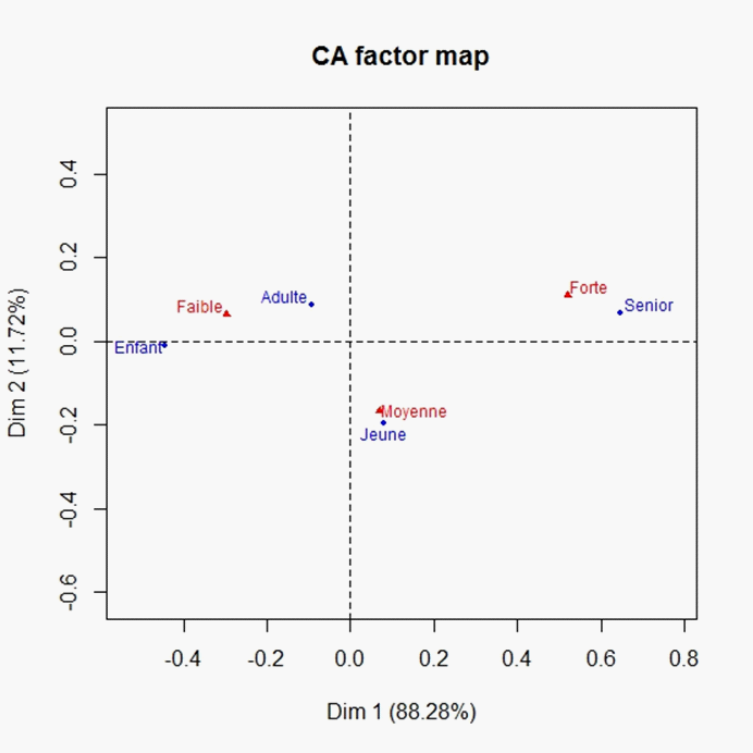
Le tableau de contingence est le suivant :

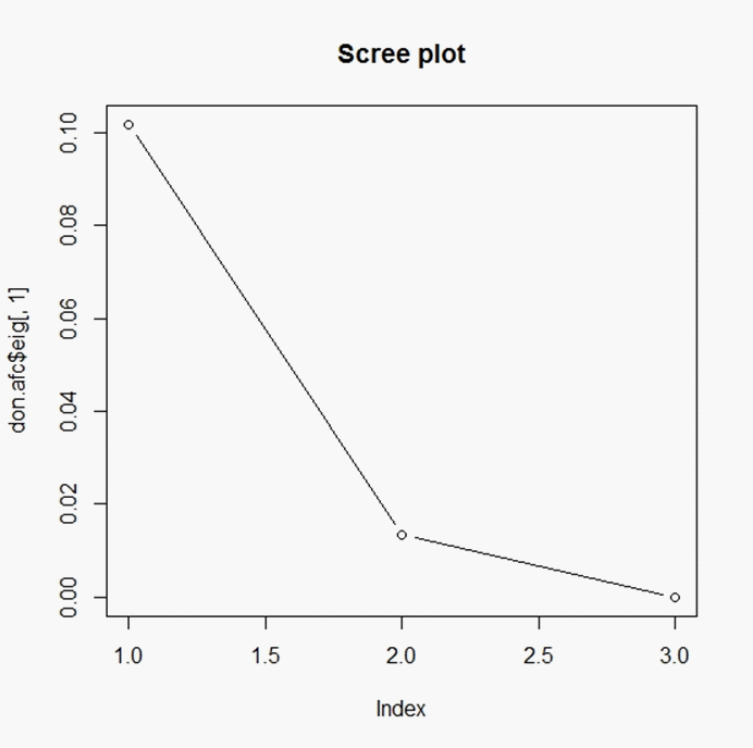
	FAIBLE	MOYENNE	FORTE
Enfant	11	5	1
Jeune	9	10	5
Adulte	23	12	9
Senior	3	5	7

Les résultats de l'AFC sont :

```
$coord
             Dim 1       Dim 2
Enfant -0.44286201 -0.01073421
Jeune   0.08118710 -0.19732018
Adulte -0.09393849  0.08830855
Senior  0.64756381  0.06883929

$contrib
           Dim 1      Dim 2
Enfant 32.781942  0.1450296
Jeune   1.555372 69.1865950
Adulte  3.817585 25.4053946
Senior 61.845101  5.2629808

$cos2
           Dim 1        Dim 2
Enfant 0.9994129 0.0005871485
Jeune  0.1447801 0.8552198940
Adulte 0.5308623 0.4691376818
Senior 0.9888255 0.0111744801

$inertia
    Enfant      Jeune     Adulte     Senior
0.03336114 0.01092638 0.00731405 0.06361166
```

On remarque que toutes les modalités ont une très bonne qualité de représentation (99%).





Première variable

On choisit ceux qui ont une contribution supèrieur ou égale à 25%.

Axe 1 :

+	-
Senior	Enfant

Axe 2 :

+	-
Adulte	Jeune

Deuxième variable

On a

$$\begin{aligned} \lambda_1 &= 0.1 \Rightarrow \sqrt{\lambda_1} = 0.31 \\ \lambda_2 &= 0.01 \Rightarrow \sqrt{\lambda_2} = 0.1 \end{aligned}$$

FORTE : $\mid a_i \mid = 0.5 > 0.31$

Faible : $\mid a_i \mid = 0.29 < 0.31$

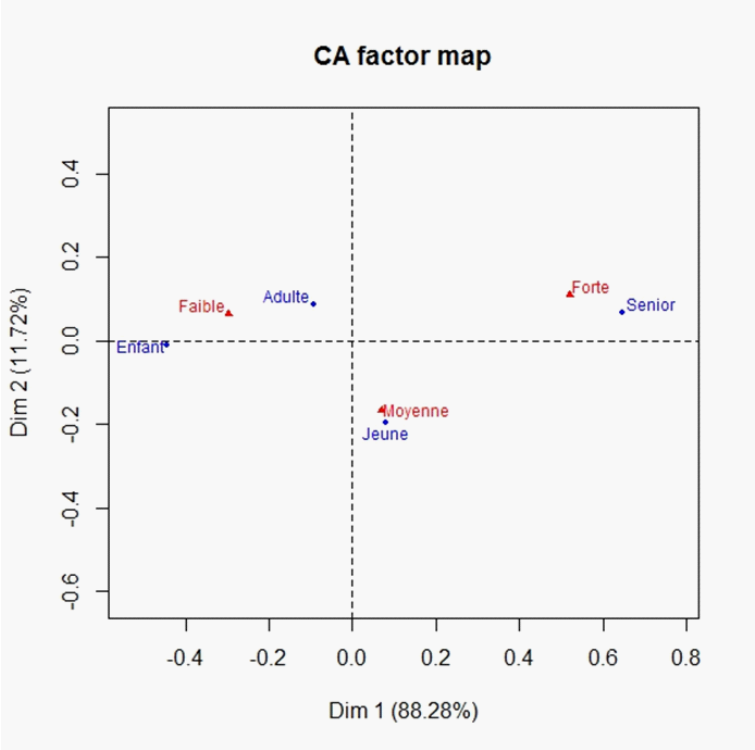
MOYENNE : $\mid a_i \mid = 0.1 = \sqrt{\lambda_2}$

Axe 1 :

+	-
FORTE	

Axe 2 :

+	-
	MOYENNE



5.2.5 Interprétation

La modalité FORTE définit le premier axe, elle est en opposition avec FAIBLE. La modalité MOYENNEdéfinit le deuxième axe

On observe aussi que les Senior sont fortement indemnisés puisqu'ils ont une très bonne qualité de représentation. Contrairement aux enfants qu'on peut dire qu'ils sont faiblement indemnisés mais ça reste équivoque puisque FAIBLE n'a pas une bonne qualité de représentation

On voit que les adultes et les jeunes sont beaucoup plus proches du deuxième axe,donc on conclut qu'il sont moyennement indemnisés.

5.2.6 Conclusion générale

Les femmes et les enfants sont faiblement indemnisés, contrairement aux hommes qui sont fortement indemnisés.

Les séniors reçoivent une indemnité forte, et les adultes reçoivent une indemnité moyenne.

Conclusion 8 *L'étude des corrélations est aujourd'hui nécessaire dans presque tous les secteurs de l'activité humaine, les méthodes statistiques qu'elle propose devraient faire partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du biologiste et de l'informaticien, car en s'appuyant sur l'existence des liens statistiques entre deux phénomènes ou plus elle permet de confirmer une théorie ou de la contredire.*

Type d'accident	Code
Ablation de la rate et fracture de cottes	AR+FRCO
Brulure de visage	BRZ
Contusion de la cuisse, poigné, jambe et cheville	CCPJCH
Contusion visage et pied	CVP
Fracture bras	FRB
Fracture de cottes	FRC
Fracture de la cheville	FRCH
Fracture de la clavicule	FRCL
Fracture de cottes	FRCO
Fracture de cottes et du bassin	FRCO+FRBA
Fracture de cottes et de la cuisse	FRCO+FRCU
Fracture du coud	FRCOU
Fracture de la cuisse	FRCU
Fracture de la cuisse et du bras	FRCU+FRB
Fracture de la cuisse et de l'index	FRCU+FRI
Fracture d'épaule	FRE
Fracture du genou	FRG
Fracture du genou et du bassin	FRG+FRBA
Fracture de la hanche	FRH
Fracture du nez omoplate, jambe et traumatisme crânien	FRN
Fracture omoplate, jambe et traumatisme crânien	FROJTRC
Fracture du poigné	FRP
Fracture des vertèbres	FRV
Fracture des vertèbres et de la hanche	FRV+FRH
Fracture vertèbres dorsale	FRVO
Luxation fracture hanche et traumatisme cranien	L+FRH+TRC
Luxation épaule	LE
Perdu la mémoire	PEM
Plaie du genou	PG
Plaies multiples	PM
Amputation de l'index	RI
Rupture de rate et fracture de la cuisse	RUR+FRC
Rupture de vessie	RV
Traumatisme du bras et du poigné	TRB+TRP
Traumatisme crânien	TRC
Traumatisme crânien et fracture de la cuisse	TRC+FRC
Traumatisme crânien et luxation épaule	TRC+LE
Traumatisme crânien et de l’épaule	TRC+TRE
Traumatisme crânien, de l’épaule et du genoux	TRC+TRE+TRG
Traumatisme du coude et du bras	TRCO+TRB
Traumatisme du genou	TRG
Traumatisme de la hanche	TRH

Traumatisme oculaire	TRO
Traumatisme thoracique	TRTH
Traumatisme du visage	TRV

Références

1. *Titre :* Probabilités, analyse des données et statistique
   Auteur : SAPORTA Gilbert

2. *Titre :* Analyse de corrélation
   *Auteur :*RiccoRakotomalala
   *Lien :*eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf

3. *Titre :* Analyse factorielle des correspondances avec R
   *Auteur :*RiccoRakotomalala
   Lien :http://eric.univ-lyon2.fr/~ricco/cours/didacticiels/R/afc_avec_r.pdf

4. *Titre :* Quelques notes sur l'interprétation d'une analyse factorielle ou canonique des correspondances
   Auteur : Daniel Brocard
   Lien :http://moodle.epfl.ch/pluginfile.php/161301/mod_folder/content/0/Borcard_AFC-ACC.pdf

5. *Titre :* Analyse quantitative des données biologiques
   Auteur : Pierre Legendre
   Lien :http://biol09.biol.umontreal.ca/Bio6077/Travaux_pratiques_en_R.pdf

6. *Titre :* Analyse canonique
   Auteur : Laurence Reboul
   Lien :http://iml.univ-mrs.fr/~reboul/canonique.pptx.pdf