**Exercice 1**

**Patie 2**  
**a)**

> E\_vE=matrix(c(.1824,.0770,.2368,.0535,.0770,.2,.1989,.0881,.2368,.1989,1.1260,-.0092,.0535,.0881,-.0092,.7461),byrow=TRUE,nrow=4)

> E\_vE

[,1] [,2] [,3] [,4]

[1,] 0.1824 0.0770 0.2368 0.0535

[2,] 0.0770 0.2000 0.1989 0.0881

[3,] 0.2368 0.1989 1.1260 -0.0092

[4,] 0.0535 0.0881 -0.0092 0.7461

**E/vE est la matrice de covariance, la matrice de corrélation associée est donc :**

> cov2cor(E\_vE)

[,1] [,2] [,3] [,4]

[1,] 1.0000000 0.4031469 0.52251659 0.14502499

[2,] 0.4031469 1.0000000 0.41913178 0.22806697

[3,] 0.5225166 0.4191318 1.00000000 -0.01003737

[4,] 0.1450250 0.2280670 -0.01003737 1.00000000

**On voit que les deux matrice donne la même relation entre des variables, par exemple dans la matrice de variance, il n’y a que la 3ème variable (manie) et la 4ème variable (dépression) qui ont une covariance négative, pareille que dans la matrice de corrélation.**

**Cela veut dire que quand on est dans la même famille, si on a la manie, ça diminue la chance d’avoir la dépression, et si on a la dépression, ça diminue la chance d’avoir la manie. Mais on remarque aussi que les corrélations entre peu n’importe 2 variable dans la même famille ne sont pas très forte.**

> H\_vH=matrix(c(.2707,.3724,-.2078,-.2814,.3724,.6862,-.2857,-.3957,-.2078,-.2857,1.1578,.4399,-.2814,-.3957,.4399,.4846),byrow=TRUE,nrow=4)

> H\_vH

[,1] [,2] [,3] [,4]

[1,] 0.2707 0.3724 -0.2078 -0.2814

[2,] 0.3724 0.6862 -0.2857 -0.3957

[3,] -0.2078 -0.2857 1.1578 0.4399

[4,] -0.2814 -0.3957 0.4399 0.4846

**H/vH est la matrice de covariance, la matrice de corrélation associée est donc :**

> cov2cor(H\_vH)

[,1] [,2] [,3] [,4]

[1,] 1.0000000 0.8640526 -0.3711801 -0.7769414

[2,] 0.8640526 1.0000000 -0.3205296 -0.6861973

[3,] -0.3711801 -0.3205296 1.0000000 0.5872804

[4,] -0.7769414 -0.6861973 0.5872804 1.0000000

**On voit qu’ici la 1ière variable (délire) et la 2ème variable (hallucination) ont une corrélation positive (forte), pareille pour la 3ème variable (manie) et la 4ème variable (dépression), tandis que la 1ière variable (délire) a une corrélation négative faible avec la 3ème variable (manie) et une corrélation négative forte avec la 4ème variable (dépression), la 2ème variable (hallucination) a une corrélation négative avec la 3ème variable (manie) et la 4ème variable (dépression).**

**Pour deux familles différents, l’existence de maladie délire dans une famille augmente la chance d’avoir l’hallucination dans une autre, identiquement pour la manie et la dépression. Contrairement, si une famille a le délire, l’autre famille a moins de chance d’avoir la manie ou la dépression, et si une famille a l’hallucination, l’autre famille a moins de chance d’avoir la manie ou la dépression.**

**Clairement ça ne donne pas le même résultat que pour la matrice E/vE, ça veut dire qu’il existe un impact de la famille sur la Schizophrénie.**

**b)**

> k=36

> N=521

> vH=k-1

> vH

[1] 35

> vE=N-k

> vE

[1] 485

> E=E\_vE\*vE

> H=H\_vH\*vH

> VP=eigen(solve(E)%\*%H)

> VP

$values

[1] 0.62625195 0.06512067 0.02621220 0.01057458

$vectors

[,1] [,2] [,3] [,4]

[1,] -0.5338181 -0.62638042 0.8489749 -0.73730962

[2,] -0.7419981 0.71146699 -0.4651062 0.05781169

[3,] 0.3279259 0.31459654 0.2002212 0.13161266

[4,] 0.2386245 0.04991311 -0.1510934 -0.66008367

> lamda=VP$values[1]

> lamda

[1] 0.626252

> delta=lamda/(1+lamda)

> delta

[1] 0.3850891

> p=4

> #Les parametres:

> s=min(vH,p)

> s

[1] 4

> m=(abs(vH-p)-1)/2

> m

[1] 15

> N\_=(vE-p-1)/2

> N\_

[1] 240

**Selon la table, la valeur critique = 0.124, et delta = 0.385 > 0.124, on rejette donc H0, la maladie est héréditaire.**

**Exercice 4**

**i. Faire l’analyse discriminante à l’aide de la fonction discriminante linéaire normale.**

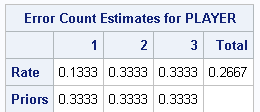
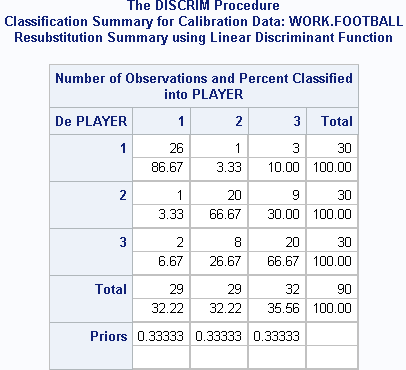
**a)**

**Méthode apparente.**

En SAS : **proc** **discrim** data=football;

class PLAYER;

**run**;



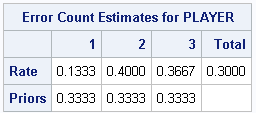
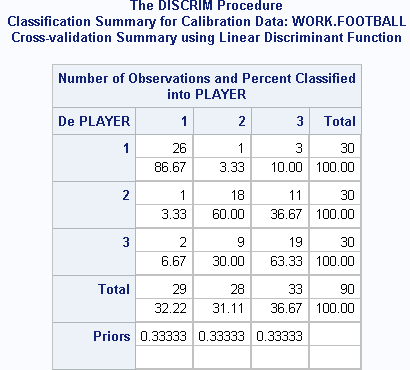
Par cette méthode, le taux d’erreur de classification = 1 – taux de classification correct = 1 – (26+20+20)/90 = 0.267.  
En total, il y a 0.267\*90 = 24 observations qui sont mal classées par cette méthode.

**Méthode de validation croisée.**

En SAS: **proc** **discrim** data=football crossvalidate;

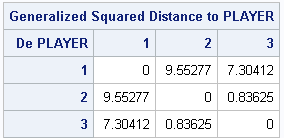
class PLAYER;

**run**;



Par cette méthode, le taux d’erreur de classification = 1 – taux de classification correct = 1 – (26+18+19)/90 = 0.3.  
En total, il y a 0.3\*90 = 27 observations qui sont mal classées pas cette méthode.

**Méthode plug-in.**



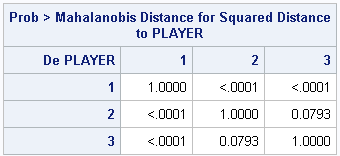
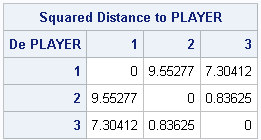
**Conclusion** : Si on fait l’analyse discriminante (analyse de classification) à l’aide de la fonction discriminante linéaire normale, le taux d’erreur de classification à l’aide d’estimé de la validation croisée est un plus grand que celui de l’estimé apparente, ainsi la méthode de la validation est peu plus robuste.   
En plus, on remarque que la méthode de plug-in sous-estime beaucoup l’erreur de classification quand les groupes sont distants, ainsi cette méthode n’est pas bonne.

**b)**

En SAS : **proc** **discrim** data=football MAHALANOBIS;

class PLAYER;

**run**;



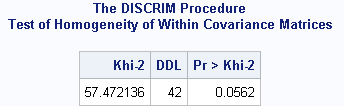
On remarque que la distance entre le groupe 2 et le groupe 3 est la plus petite(0.836), sa p-value est de 0.079 > 0.05, on rejette donc H0, et on peut conclure que le groupe 2 et le groupe 3 sont proches, ainsi c’est le groupe 1 qui est significativement différent que les autres groupes. En plus, on trouve que le p-value de H0 : d(1,2) = 0 est presque 0, ainsi pour le p-value de H0 : d(1,3) = 0, on rejette donc ces 2 hypothèses nulles. Cela confirme notre conclusion, le groupe 1 est significativement différent que les autres groupes.

**c)**

En SAS : **proc** **discrim** data=football pool=test SLpool=**0.01**;

class PLAYER;

**run**;



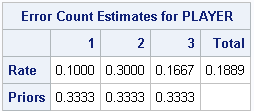
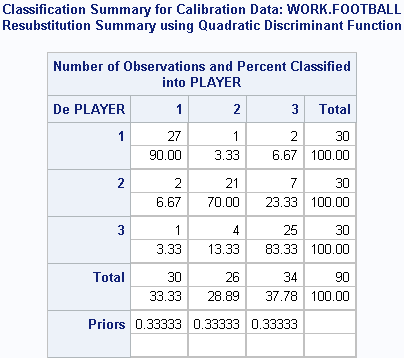
On a le p-value = 0.056 > 0.05, il n’y donc pas assez d’évidences à rejeter l’hypothèse nulle d’homogénéité, il existe l’homogénéité des matrices de variance-covariance entre des groupes.

**ii.** **Faire l’analyse discriminante à l’aide de la fonction discriminante quadratique.**

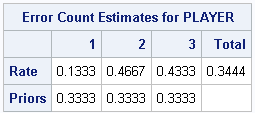
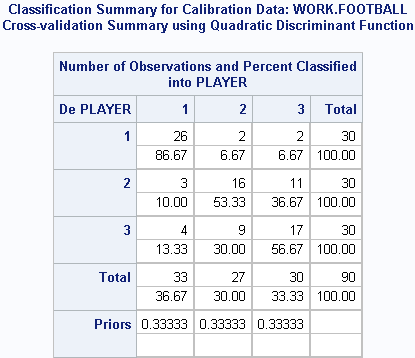
En SAS : **proc** **discrim** data=football pool=no crossvalidate;

class PLAYER;

**run**;



Le taux d’erreur de classification à l’aide d’estimé apparente = 1-taux de classification correct = 1-(27+21+25)/2 = 0.1889.  
En total, il y a 0.1889\*90 = 17 observations qui sont mal classées par cette méthode.



Le taux d’erreur de classification à l’aide d’estimé de la validation croisée = 1-taux de classification correct = 1-(26+16+17)/2 = 0.3444.  
En total, il y a 0.3444\*90 = 30 observations qui sont mal classées par cette méthode.

**Conclusion** : On remarque que le taux d’erreur de classification à l’aide de l’estimé de la validation croisée est beaucoup plus élevé que celui à l’aide de l’estimé apparente, la méthode de la validation est donc beaucoup plus robuste.

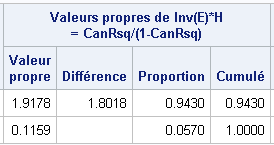
**iii.**

En SAS : **proc** **candisc** data=football out=CAND pcov;

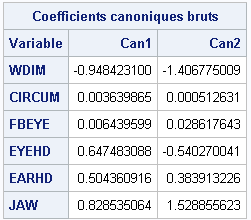
class PLAYER;

**run**;

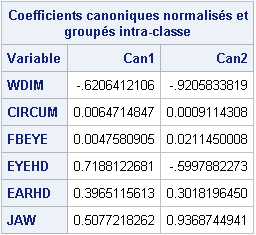
On obtient les valeurs propres de  :



Les deux valeurs propres de sont 1.9178 et 0.1159, et les vecteurs propres **a** associés à ces deux valeurs sont :



On standardise ces vecteurs par **a**\* = **a**, et on a les vecteurs standardisés :



On a donc que pour **a**1, la 4ème variable(EYEHD), la 1ère variable(WDIM), la 6ème variable(JAW) et la 5ème variables(EARHD) contribuent les plus pour la séparation des groupes en ordre décroissante, en présence de ces variables, les 2 autres variables(CIRCUM et FBEYE) ne sont pas utiles pour séparer les groupes.

**iv.** **De i, on a trouvé que les deux groupes homogènes sont 2 et 3, on enlève donc les données du groupe 1 pour faire ACP.**

En SAS : **data** football2;

set football;

if PLAYER = **1** then delete;

**run**;

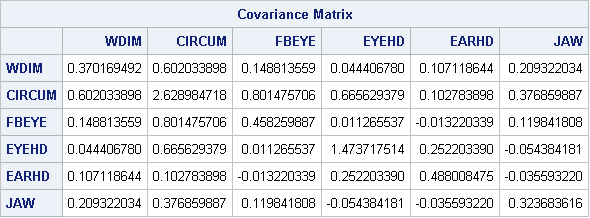
1. **En utilisant la matrice de variance-covariance.**

En SAS : **proc** **princomp** data=football2 out=res COV;

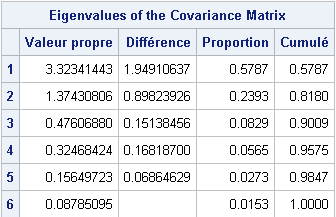
var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;

On obtient la matrice de covariance **S** :



On a les valeurs propres de **S** et les vecteurs propres associés:



On a donc les deux premiers composantes principales :   
z1 = **a**1**y** = 0.207y1 + 0.873 y2 + 0.261 y3 + 0.326y4 + 0.066y5 + 0.128y6,  
z2 = **a**2**y** = −0.142y1 − 0.219y2 − 0.231 y3 + 0.891 y4 + 0.222y5 − 0.187y6.

On Remarque que les plus grands coefficients dans z1 et dans z2 sont respectivement 0.873 et 0.891, ce sont des deux plus grandes variances en diagonal de matrice **S**. Les deux variables associées à ces variances sont y2 et y4, ils ont donc une influence notable sur les deux premières composantes principales.

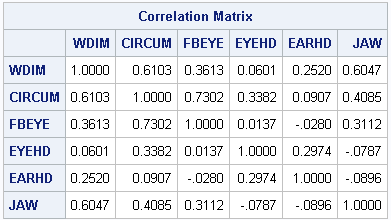
**En utilisant la matrice de corrélation.**

En SAS : **proc** **princomp** data=football2 out=res\_corr;

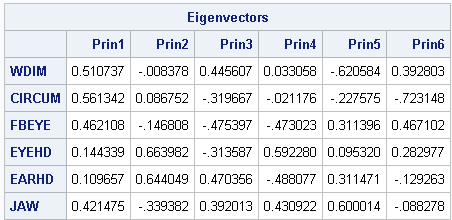
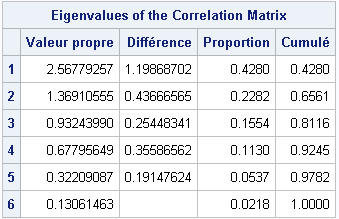
var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;

On obtient la matrice de corrélation **R** :



On a les valeurs propres de **R** et les vecteurs propres associés:



On a donc les deux premières composantes principales :   
z1 = **a’**1**y** = 0.511(y1-) + 0.561(y2-) + 0.462(y3-) + 0.144(y4-) + 0.110(y5-) + 0.421(y6-),  
z2 = **a’**2**y** = −0.008(y1-) + 0.087(y2-) − 0.147(y3-) + 0.664(y4-) + 0.644(y5-) − 0.339(y6-).

Les deux variables qui contribuent les plus en z1 et z2 sont CIRCUM sur z1 et EYEHD sur z2.

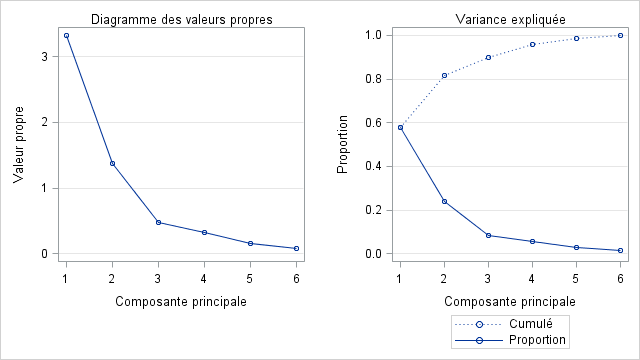
1. **En utilisant la matrice de variance-covariance.**

1ère règle : Les 2 premières composantes principales expliquent (*λ*1 + *λ*2)/ = 81.8% de la variation, on garde donc les 2 premières composantes.

2ème règle : = = 0.957

On a donc *λ*1, *λ*2 qui sont plus grand ou égale à, on garde donc z1 et z2.

3ème règle : Selon les graphes des valeurs propres, on a :



Il faut garder les *λ* précédant le pied de l’éboulis, on garde donc *λ*1 et *λ*2, ainsi les deux premières composantes pricipales.

4ème règle : comme *λ*1 et *λ*2 sont les deux valeurs propres plus grandes à 1, on les garde, ainsi on garde z1 et z2.

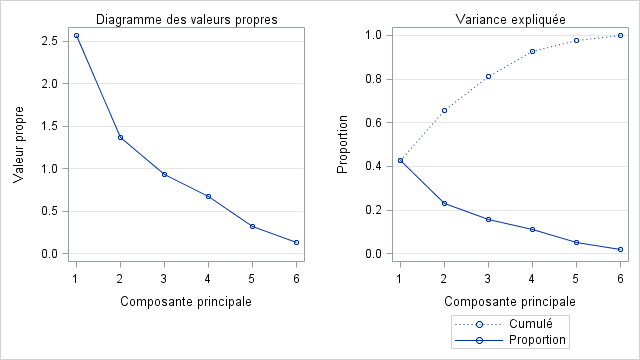
Conclusion : on garde les 2 premières composantes principales.

**En utilisant la matrice de corrélation.**

1ère règle : Cette règle n’applique pas pour la matrice de corrélation.

2ème règle : On a *λ*1, *λ*2 qui sont plus grand ou égale à 1, on garde donc z1 et z2.

3ème règle : Selon les graphes des valeurs propres, on a :



Il faut garder les *λ* précédant le pied de l’éboulis, mais le dessin n’est pas très claire, on peut garder *λ*1 et *λ2*.

4ème règle : Dans ce cas, il est pareil que la 2ème règle, on garde z1 et z2.

Conclusion : on garde les 2 premières composantes principales.

1. ACP en R est recommandé quand des variances des variables sont très différentes ou les unités de mesures des variables sont différentes, ici ce n’est pas le cas, on choisit donc ACP en S pour mieux interpréter notre analyse.
2. Les vecteurs propres de la matrice de covariance qu’on a sorti en a) nous montre que la deuxième composante principale diminue quand les variables WDIM, CIRCUM, FBEYE et JAW augmentent, et augment quand les variables EYEHD et EARHD augmentent. Cela répondre déjà notre question, mais pour être sure, on fait le test de corrélation pour la matrice de covariance et la matrice de corrélation.

En SAS : **proc** **corr** data=res;

var prin1 prin2;

with WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;

**proc** **corr** data=res\_corr;

var prin1 prin2;

with WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;



En regardant les coefficients de corrélation de la matrice de corrélation, on voit que pour la première composante, les p-value pour les variables WDIM, CIRCUM, FBEYE et JAW sont très petits, ainsi ces 4 variables sont des mesures verticales, tandis que les variables EYEHD et EARHD ont un p-value très petit pour la deuxième composante principale, on dit donc ces 2 variables sont des mesures horizontales. Ainsi cette analyse reflète le pattern des 6 variables.

**v.**

**a) Méthode composante principale**

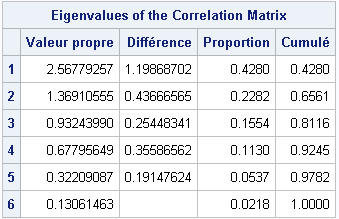
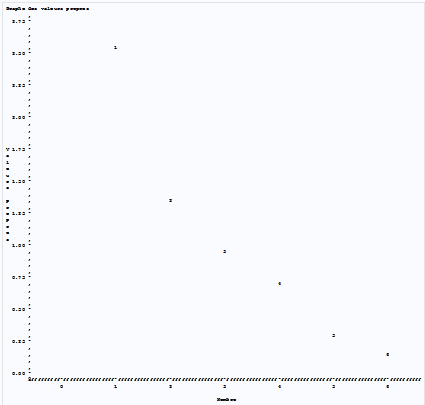
On fait une analyse en composante principale de la matrice de corrélations.

En SAS : **proc** **factor** data=football2 method=prin scree ;

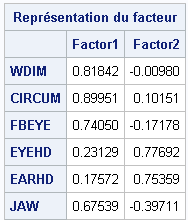
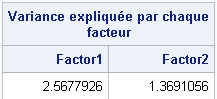
var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;

O a les vecteurs propres de la matrice de corrélation et le graphique de ces valeurs propres.

Par la 2ème règle, on garde donc les deux premiers facteurs.  
Par la 3ème règle, on garde aussi les deux premiers facteurs.

Toutes les variables sont bien interpreter par les deux facteurs, dont la variable CIRCUM est le mieux.

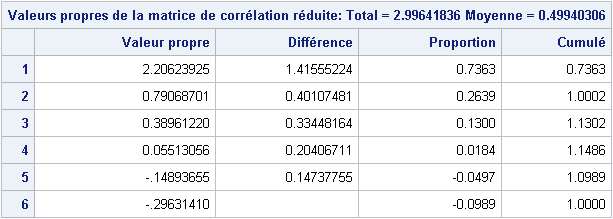
**Méthode principal factor analysis avec itérations**

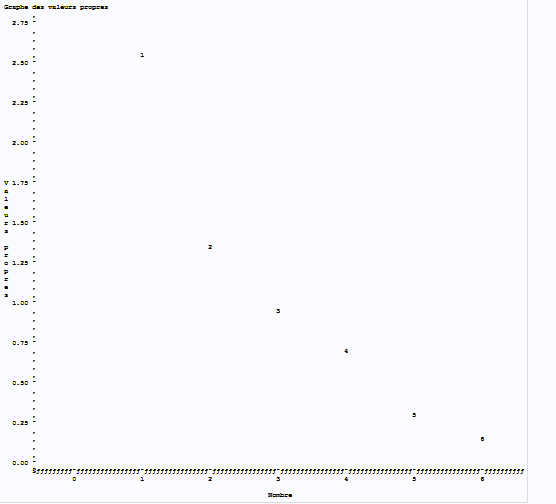
En SAS : **proc** **factor** data=football2 method=prinit scree ;

var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

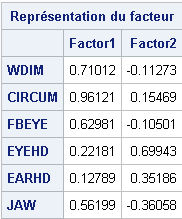
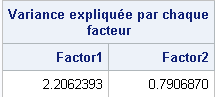
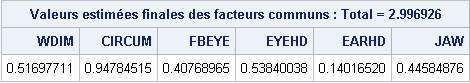
**run**;

O a les vecteurs propres de la matrice de corrélation et le graphique de ces valeurs propres.





Par la 2ème règle, comme il y a deux valeurs propres (2.206 et 0.791) qui sont plus grandes que la moyenne des valeurs propres(0.497), on garde donc 2 facteurs.  
Par la 3ème règle, on garde aussi les deux premiers facteurs.

Les variables FBEYE, EARHD et JAW ne sont pas bien interpreter par les deux facteurs.

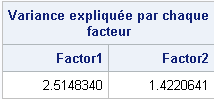
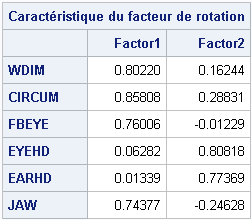
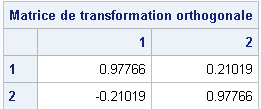
**b.**

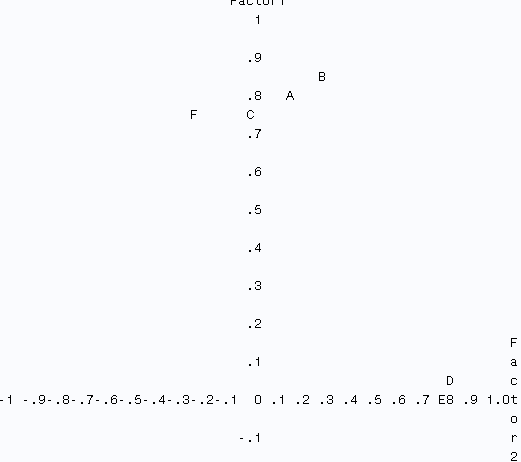
**Rotation varimax (Principal component analysis)**

En SAS: **proc** **factor** data=football2 rotate=varimax preplot plot;

var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;







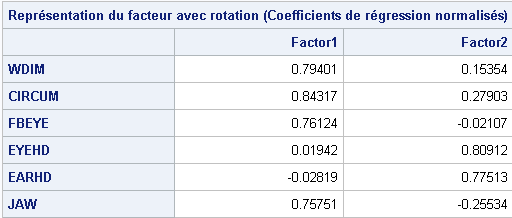
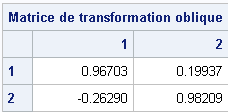
Selon les tableaux et le graphique, on peut dire que le facteur 1 regroupe les variables WDIM, CIRCUM, FBEYE et JAW, le facteur 2 regroupe les variables EYRHD et EYEHD.

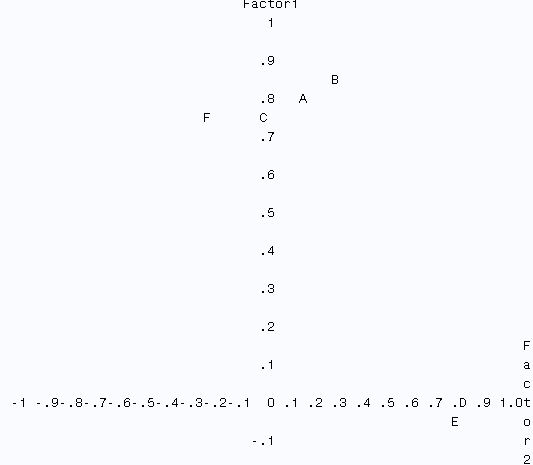
**Rotation oblique (Principal component analysis)**

En SAS : **proc** **factor** data=football2 method=prinit rotate=obvarimax preplot plot;

var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;





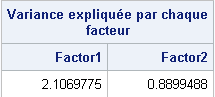
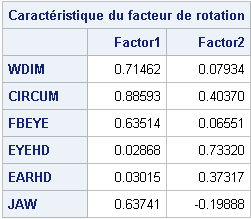
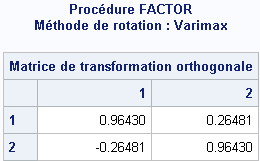
On a le même regroupement que la rotation varimax. Si on devrait choisir une méthode entre ces deux, on choisit la rotation varimax car elle garde toujours des axes des facteurs perpendiculaires.

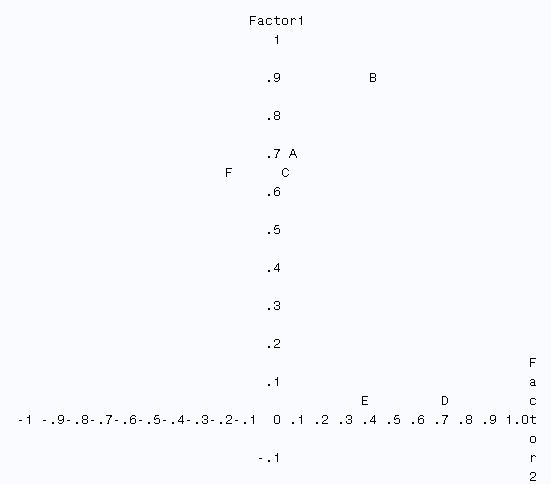
**Rotation varimax itérée (Principal factor analysis)**

En SAS : **proc** **factor** data=football2 method=prinit rotate=varimax preplot plot;

var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;





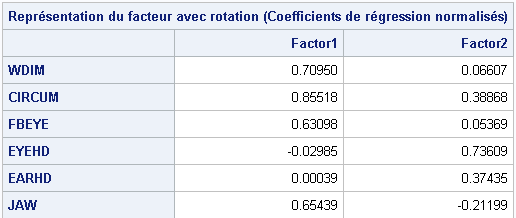
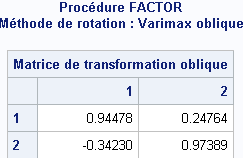
Selon les tableaux et le graphique, on peut dire que le facteur 1 regroupe les variables WDIM, CIRCUM, FBEYE et JAW, le facteur 2 regroupe les variables EYRHD et EYEHD.

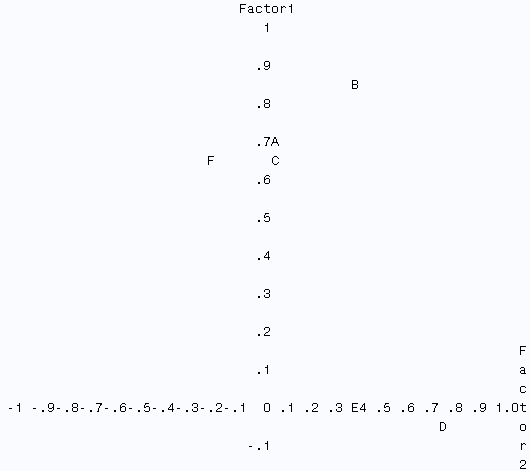
**Rotation oblique itérée (Principal factor analysis)**

En SAS : **proc** **factor** data=football2 method=prinit rotate=obvarimax preplot plot;

var WDIM CIRCUM FBEYE EYEHD EARHD JAW;

**run**;





On a le même regroupement que la rotation varimax itérée. Si on devrait choisir une méthode entre ces deux, on choisit la rotation varimax car elle garde toujours des axes des facteurs perpendiculaires.

Les deux différentes méthodes « principal component analysis » et « principal factors analysis » donnent le même résultat ici, on choisit la méthode principal component analysis car par cette méthode la représentation des facteurs sont plus importantes.

**c**. Dans cet exemple, on a le même pattern des variables en analyse factorielle que en analyse des composantes principales.