2. Analyse en composante principale

2.1. Exercice théorique

Cet exercice porte sur l’étude des correcteurs du jeu de données notes. Le nouveau jeu de donnée formé compte 8 individus : les correcteurs, et 4 variables : les moyennes et écart types des notes du médian et du final. Pour les correcteurs 2 et 8 certaines notes manquent, ainsi, dans une première partie l’étude se concentrera sur les 6 autres individus, puis dans une seconde partie une solution sera apportée pour pallier aux valeurs manquantes.

1. Calculer les axes factoriels de l’ACP du nuage de points défini par les quatre variables quantitatives. Quels sont les pourcentages d’inertie expliquée par chacun de ces axes ?

Le calcul des axes factoriels de l’ACP se fait par étapes pour la matrice suivante.

moy.median std.median moy.final std.final

Cor1 10.71 3.90 10.94 4.58

Cor3 10.23 3.04 13.43 4.34

Cor4 10.98 4.41 11.83 3.97

Cor5 11.50 4.30 13.41 4.88

Cor6 10.12 4.03 11.90 4.44

Cor7 10.74 4.65 11.40 4.87

La première étape consiste à centrer la matrice. Xc=

moy.median std.median moy.final std.final

Cor1 -0.01 -0.16 -1.21 0.07

Cor3 -0.48 -1.01 1.28 -0.17

Cor4 0.27 0.36 -0.32 -0.54

Cor5 0.79 0.25 1.26 0.36

Cor6 -0.59 -0.03 -0.25 -0.07

Cor7 0.03 0.59 -0.76 0.36

Ensuite, nous calculons la matrice d’inertie ou matrice de variances Xv grâce à la formule suivante :

Xv = (1/n)\*XcT\*Xc

Xv=

moy.median std.median moy.final std.final

moy.median 0.21 0.13 0.07 0.05

std.median 0.13 0.26 -0.23 0.05

moy.final 0.07 -0.23 0.91 0.01

std.final 0.05 0.05 0.01 0.10

Le calcul des valeurs propres : λ1 = 0.98, λ2 = 0.37, λ3 = 0.08, λ4 = 0.05 de la matrice de variances permet d’obtenir leurs vecteurs propres associés ou axes principaux d’inertie présentés dans la matrice U suivante :

U1 U2 U3 U4

[1,] -0.04 -0.70 -0.23 0.67

[2,] 0.29 -0.65 -0.09 -0.70

[3,] -0.96 -0.17 -0.02 -0.24

[4,] 0.00 -0.24 0.97 0.09

Il est ensuite possible de calculer le pourcentage d’inertie expliquée de chaque axe (dans l’ordre) : 66.10, 24.79, 5.61 et 3.51. Le pourcentage d’inertie expliquée par le premier plan factoriel (formé par les deux premiers axes) est 90.88.

2. Calculer les composantes principales ; en déduire la représentation des six individus dans le premier plan factoriel.

La matrice des composantes principales C se calcule à partir de la matrice centrée Xc et de la matrice des axes factoriels :

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

Cor1 1.11 0.30 0.11 0.40

Cor3 -1.50 0.81 0.01 0.06

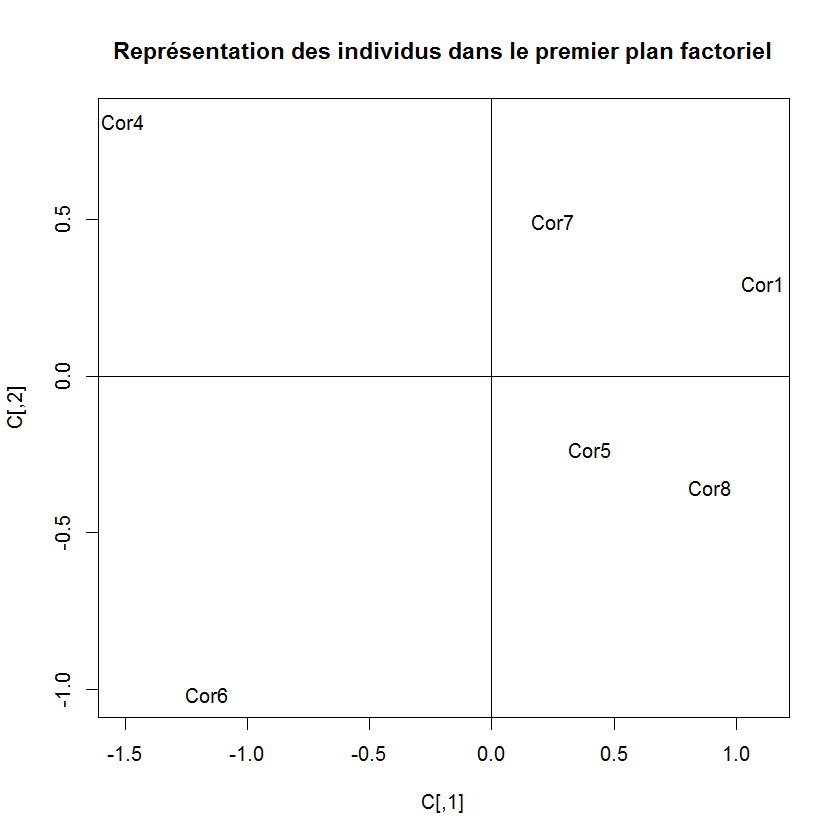
Cor4 0.40 -0.23 -0.61 -0.04

Cor5 -1.16 -1.02 0.12 0.08

Cor6 0.25 0.49 0.08 -0.32

Cor7 0.90 -0.35 0.30 -0.18

Ces composantes principales permettent d’obtenir la représentation des six individus dans le premier plan factoriel



La proximité des correcteurs 7 et 1 puis des correcteurs 5 et 8 sur cette représentation peut être interprétée comme le fait qu’ils ont un comportement similaire.

3. Tracer la représentation des quatre variables dans le premier plan factoriel.

Tout comme cela a été fait pour les individus, il est possible de représenter les variables en fonction des individus et de les analyser. Il n’est pas nécessaire de refaire tous les calculs fait précédemment, les axes factoriels et ainsi les composantes principales pour les variables se déduisent des axes factoriels de l’analyse de nuage de points-individus. Pour cela, la corrélation entre les vecteurs variable et les composantes principales normées des individus est étudiée et permet le calcul de la matrice suivante :

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

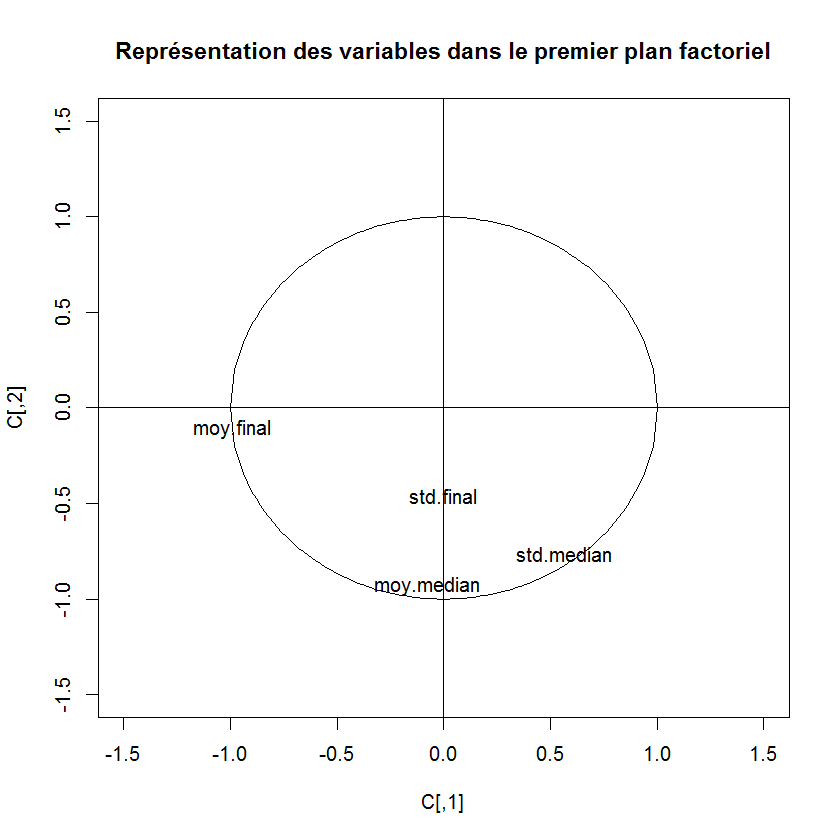
moy.median -0.08 -0.93 -0.15 0.33

std.median 0.57 -0.76 -0.05 -0.31

moy.final -0.99 -0.11 -0.01 -0.06

std.final 0.00 -0.46 0.89 0.06

A partir de cette matrice, il est possible de représenter les quatre variables dans le premier plan factoriel.



Les trois variables moy.median, std median et moy.final se trouvent sur le cercle de corrélation ce qui indique qu’elles sont bien représentées, ce qui n’est pas le cas de la variable std.final. Les variables moy.median et moy.final sont positionnées à angle droit ce qui indique qu’elles ne sont pas du tout corrélées.

4. Calculer l’expression pour les valeurs k = 1; 2 et 3. À quoi correspond cette somme lorsque k = 4 ?

K=1

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

[1,] -0.11 0.32 -1.06 0.07

[2,] 0.15 -0.44 1.43 -0.09

[3,] -0.04 0.12 -0.38 0.03

[4,] 0.12 -0.34 1.10 -0.07

[5,] -0.03 0.07 -0.24 0.02

[6,] -0.09 0.26 -0.85 0.06

K=2

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

[1,] 0.17 0.43 -1.05 0.11

[2,] 0.91 -0.16 1.44 0.02

[3,] -0.26 0.04 -0.39 -0.01

[4,] -0.83 -0.69 1.09 -0.21

[5,] 0.43 0.24 -0.23 0.08

[6,] -0.42 0.14 -0.86 0.01

K=3

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

[1,] 0.13 0.47 -1.03 0.20

[2,] 0.90 -0.15 1.44 0.03

[3,] -0.08 -0.24 -0.52 -0.51

[4,] -0.86 -0.63 1.11 -0.12

[5,] 0.41 0.28 -0.21 0.15

[6,] -0.51 0.27 -0.79 0.25

K = 4

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

[1,] 0.22 0.16 -1.12 0.42

[2,] 0.91 -0.20 1.43 0.06

[3,] -0.09 -0.21 -0.52 -0.53

[4,] -0.85 -0.70 1.09 -0.07

[5,] 0.34 0.53 -0.14 -0.04

[6,] -0.54 0.41 -0.75 0.15

La matrice obtenue avec k = 4 correspond à la matrice d’origine Xc.

5. On souhaite représenter les individus initialement écartés de l’ACP. Remplacer chacune de leurs valeurs manquantes par la moyenne de la variable correspondante (imputation par la moyenne), puis représenter ces individus dans les deux premiers plans factoriels.

Afin de pouvoir réaliser une ACP avec les 8 individus, il est possible de remplacer les valeurs manquantes par la moyenne de la variable correspondante. La matrice suivante présente les données ayant subi ce traitement et pour laquelle nous allons calculer les axes factoriels.

moy.median std.median moy.final std.final

Cor1 10.71 3.90 10.94 4.58

Cor2 12.01 3.71 12.21 4.39

Cor3 10.23 3.04 13.43 4.34

Cor4 10.98 4.41 11.83 3.97

Cor5 11.50 4.30 13.41 4.88

Cor6 10.12 4.03 11.90 4.44

Cor7 10.74 4.65 11.40 4.87

Cor8 10.90 4.01 12.57 3.65

L’obtention de la matrice des composantes principales avec ces données se fait de la même manière que précédemment, avec les 6 individus. Après calculs on obtient la matrice suivante qui nous permet de représenter tous les individus dans les deux premiers plans factoriels :

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

Cor1 1.21 -0.21 -0.12 0.43

Cor2 -0.20 0.93 -0.46 0.45

Cor3 -1.38 -0.93 -0.01 0.31

Cor4 0.45 0.15 -0.27 -0.45

Cor5 -1.08 0.75 0.62 -0.10

Cor6 0.38 -0.71 0.21 -0.08

Cor7 1.01 0.12 0.54 -0.07

Cor8 -0.39 -0.10 -0.52 -0.50

