Sprint Groupe 1 :

Léo Pichard, Thibault Breton, Thomas Chevalier, Mathieu Clavié

Question 1

1.1. Statistiques descriptives: typologie du travail à temps partiel chez les actifs

Limitez l'étude à la sous-population des actifs selon la variable FI, de 20 à 50 ans selon la variable AGCM. Combien reste-t-il d'individus?

Il reste 5026 individus, cf Annexe 1.1

1.2. Il est parfois nécessaire de recoder les facteurs comprenant de nombreuses modalités pour améliorer lisibilité et pertinence des synthèses:

(a) Recodez Région (RG) en 7 ou 8 "grandes régions" en perdant le moins possible l'information.

On a regroupé régions en 8 super-régions : Nord, Ouest, Sud_Ouest, Sud_Est, Est, Centre et IDF. l'île de france regroupant presque 20% des individus, elle se suffit à elle même. Annexe 1.2.1

(b) Recodez de même NR (Nationalité) en 4 "groupes" de nationalités cohérents.

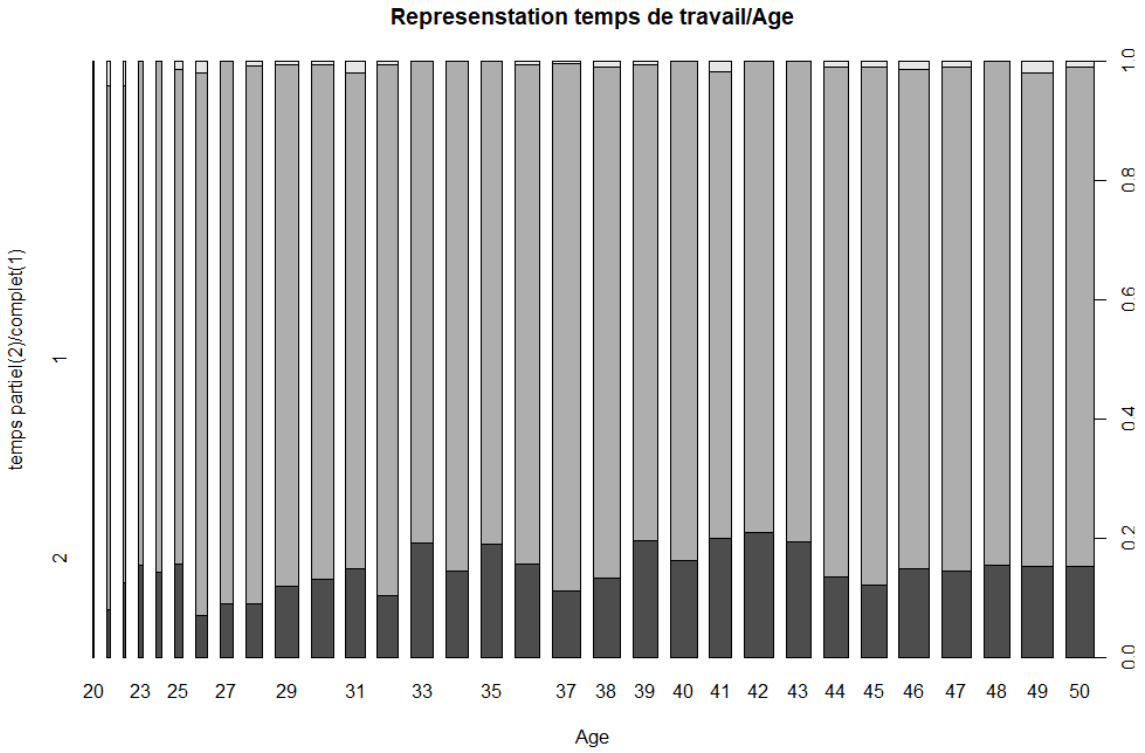
On regarde la réparation des individus selon leurs nationalité, et on se dit que l'on peut les regrouper en 4 groupe : Francais, Europe (= Italien, Espagnol, Portugais, Autre membre de la CEE, Europe de l'est), Afrique (= Afrique du nord, Autre pays d'Afrique) et autre (=Turc, Autre pays). cf Annexe 1.2.2

1.3. Analysez, avec des outils numériques et/ou graphiques du chapitre 1 du cours, les liens entre temps partiel/complet (TP) et les variables âge (AGCM), sexe (S), niveau de diplôme (DDIPL), nationalité (NR), région (RG).

Pour analyser les liens qu'il peut y avoir entre le temps de travail avec les différentes variables, on choisit d'utiliser une approche numérique combinée à une approche graphique. On réalise donc des tests du chi2 pour chaque comparaison, ainsi qu'une table de contingence que l'on représente ensuite par un spineplot

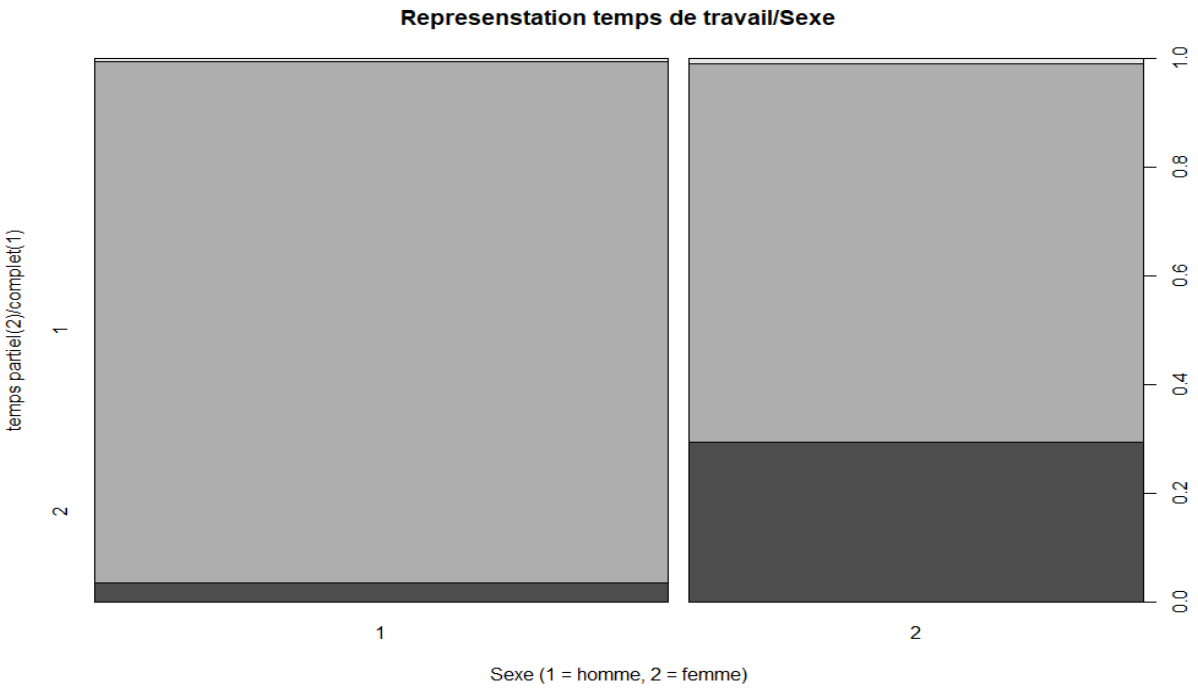
Lien entre temps partiel/complet et …

- … age :



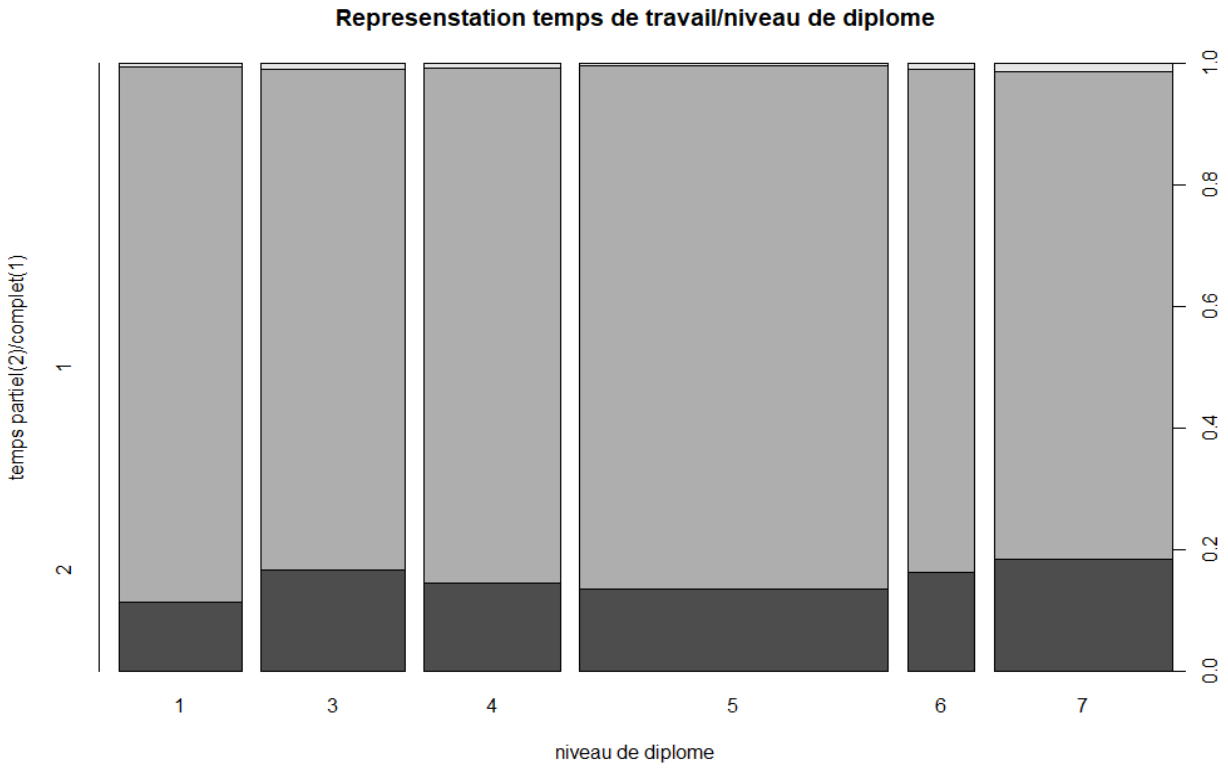
p-valeur chi2 : 0.04216 (on peut y voir une faible dépendance entre l'âge et le temps de travail, on peut dégager une tendance moyenne avec des exceptions peu prononcées selon les tranches d'âge)

- … sexe :



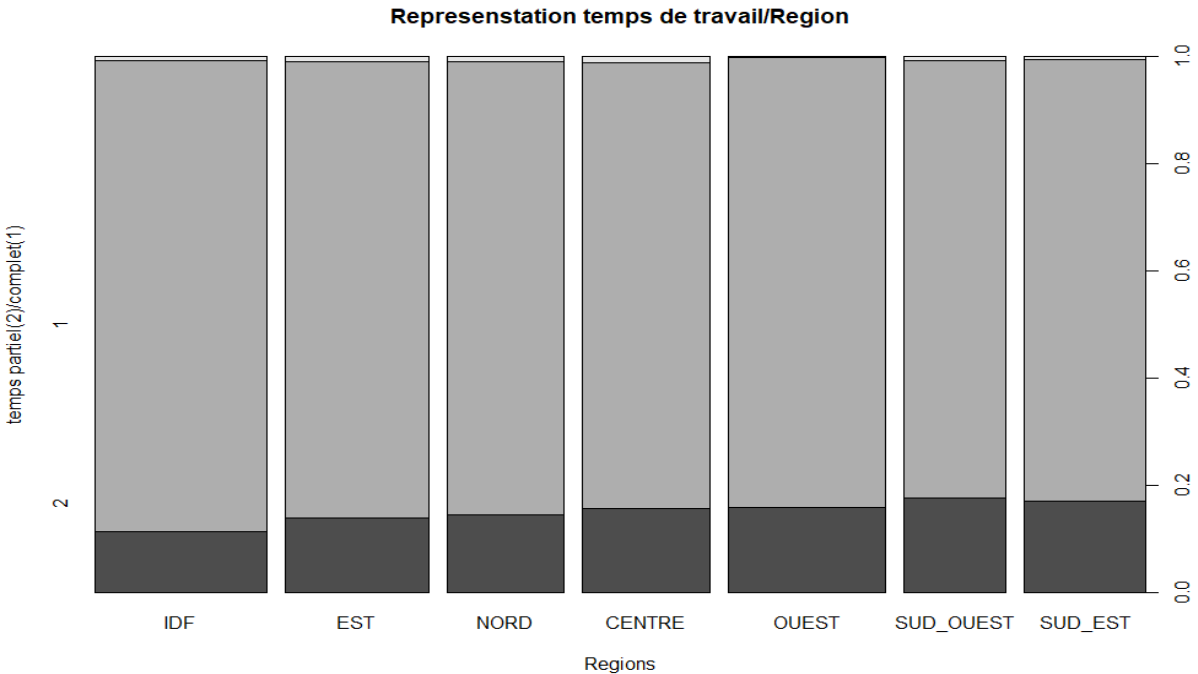
p-valeur chi2 : 2.2e-16()
On voit une très forte corrélation entre le sexe et le temps de travail, près de 95% des hommes ont un travail à temps complet contre 75% des femmes

- … niveau de diplôme :



p-valeur chi2 : 0.002633()
On peut voir, que le niveau de diplôme est proportionnel au temps de travail ; plus le niveau est élevé plus la part des temps partiels augmente. Ce niveau de corrélation est confirmé par le test du chi2

- … région :



p-valeur chi2 : 0.02062()
Il y a une faible corrélation entre les régions et le temps de travail, on peut voir que la région la plus active économiquement (IDF) est celle comprenant le moins de temps partiels.

- … nationalité :

p-valeur chi2 : 0.6206()
On ne voit aucune corrélation entre la nationalité et le temps de travail, la tendance visible dans le groupe Afrique est négligeable de par la faible quantité et donc représentativité de ses individus sur les données totales.
Annexe 1.3

1.4. Effectuez le ou les recodages appropriés afin de réaliser des tests du χ 2 pour l'hypothèse nulle "pas de lien" entre TP et chacun de ces facteurs. Interprétez les résultats afin de déterminer les facteurs les plus significatifs.

On choisi de faire le recodage de la catégorie Âge en 3 parties (20-30 ans, 30-40 ans et 40-50 ans) en utilisant les mêmes techniques que pour les questions précédentes, on

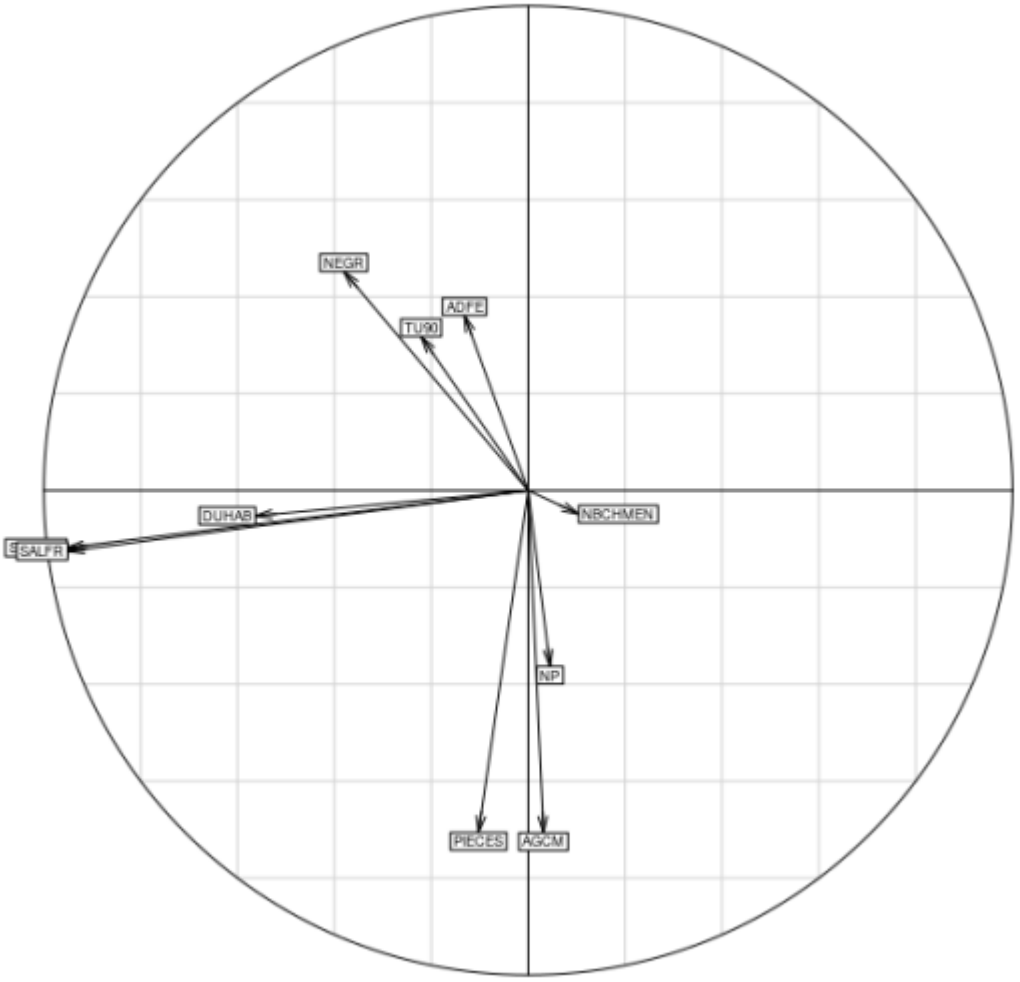
effectue à nouveau un test du chi2 avec les nouvelles valeurs et on obtient : 0.005996, on confirme donc la même analyse que lors de la question précédente, on rejette l'hypothèse d'indépendance entre l'Âge et le temps de travail. Le temps de travail est dépendant de la tranche d'âge.

<u>Question 2. ACP des variables quantitatives et qualitatives ordinales</u>
Il y a peu de variables quantitatives dans ces données, mais certaines variables qualitatives ont des relations
d'ordre naturelles entre modalités, ce qui permet de les considérer comme quantitatives. Ainsi on peut proposer
une ACP des 10 variables: AGCM, SALRED, SALFR, DUHAB, NP, NBCHMEN, PIECES, TU90, ADFE, NEGR, afin d'étudier les éventuels liens entre celles-ci et quelques facteurs qualitatifs déjà vus dans la question précédente. Un code de démarrage SPRINT Q2.R, qui a généré le fichier de données spécifique EmploiQ2.Rdata
vous est fourni pour cette question; dans ce code:

- **les variables utiles ont été sélectionnées, et les qualitatives ordinales converties en numérique**
- **Les lignes contenant des individus non renseignés (NA) pour les variables numériques ont été supprimées**

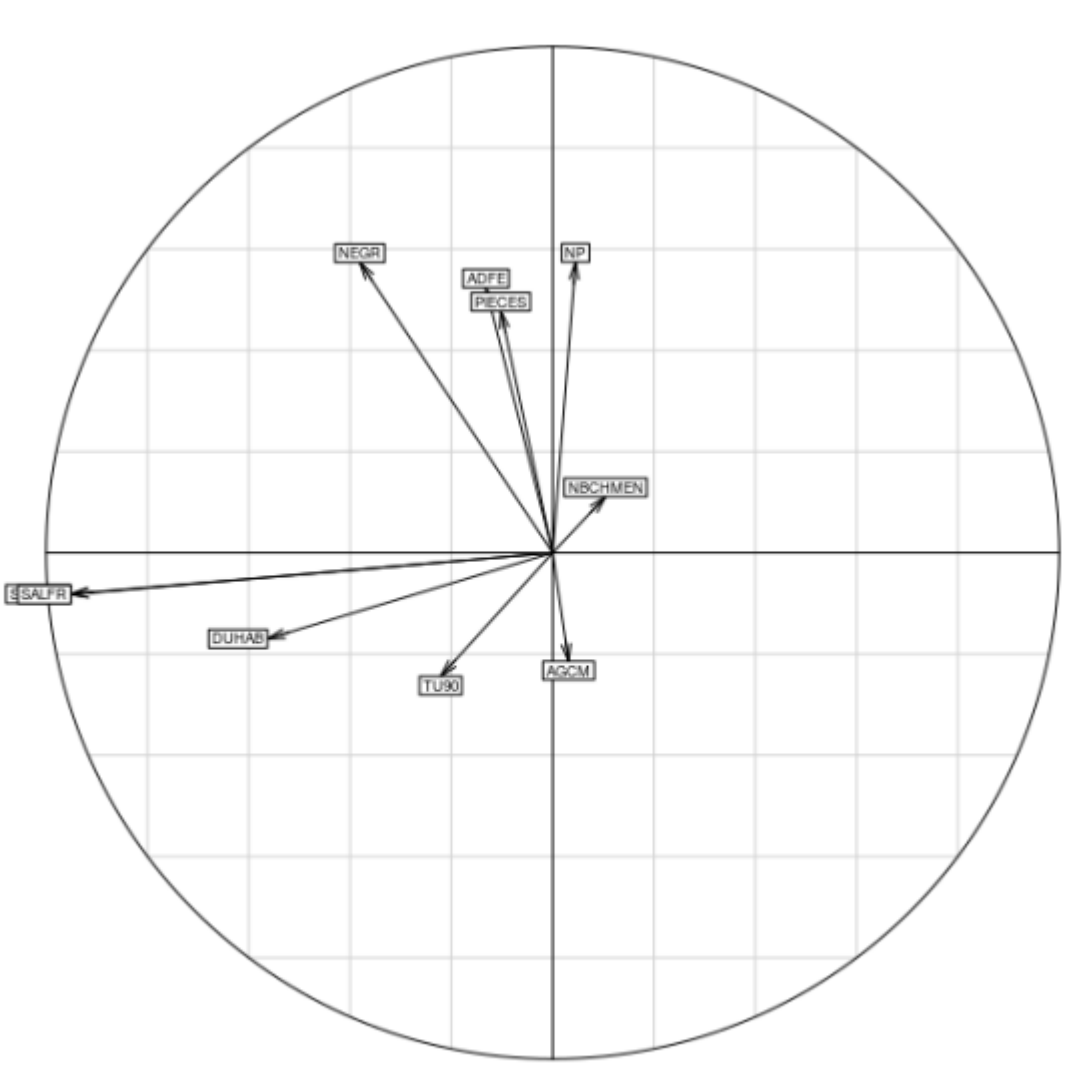
2.1. Sélectionner la sous-population composée des actifs selon la variable FI.
Annexe 2.1
2.2. Supprimer les individus correspondants à des salaires extrêmes, supérieurs au quantile d'ordre 99% de la
variable SALRED. Combien cela supprime-t-il d'individus? A votre avis, pourquoi faire cela?
Annexe 2.2
42 individus sont supprimés en ce faisant. Des données extrêmes sont aberrantes ou simplement fausses. Les prendre en compte induirait des moyennes et des écarts types incohérents. Sur l'exemple des salaires, un seul milliardaire a autant de poids dans les données, dans les calculs de moyenne, qu'un million de personnes au smic, un tel écart donnerait des interprétations non représentatives de la réalité.

2.3. Réaliser l'ACP sur la table résultat de ces traitements et interpréter les sorties.
Génération des ACP en Annexe 2.3



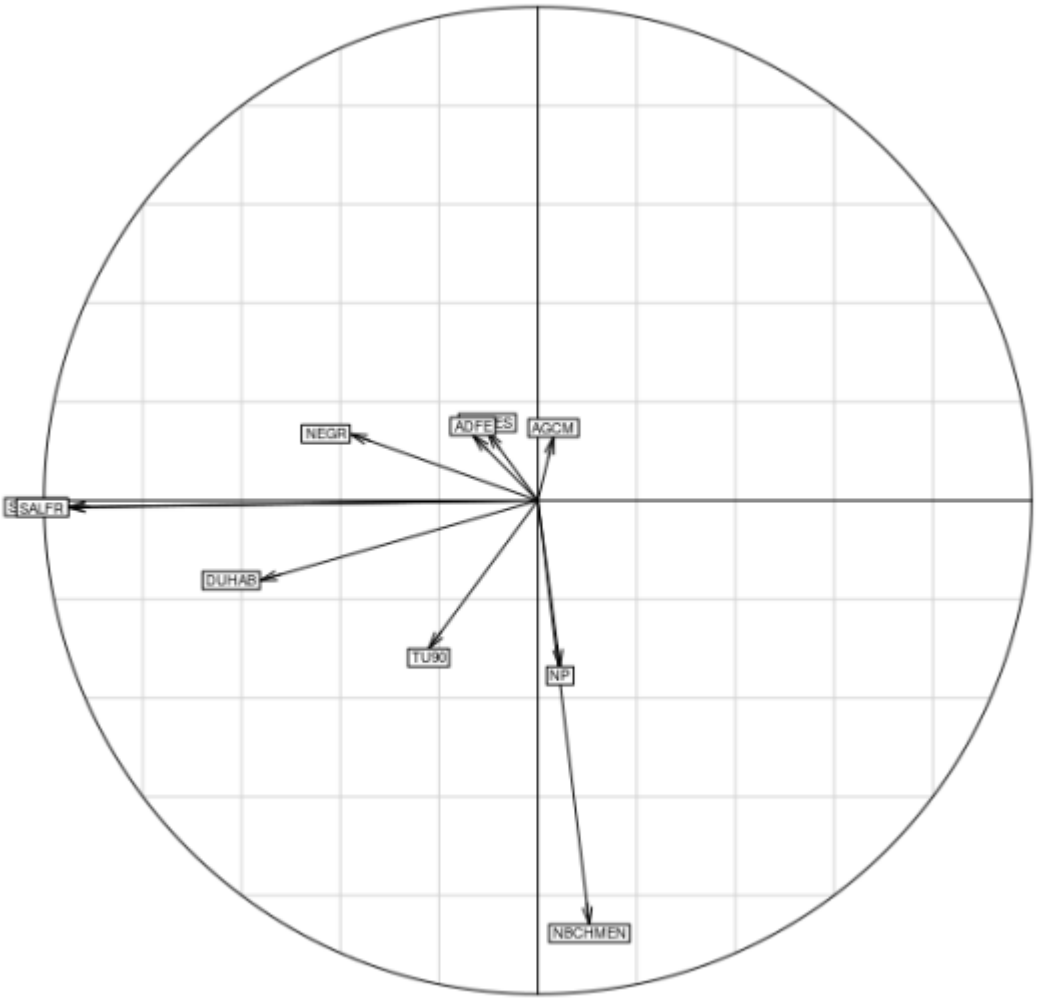
cercle de corrélations 1-2

On peut voir une très forte corrélation entre SALFR et SALRE et un manque total d’association de ces valeurs avec AGCM, avec laquelle le tracé des flèches forme un angle proche des 90°.



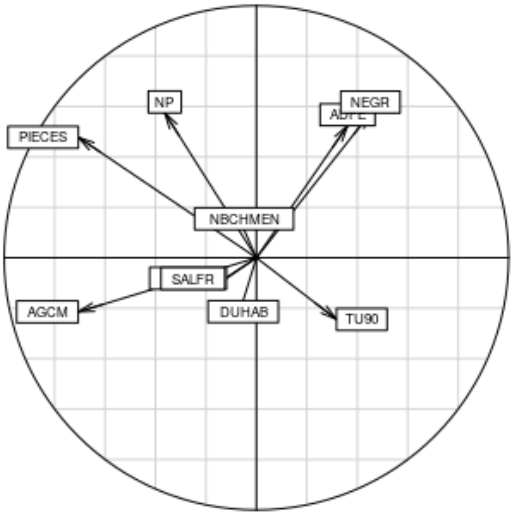
cercle de corrélations 1-3

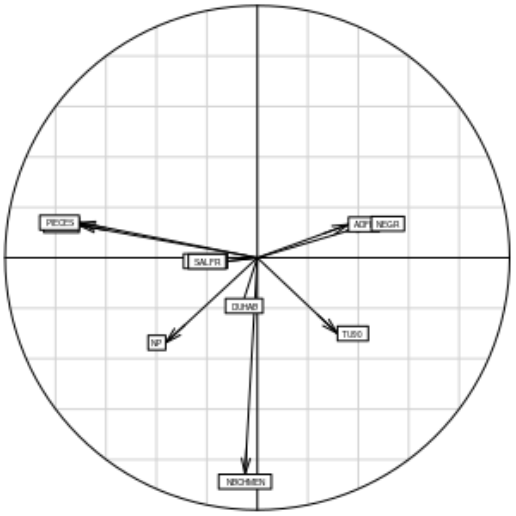
On peut retrouver la même interprétation sur la corrélation de SALFR et SALRE, mais les autres valeurs sont de poids trop faible pour être interprétables.



cercle de corrélations 1-4

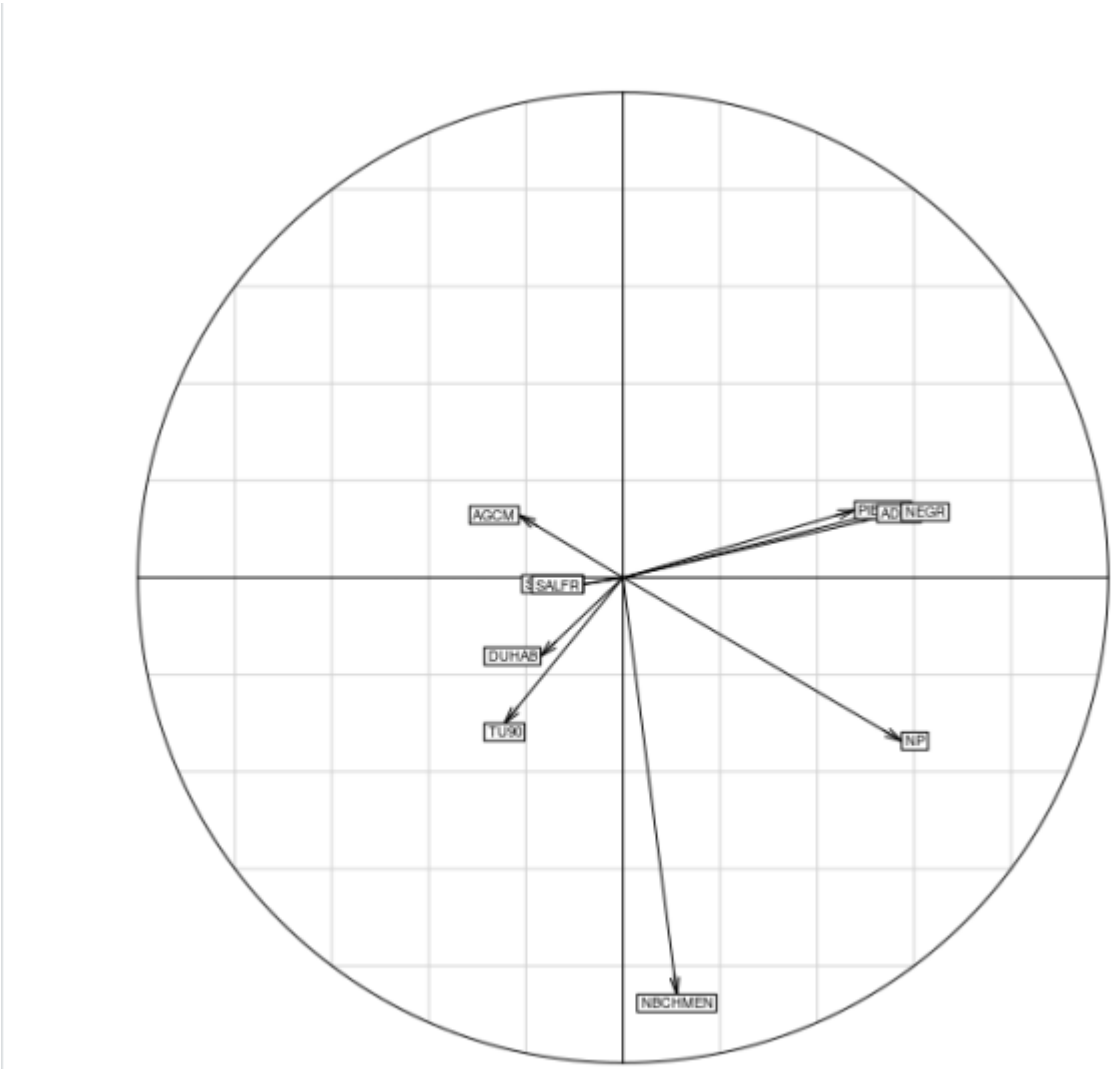
Sur ce cercle on retrouve l'interprétation de la corrélation de SALFR et SALRE, mais ici NBCHMEN est suffisamment représentatif pour être interprétable, et présente une forte dissociation avec SALFR et SALRE.





cercle de corrélations 2-3 cercle de corrélations 2-4

Sur l’axe 2 les données sont trop dispersées pour pouvoir en extraire une interprétation pertinente.



cercle de corrélations 3-4

On peut voir ici qu'une seul valeur est interprétable (NBCHMEN) mais n'est pas comparable aux autres car les poids des autres valeurs ne sont pas exploitables, ce qui se voit par leurs courtes flèches. On ne compare pas d'élément seul.

2.4. Représenter sur le plan principal (ψ 1 , ψ 2) et le plan factoriel (ψ 3 , ψ 4) les barycentres des projetés des individus par modalités des trois facteurs S, DDIPL, RG recodés comme pour la Question 1 (cf. techniques vues en TP).

image 4.1

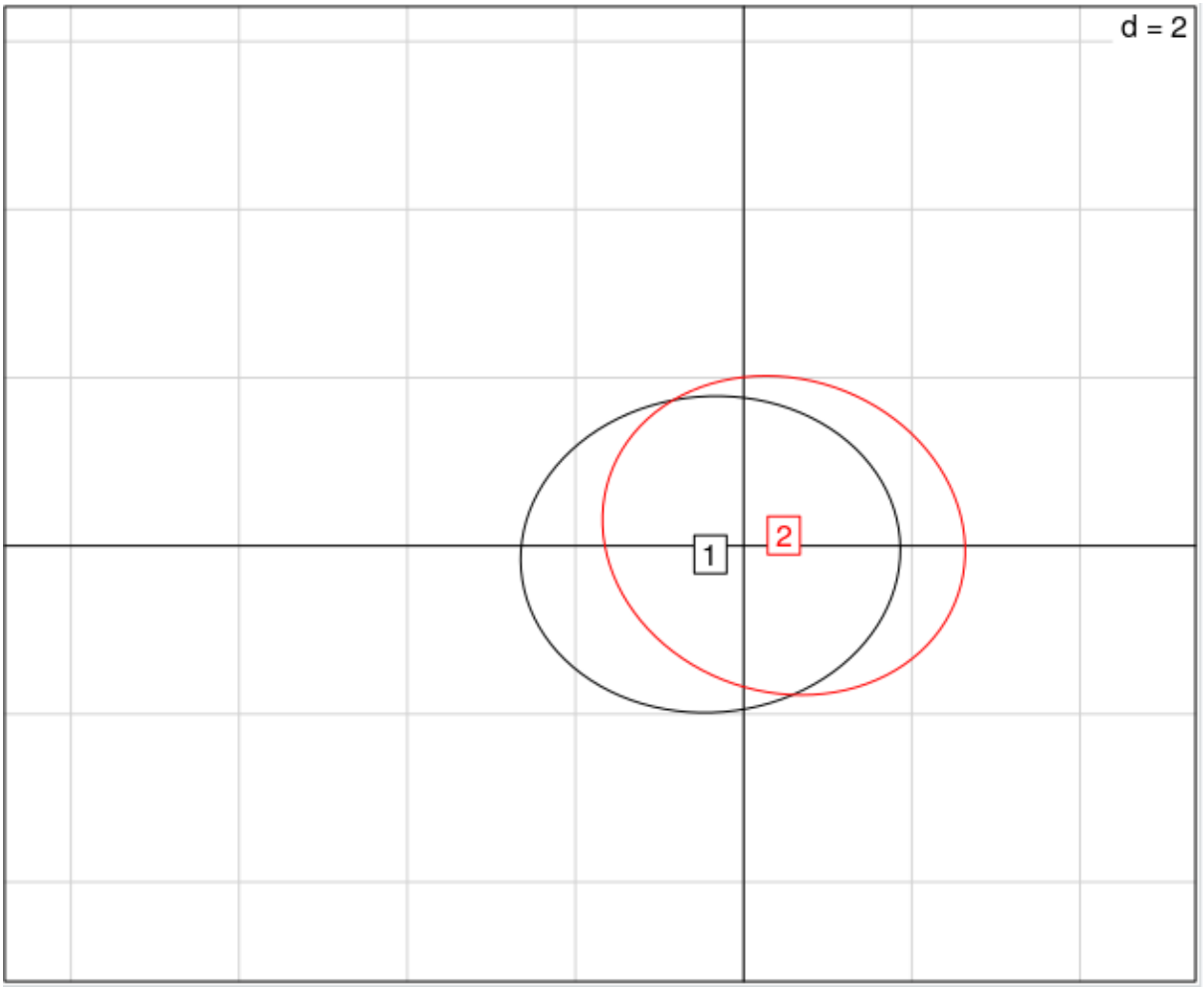


image 4.2

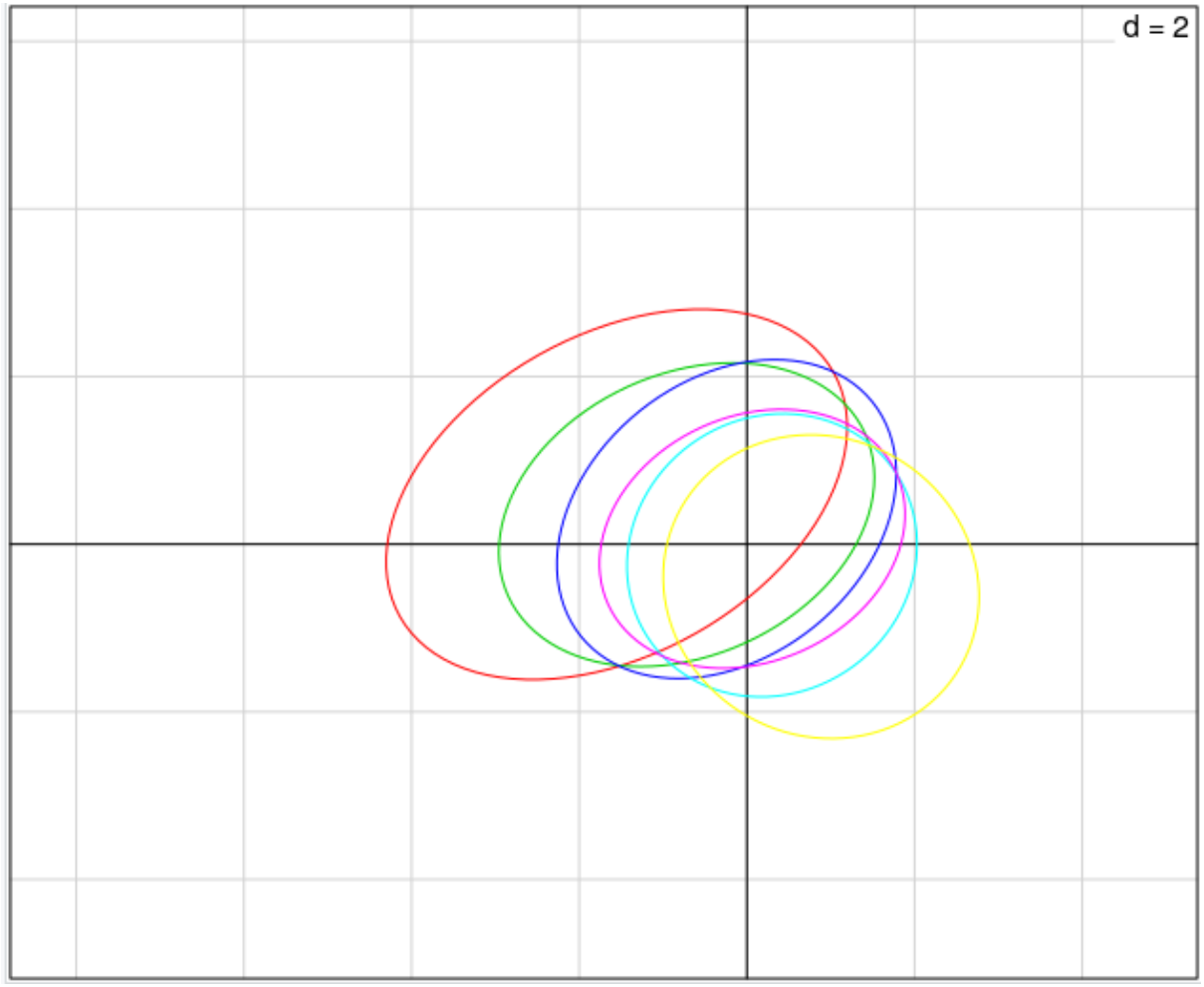
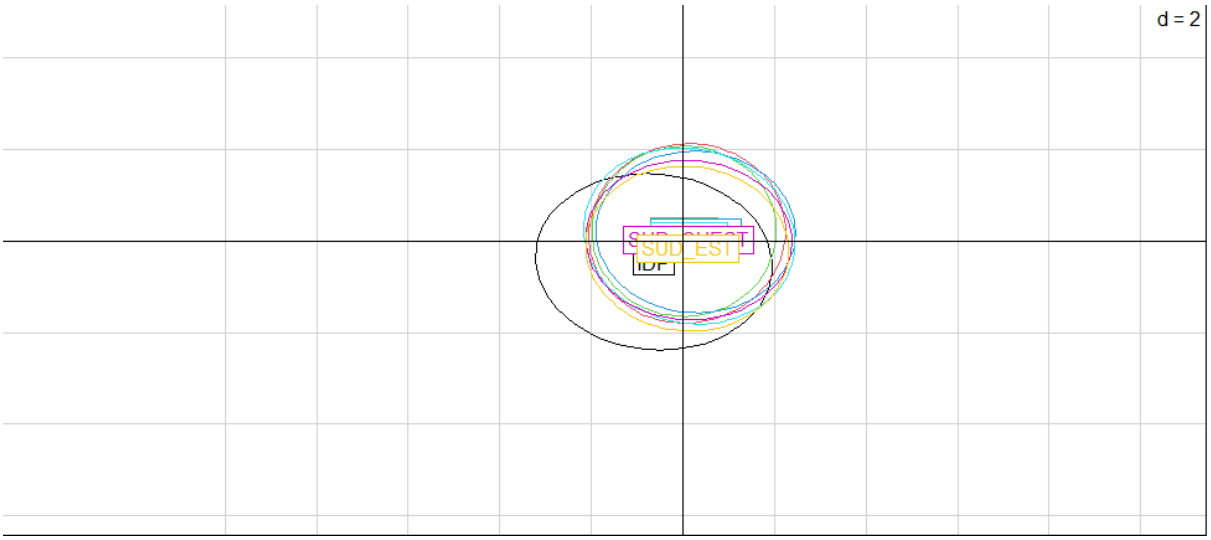


image 4.3



2.5. Interprétations et synthèse

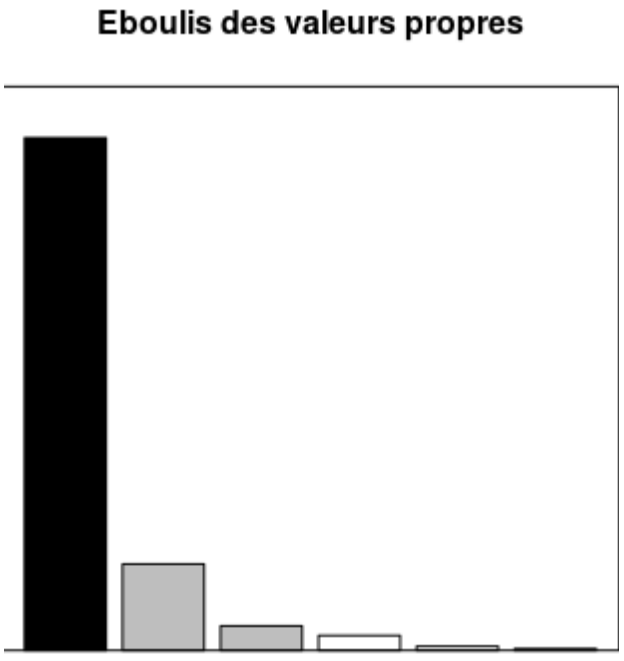
Les Corcircle sur les axes (2,X) et (3,X) n'étant pas exploitables, les résultats des barycentres seront interprétés selon les axes (1,X).
Sur le premier graphe, (4.1), le cercle des hommes est un peu plus décalé sur la gauche par rapport aux femmes. On peut en déduire une faible différence entre le salaire des hommes et des femmes. Les hommes ayant un salaire moyen légèrement supérieur aux hommes. Le fait que les 2 cercles partagent la grande majorité de leur aire et la forte proximité de leurs centres montrent une différence très peu prononcée.

Sur le dernier graphe (4.3) IDF (ellipse noire) se démarquant un peu du reste indique le plus grand niveau de salaire, mais la superposition des ellipses quelque soit la région indique une homogénéité des niveaux de salaire, qui présentent des caractéristiques, une répartition identique quelque soit la région.

3. AFC: liens entre diplôme et région

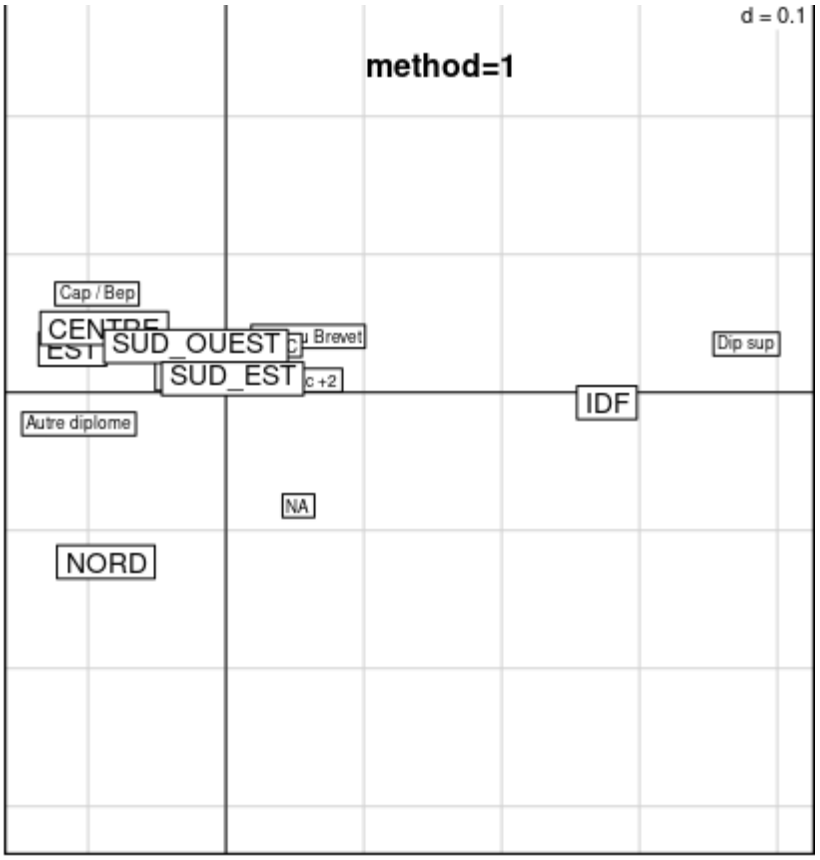
Voir annexe 3

Dans cette partie on s'intéresse à la sous-population composée des actifs, chômeurs, étudiants, retraités, selon la variable FI. Le code SPRINT Q3.R vous donne les instructions pour convertir correctement la table de contingence que vous obtiendrez vous-même, en data.frame nécessaire pour l'AFC. Réalisez une Analyse Factorielle des Correspondances entre les facteurs Diplôme (DDIPL) et région (RG) recodée comme pour la question 1. Faites-en l'interprétation et la synthèse.

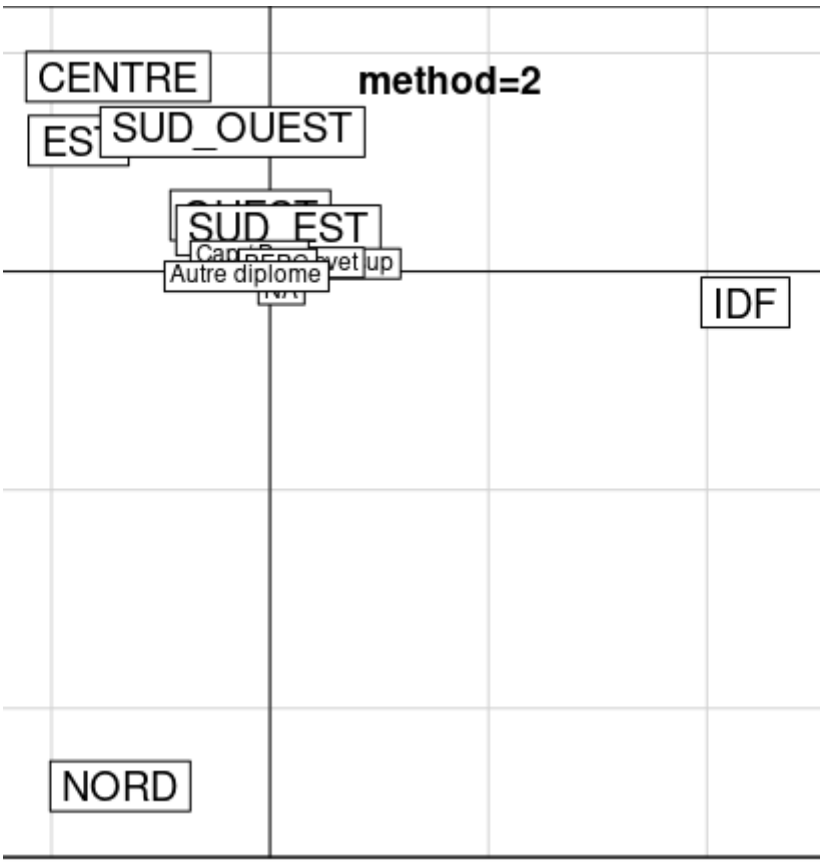


barplot 2.3

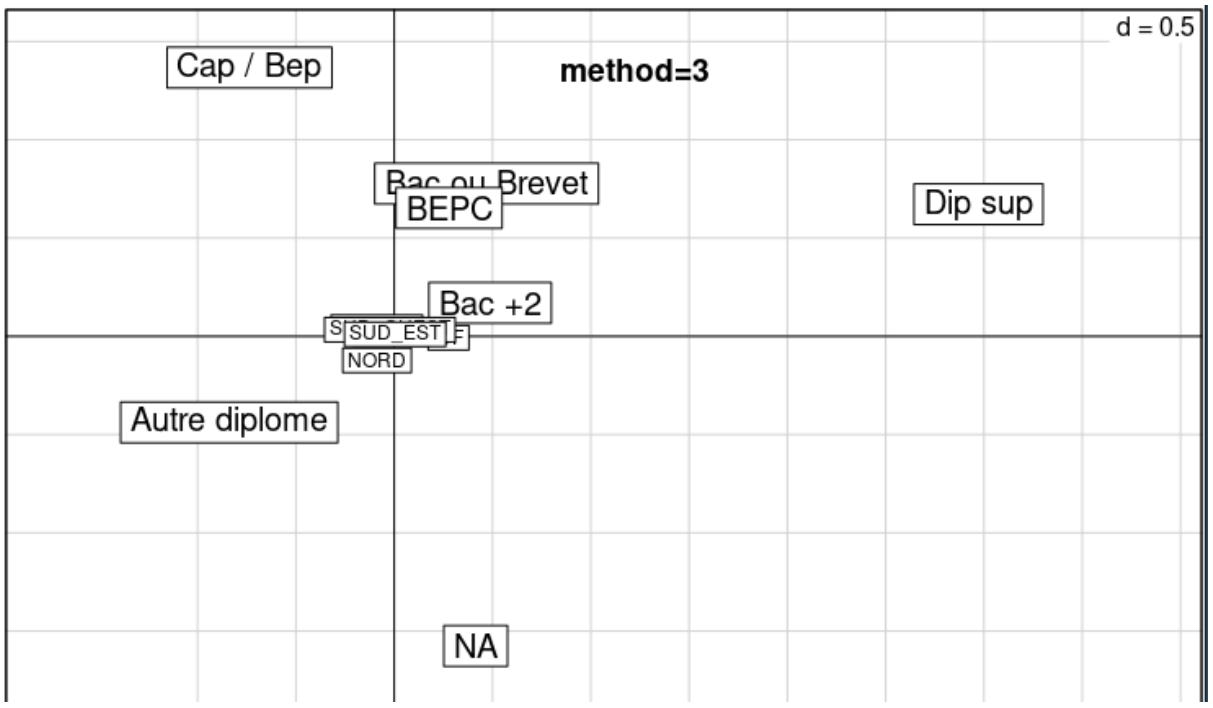
Ce graphe nous donne le nombre de catégories pertinentes permettant de classer les valeurs. On peut voir ici que 6 catégories sont pertinentes, l'une d'entre elle comprendra la majorité des valeurs tandis que les deux dernières seront des regroupements de valeurs minoritaires/exceptionnelles. On peut réduire notre analyse aux 3 axes de droite seulement, la perte des informations complémentaires apportées par les autres axes est acceptable.



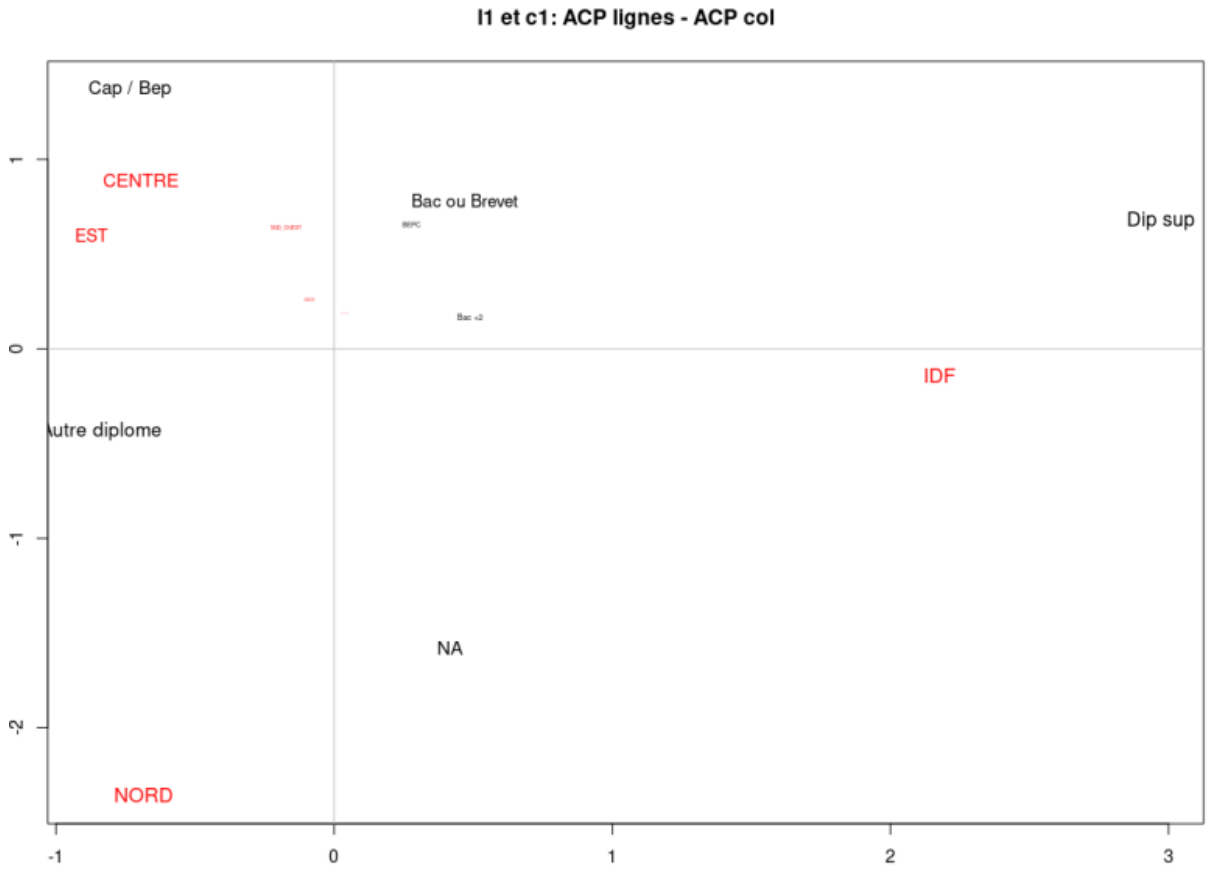
On peut voir sur ce graphe que selon la méthode 1, le diplôme Dip sup est fortement corrélé à l'IDF, tandis que les autres ne présentent que de très faible corrélations difficilement interprétables



Avec la méthode 2 on peut voir une très forte concentration des diplômes au centre et une faible disparité des Régions sur le graphe, la méthode 2 n'est pas interprétable au delà du statut "à part" de IDF et NORD.



On ne peut voir sur ce graphe que des déductions logiques, c'est-à-dire une opposition entre les possesseurs de dip sup et autres diplômes ou NA. c'est à dire que ceux en possession d'un diplôme de haut niveau est opposé aux "autre diplôme" ou des "NA". Aucune information géographique n'est interprétable ici.



On peut voir avec une acp des lignes et colonnes une corrélation entre IDF et Dip Sup, ainsi qu'une faible corrélation entre CENTRE, EST et le CAP/BEP. Il paraît aussi une corrélation inverse entre le NORD et le diplôme Dip Sup.

Annexe 1:

Annexe 1.1 :

```
# Pour manipuler plus simplement les variables
attach(dataQ1)
# On stocke les actifs dans un tableau
tab <- FI=="actifs"
# On crée un tableau de booleans selon la présence de valeur ou non des
individus (NA)
tab[is.na(tab)]<- FALSE
# Dans une variable on met le résultat du tri de dataQ1 selon les ligne
de notre tableau de booleans (si TRUE alors on garde la ligne)
dataQuetion1 <- dataQ1[tab,]
# On ne garde que les individus avec un âge (AGCM) compris entre 20 ans
et 50 ans
dataquestion1P2 <- dataQuetion1[dataQuetion1$AGCM <= 50 &
dataQuetion1$AGCM >= 20 ,]
# Affiche le résumé des valeurs, on peut y voir les 5026 actifs
summary(dataquestion1P2)
```

Annexe 1.2.1 :

```
#Découpage par région :
#Ile De France reste seule
IDF = dataquestion1P2[dataquestion1P2$RG==11,]
#SUD_OUEST regroupe Aquitaine, Midi Pyrénées et Languedoc-Roussillon
SUD_OUEST =
dataquestion1P2[dataquestion1P2$RG==72|dataquestion1P2$RG==73
|dataquestion1P2$RG==91,]
#SUD_EST regroupe Provence-Côte d'azur-corse, Rhône-Alpes
SUD_EST =
dataquestion1P2[dataquestion1P2$RG==92|dataquestion1P2$RG==82,]
#OUEST regroupe la Bretagne, Basse-Normandie, Pays de la Loire,
Poitou-Charente
OUEST =
dataquestion1P2[dataquestion1P2$RG==53|dataquestion1P2$RG==25|dataquesti
on1P2$RG==52|dataquestion1P2$RG==54,]
#Nord regroupe Haute Normandie, Nord pas de Calais, Picardie
NORD =
dataquestion1P2[dataquestion1P2$RG==23|dataquestion1P2$RG==31|dataquesti
on1P2$RG==22,]
#Centre regroupe Centre, Limousin, Auvergne, Bourgogne
CENTRE =
dataquestion1P2[dataquestion1P2$RG==24|dataquestion1P2$RG==74|dataquesti
on1P2$RG==83|dataquestion1P2$RG==26,]
#EST regroupe Lorraine, Alsace, Franche-Comté, Champagne-Ardenne
EST =
```

```
dataquestion1P2[dataquestion1P2$RG==41|dataquestion1P2$RG==42|dataquesti
on1P2$RG==43|dataquestion1P2$RG==21,]
```

Annexe 1.2.2 :

```
RG2 <- dataquestion1P2$RG
RG2 <- factor( RG2,labels=c("IDF", "EST", "NORD", "NORD", "CENTRE",
"OUEST", "CENTRE" ,"NORD", "EST", "EST", "EST", "OUEST", "OUEST",
"OUEST", "SUD_OUEST", "SUD_OUEST", "CENTRE", "SUD_EST", "CENTRE",
"SUD_OUEST", "SUD_EST"))
dataquestion1P2$RG <- RG2

# On regarde le nombre d'individu par nationalité
table(dataquestion1P2$NR)
# On stocke la répartition des individus par nationalité
NR2 <- dataquestion1P2$NR
table(NR2)
# On recode les facteurs nationalités en 4 groupes
NR2 <- factor(NR2, labels = c("France", "Afrique", "Afrique", "Europe",
"Europe", "Europe", "Europe", "Europe", "Autres", "Autres"))
# On contrôle notre répartition
table(NR2)
# On contrôle la répartition avant le recodage
table(dataquestion1P2$NR, NR2)
# On remplace la variable NR par notre nouvelle variable
dataquestion1P2$NR <- NR2
# Contrôle que la nouvelle variable soit bien dans la table
summary(dataquestion1P2$NR)
```

Annexe 1.3

```
# création de la table de contingence
contingence = table(dataquestion1P2$TP, dataquestion1P2$AGCM )
#test du chi2 pour voir si on rejette ou non l'hypothèse d'indépendance
chisq.test(dataquestion1P2$TP,dataquestion1P2$AGCM)
# affichage de la répartition générée par le tableau de contingence du
temps de travail en fonction de l'âge
spineplot(t(contingence), xlab = "Age", ylab = "temps
partiel(2)/complet(1)")
title("Representation temps de travail/Âge")

# affichage de la répartition générée par le tableau de contingence du
temps de travail en fonction du sexe
contingence2 = table(dataquestion1P2$TP, dataquestion1P2$S )
chisq.test(dataquestion1P2$TP,dataquestion1P2$S)
spineplot(t(contingence2), xlab = "Sexe (1 = homme, 2 = femme)", ylab =
"temps partiel(2)/complet(1)")
```

```
title("Representation temps de travail/Sexe")

# affichage de la répartition générée par le tableau de contingence du
temps de travail en fonction du niveau de diplôme
contingence3 = table(dataquestion1P2$TP, dataquestion1P2$DDIPL )
chisq.test(dataquestion1P2$TP,dataquestion1P2$DDIPL)
spineplot(t(contingence3), xlab = "niveau de diplome", ylab = "temps
partiel(2)/complet(1)")
title("Représentation temps de travail/niveau de diplôme")

# affichage de la répartition générée par le tableau de contingence du
temps de travail en fonction de la région en france
contingence4 = table(dataquestion1P2$TP, dataquestion1P2$RG)
chisq.test(dataquestion1P2$TP,dataquestion1P2$RG)
spineplot(t(contingence4), xlab = "Regions", ylab = "temps
partiel(2)/complet(1)")
title("Représentation temps de travail/Région")

# affichage de la répartition générée par le tableau de contingence du
temps de travail en fonction de la nationalité
contingence5 = table(dataquestion1P2$TP, dataquestion1P2$NR )
chisq.test(dataquestion1P2$TP,dataquestion1P2$NR)
spineplot(t(contingence5),  xlab = "Nationalités", ylab = "temps
partiel(2)/complet(1)")
title("Représentation temps de travail/Nationalité")
```

Annexe 1.4

```
AgeCut <- cut(as.numeric(dataquestion1P2$AGCM), breaks = c(20, 30, 40,
Inf), labels = c("20-30 ans", "30-40 ans", "40-50 ans")) #création des 3
classes d’âges

#affectation des nouvelles classes d’âge
dataquestion1P2$AGCM = AgeCut
chisq.test(dataquestion1P2$TP, dataquestion1P2$AGCM)
```

Annexe 2

Annexe 2.1 et 2.2:

```
On ne sélectionne que les actifs
dataquestion1 = dataQ2[dataQ2$FI=="actifs",]
#parmi eux on retire les 1% des plus haut revenus (30000 et plus par
mois)
qd = quantile(dataquestion1$SALRED,0.99)
```

```
qd
#on applique cette sélection aux données
dataquestion2 = dataquestion1[dataquestion1$SALRED<=qd,]
nrow(dataquestion1)
nrow(dataquestion2)
#on affiche le boxplot, la boîte à moustache, les quartiles de ces
données.
boxplot(dataquestion2$SALRED)
# On a une différence de 42 personnes, ces valeurs peuvent être
considérées aberrantes ou fausses.
```

Annexe 2.3

```
#sur le barplot 2.3 on voit un plateau à 4-5 -> on prend nf = 4

acp=dudi.pca(dataquestion2[,c(1:10)],scannf=F, nf=4)
barplot(acp$eig)#création des cercles de corrélation
par(mfrow=c(2,2))
s.corcircle(acp$co, 1,2 , clabel = 0.5, sub = "cercle de corrélations
1-2")
s.corcircle(acp$co, 1,3 , clabel = 0.5, sub = "cercle de corrélations
1-3")
s.corcircle(acp$co, 1,4 , clabel = 0.5, sub = "cercle de corrélations
1-4")
par(mfrow=c(1,2))
s.corcircle(acp$co, 2,3 , clabel = 0.5, sub = "cercle de corrélations
2-3")
s.corcircle(acp$co, 2,4 , clabel = 0.3, sub = "cercle de corrélations
2-4")
#Angle entre NBCHMEN & PIECES &
par(mfrow=c(1,1))
s.corcircle(acp$co, 3,4 , clabel = 0.5, sub = "cercle de corrélations
3-4")
#
```

Annexe 2.4

```
#on crée des représentations selon différentes sélection de colonnes
#image 4.1
s.class(acp$li,fac=dataquestion2$S,cstar=0,cpoint=0,col =
coltype,axesell = FALSE)
#image 4.2
s.class(acp$li,fac=dataquestion2$DDIPL,cstar=0,cpoint=0,col =
seq(length(levels(dataquestion2$DDIPL))),axesell = FALSE)
```

```r
## Enquête emploi 2001 INSEE
## sujet SPRINT 2021
#####################################################
## Code de démarrage pour la Question 2
## ACP des variables quantitatives et qualitatives ordinales

## Sélection des colonnes
rm(list = ls()) # vider le workspace (repart de zéro)
load("Emploi01.Rdata") # Charger le fichier de data
# Sélection des variables utiles pour la Question 2
# 10 num (ordinales) et 5 facteurs qual suppl
varQ2 <- c("AGCM", "SALRED", "SALFR", "DUHAB", "NP", "NBCHMEN",
           "PIECES", "TU90", "ADFE", "NEGR",
           "S", "FI", "DDIPL", "RG")
dataQ2 <- Emploi01[, varQ2] # sélection des variables

dataQ2
## recodage de FI cf Q1
FI2 <- dataQ2$FI; FI2[FI2 == ""] <- NA
FI2 <- factor(FI2, labels = c("actifs", "chômeur", "étudiant",
"militaire",
                              "retraité", "retiré", "ffoyer",
"inactif"))
table(dataQ2$FI, FI2, useNA = "ifany") # vérifier le recodage
dataQ2$FI <- FI2

summary(dataQ2) # nombreux NA dans SALRED...

### recodage des factors ordinaux en "numeric":
colnum <- 1:10   # colonnes concernées
for (j in colnum) dataQ2[,j] <- as.numeric(as.character(dataQ2[,j]))

# suppression des NA:
dataNA <- is.na(dataQ2)
NArow <- apply(dataNA, 1, sum)
sum(NArow == 0) # 4159 lignes (obs) complètes = sans NA
dataQ2 <- dataQ2[NArow == 0, ]
summary(dataQ2)
save(dataQ2, file="EmploiQ2.Rdata")
dataQ2

dataquestion1 = dataQ2[dataQ2$FI=="actifs",]
qd = quantile(dataquestion1$SALRED,0.99)
qd
#qd = quantile(dataquestion1$SALRED,0.01)
dataquestion2 = dataquestion1[dataquestion1$SALRED<=qd,]
```

```r
RG2 <- dataquestion2$RG
RG2 <-
factor(RG2,labels=c("IDF","EST","NORD","NORD","CENTRE","OUEST","CENTRE",
"NORD","EST","EST","EST","OUEST","OUEST","OUEST","SUD_OUEST","SUD_OUEST"
,"CENTRE","SUD_EST","CENTRE","SUD_OUEST","SUD_EST"))
dataquestion2$RG <- RG2

nrow(dataquestion1)
nrow(dataquestion2)
boxplot(dataquestion2$SALRED) # On a une différence de 42 personnes, ces
valeurs peuvent être considérées aberrantes ou fausses.
library(ade4)
dataquestion2
dataquestion2[,c(1:10,13)]

newtab = dataquestion2[,c(1:10)]
newtab[c(1,4:10)]
?dudi.pca

apply(newtab,2,sd)
x=scale(newtab,center = TRUE,scale = TRUE)

par(mfrow=c(1,2))
newtab
acp = dudi.pca(x)

Imen <- inertia.dudi(acp, col.inertia=F, row.inertia=TRUE)
Imen$TOT
print(acp$eig/3,3)
(acp$eig/10)
cumsum(acp$eig/10)

acp=dudi.pca(x,scannf=F,nf=6)
Imen <- inertia.dudi(acp, col.inertia=F, row.inertia=TRUE)
Imen$TOT

plot(acp$eig,type="l")
s.corcircle(acp$co,clabel = 0.5)

##4
dataquestion2
datapouracp = dataquestion2[,1:10]
acp = dudi.pca(df = datapouracp, scannf = FALSE, nf = 3,scale = TRUE)
plot(acp$li$Axis1,acp$li$Axis2,type="n")
coltype = seq(length(levels(dataquestion2$S)))
par(mfrow=c(1,1))
```

```
#ici est créé le barycentre des cercles avec le premier axe
s.class(acp$li,fac=dataquestion2$S,cstar=0,cpoint=0,col =
coltype,axesell = FALSE)
#ici est créé le barycentre des niveaux des diplômes
s.class(acp$li,fac=dataquestion2$DDIPL,cstar=0,cpoint=0,col =
seq(length(levels(dataquestion2$DDIPL))),axesell = FALSE)

#ici est créé le barycentre des différentes Regions
s.class(acp$li,fac=dataquestion2$RG,cstar=0,cpoint=0,col =
seq(length(levels(dataquestion2$RG))),axesell = FALSE)


#plot(acp$li$Axis1, acp$li$Axis2, type="n",xlab = "Axe principal  1",
ylab = "Axe principal 2")

s.class()
dataquestion1$RG
?s.class
abline(h=0); abline(v=0) # barycentre du nuage
```

Annexe 3 :

```
# On recode les régions en 7 groupe comme à la question 1
varQ1 <- c("FI", "TP", "AGCM", "S", "DDIPL", "NR", "RG")
dataQ1 <- Emploi01[, varQ1]
attach(dataQ1)
RG2 <- RG
RG2 <-
factor(RG2,labels=c("IDF","EST","NORD","NORD","CENTRE","OUEST","CENTRE",
"NORD","EST","EST","EST","OUEST","OUEST","OUEST","SUD_OUEST","SUD_OUEST"
,"CENTRE","SUD_EST","CENTRE","SUD_OUEST","SUD_EST"))
RG <- RG2

#actifs, chômeurs, étudiants, retraités
#1 2 3 5
# On conseve que les individus qui sont : soit actifs, soit chômeurs,
soit étudiants, soit retraités
dataQ1$FI
dataQ1 <- dataQ1[dataQ1$FI == 1 | dataQ1$FI == 2 | dataQ1$FI == 3 |
dataQ1$FI == 5 ,]
dataQ1
# recode les diplômes
DDIPL2 <- DDIPL
DDIPL2 <- factor(DDIPL2,labels=c("NA", "Dip sup", "Bac +2", "Bac ou
Brevet", "Cap / Bep", "BEPC", "Autre diplome"))
```

```r
DDIPL <- DDIPL2



tb <- table(DDIPL, RG)
tbdf <- as.data.frame(unclass(tb))  # convertir en data frame pour afc !
class(tbdf) # "data.frame"
tbdf # doit redonner la table de contingence; mais autre format

## puis AFC...



afc <-dudi.coa(tbdf, scan = FALSE) # df as input
summary(afc) # donne la répartition de l'inertie (call inertia.dudi)

afcin <- inertia.dudi(afc,col.inertia=T,row.inertia=T)
# Decomposition de l'inertie
afcin$tot.inertia

# On garde 3 colonne pour avoir 96% d'inertie
afc <-dudi.coa(tbdf, scan = FALSE, nf=3)
afcin <- inertia.dudi(afc,col.inertia=T,row.inertia=T)
afcin$tot.inertia

# Contrôle visuel de la règle du code, car on a choisi avant 3 colonnes
(valeur de boulies)
scatterutil.eigen(afc$eig,nf=3,box=T,sub="")
title("Eboulis des valeurs propres")



# Test chi^2
tind <- chisq.test(tbdf)
tind$stat/afc$N
sum(afc$eig)
# on est a 0.020 donc on rejette l'hypothèse H0, il n'y a pas de
indépendance


# REPRESENTATIONS PLANS FACTORIELS: defaults functions from ade4

par(mfrow=c(1,1))

scatter(afc, method=1, posieig="none"); title("method=1")
scatter(afc, method=2, posieig="none"); title("method=2")
```

```r
scatter(afc, method=3, posieig="none"); title("method=3")
```

```r
## Voir le contenu de l'objet afc:

# poids des lignes et des colonnes (cf cours: marginales X et Y)
afc$lw
afc$cw
## interpréter cf table

## coordonnées: par exemple
afc$l1
afc$c1

# stockage des cos2 cumulés dans le plan principal (colonne 2)
cos2li <- afcin$row.cum[,2] # cos2 cumulés 1-2 pour les lignes
cos2co <- afcin$col.cum[,2] # idem colonnes

# adapter taille des points si besoin
cos2li <- cos2li/90
cos2co <- cos2co/90

par(mfrow=c(1,1))
## $li et $c1:
## echelle dépend des 2 nuages: assembler les 2 nuages avant plot()
mt <- rbind(as.matrix(afc$li), as.matrix(afc$c1))
plot(mt[,1], mt[,2], type="n", xlab="Axe 1", ylab="Axe 2")
text(afc$li$Axis1, afc$li$Axis2, row.names(afc$li), cex=cos2li)
text(afc$c1$CS1, afc$c1$CS2, row.names(afc$c1), col=2, cex=cos2co)
title("li et c1: ACP col - barycentres lignes") # Yeux sont à
l'intérieur des nuages => ACP
abline(h=0, col=8); abline(v=0, col=8)

## $l1 et $co:
mt <- rbind(as.matrix(afc$l1), as.matrix(afc$co))
plot(mt[,1], mt[,2], type="n", xlab="Axe 1", ylab="Axe 2")
text(afc$l1$RS1, afc$l1$RS2, row.names(afc$l1), cex=cos2li)
text(afc$co$Comp1, afc$co$Comp2,row.names(afc$co), col=2, cex=cos2co)
title("l1 et co: ACP lignes - barycentres col")
abline(h=0, col=8); abline(v=0, col=8)


## $l1 et $c1:
```

```r
mt <- rbind(as.matrix(afc$l1), as.matrix(afc$c1))
plot(mt[,1], mt[,2], type="n", xlab="Axe 1", ylab="Axe 2")
text(afc$l1$RS1, afc$l1$RS2, row.names(afc$l1), cex=cos2li)
text(afc$c1$CS1, afc$c1$CS2,row.names(afc$c1), col=2, cex=cos2co)
title("l1 et c1: ACP lignes - ACP col") # simultanée ACP col OK
abline(h=0, col=8); abline(v=0, col=8)
```