**Analyse de Données**

Équipe n°3

Membres de l’équipe : LAPOSTOLET Arsène, LACAZE Thomas, REMEUR Jean-Michel, KERROUE Sébastien

Pour information le rendu est composé de différents fichiers :

1. Ce rapport regroupant les deux études menées
2. Un script .Rmd : script. R de l’étude 1 avec des commentaires (un fichier .R est aussi données au cas où, vous ne disposez pas des extensions nécessaires).
3. Un script .Rmd : script. R de l’étude 2 avec des commentaires (un fichier .R est aussi données au cas où, vous ne disposez pas des extensions nécessaires).
4. Les différents fichier csv utilisés pour les études.
5. Notre script R permettant de traiter les données afin qu’elles soient lisible pour notre script de référence
6. Notre script R de référence permettant de procéder à une ACP sur n’importe quelles données lisibles : une matrice CSV avec labels pour les lignes et les colonnes (UTF-8 pris en charge), nous avons testé avec les bilans financier des groupes.

Table des matières

[Introduction 3](#_Toc1)

[Script R 3](#_Toc2)

[Étude 1 - évolution de l'emploi selon la région de France de 1989 à 2018 3](#_Toc3)

[Lecture des données 3](#_Toc4)

[Nombre de colonnes 3](#_Toc5)

[Nombre de lignes 3](#_Toc6)

[Affichage des 10 premières lignes (pour uniquement 2 colonnes) 3](#_Toc7)

[Informations basiques 4](#_Toc8)

[Résumé (pour uniquement 2 colonnes) 4](#_Toc9)

[Covariance (pour uniquement 2 colonnes) 4](#_Toc10)

[Variance (pour uniquement 2 colonnes) 4](#_Toc11)

[corrélation (pour uniquement 2 colonnes) 4](#_Toc12)

[Données centrées réduites 4](#_Toc13)

[Covariance (pour uniquement 2 colonnes) 5](#_Toc14)

[Variance (pour uniquement 2 colonnes) 5](#_Toc15)

[Corrélation (pour uniquement 2 colonnes) 5](#_Toc16)

[Analyse en composante principale 5](#_Toc17)

[Valeurs propres 5](#_Toc18)

[Graphique des valeurs propres (éboulis et coude) 5](#_Toc19)

[Composantes principales 6](#_Toc20)

[Cercle de corrélation 6](#_Toc21)

[Graphe 2D 7](#_Toc22)

[Étude 2 - Comparaison de l’évolution du nombre de salariés par sexe entre 1989 et 2018 7](#_Toc23)

[Introduction 7](#_Toc24)

[Lecture des données 7](#_Toc25)

[Nombre de colonnes 8](#_Toc26)

[Nombre de lignes 8](#_Toc27)

[Affichage des 10 premières lignes (pour uniquement 2 colonnes) 8](#_Toc28)

[Informations basiques 8](#_Toc29)

[Résumé (pour uniquement 2 colonnes) 8](#_Toc30)

[Covariance (pour uniquement 2 colonnes) 8](#_Toc31)

[Variance (pour uniquement 2 colonnes) 8](#_Toc32)

[corrélation (pour uniquement 2 colonnes) 8](#_Toc33)

[Données centrées réduites 9](#_Toc34)

[Covariance (pour uniquement 2 colonnes) 9](#_Toc35)

[Variance (pour uniquement 2 colonnes) 9](#_Toc36)

[corrélation (pour uniquement 2 colonnes) 9](#_Toc37)

[Analyse en composante principale 10](#_Toc38)

[Valeurs propres 10](#_Toc39)

[Graphique des valeurs propres (éboulis et coude) 10](#_Toc40)

[Composantes principales 10](#_Toc41)

[Cercle de corrélation 10](#_Toc42)

[Graphe 2D 11](#_Toc43)

# Introduction

*Étude 1 : évolution de l'emploi selon la région de France de 1989 à 2018*

*Étude 2 : évolution de l'emploi selon le sexe en France de 1989 à 2018*

Afin de facilement générer des études, nous avons mis en place un script permettant de mettre en forme les données fournies par le projet afin qu’elles soient lisibles par notre script de référence.

Ainsi les données entrées sont celles fournies dans le sujet du projet, et deux fonctions sont présentes pour générer un data frame par région ou par département.

Pour générer nos deux études nous avons utilisé deux scripts (.rmd) qui permettent d’exécuter des commandes R et d’y ajouter des commentaires en markdown: <https://rmarkdown.rstudio.com/authoring_quick_tour.html>

Tout en affichant les résultats graphiques (utiles pour les différents graphiques de l’ACP)

Ainsi le fichier final, est uniquement le regroupement de :

* ‘3—Rapport—Regions.rmd’
* ‘3—Rapport—Femme—Homme.rmd’

# Script R

Présentation / description du script

# Étude 1 - évolution de l'emploi selon la région de France de 1989 à 2018

## Lecture des données

x\_matrix <- read.csv("CSV/generated/reg-e-t-format.csv", header = T, sep = ";", row.names = 1)

### Nombre de colonnes

ncol(x\_matrix)

## [1] 14

### Nombre de lignes

nrow(x\_matrix)

## [1] 30

### Affichage des 10 premières lignes (pour uniquement 2 colonnes)

x\_matrix[1:10,1:2]

## DOM Auvergne.Rhone.Alpes  
## X2018.p. 550802 3022364  
## X2017.p. 546379 2998833  
## X2016 541539 2954646  
## X2015 541416 2928783  
## X2014 532529 2915434  
## X2013 524075 2911376  
## X2012 519294 2888870  
## X2011 520442 2885179  
## X2010 514807 2876699  
## X2009 507328 2850812

## Informations basiques

### Résumé (pour uniquement 2 colonnes)

summary(x\_matrix[,1:2])

## DOM Auvergne.Rhone.Alpes  
## Min. :335232 Min. :2388052   
## 1st Qu.:398155 1st Qu.:2501314   
## Median :467815 Median :2806099   
## Mean :458028 Mean :2729311   
## 3rd Qu.:518172 3rd Qu.:2887947   
## Max. :550802 Max. :3022364

### Covariance (pour uniquement 2 colonnes)

cov(x\_matrix[,1:2])

## DOM Auvergne.Rhone.Alpes  
## DOM 4868469241 14502233502  
## Auvergne.Rhone.Alpes 14502233502 44666371457

### Variance (pour uniquement 2 colonnes)

var(x\_matrix[,1:2]);

## DOM Auvergne.Rhone.Alpes  
## DOM 4868469241 14502233502  
## Auvergne.Rhone.Alpes 14502233502 44666371457

### corrélation (pour uniquement 2 colonnes)

cor(x\_matrix[,1:2])

## DOM Auvergne.Rhone.Alpes  
## DOM 1.0000000 0.9834411  
## Auvergne.Rhone.Alpes 0.9834411 1.0000000

## Données centrées réduites

centree\_reduite <- scale(x\_matrix, center = T, scale = T);

summary(centree\_reduite[,1:2])

## DOM Auvergne.Rhone.Alpes  
## Min. :-1.7599 Min. :-1.6147   
## 1st Qu.:-0.8581 1st Qu.:-1.0788   
## Median : 0.1403 Median : 0.3633   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.8620 3rd Qu.: 0.7506   
## Max. : 1.3296 Max. : 1.3866

### Covariance (pour uniquement 2 colonnes)

cov(centree\_reduite[,1:2])

## DOM Auvergne.Rhone.Alpes  
## DOM 1.0000000 0.9834411  
## Auvergne.Rhone.Alpes 0.9834411 1.0000000

### Variance (pour uniquement 2 colonnes)

var(centree\_reduite[,1:2]);

## DOM Auvergne.Rhone.Alpes  
## DOM 1.0000000 0.9834411  
## Auvergne.Rhone.Alpes 0.9834411 1.0000000

### Corrélation (pour uniquement 2 colonnes)

cor(centree\_reduite[,1:2])

## DOM Auvergne.Rhone.Alpes  
## DOM 1.0000000 0.9834411  
## Auvergne.Rhone.Alpes 0.9834411 1.0000000

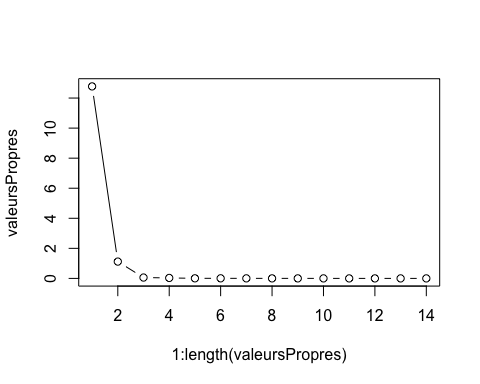
## Analyse en composante principale

### Valeurs propres

propres <- eigen(cor(centree\_reduite));  
valeursPropres <- propres$values;  
vecteursPropres <- propres$vectors;

### Graphique des valeurs propres (éboulis et coude)

plot(1:length(valeursPropres), valeursPropres, type = "b");



### Composantes principales

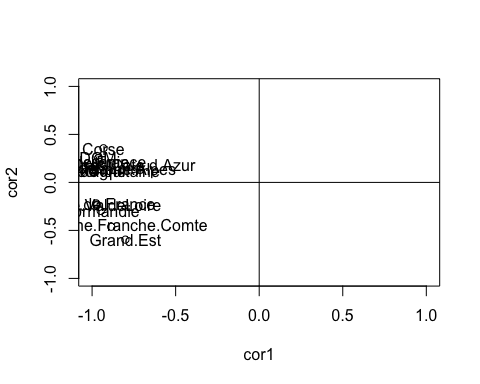
data\_acp <- centree\_reduite %\*% vecteursPropres;  
composante\_principale\_1 <- data\_acp[, 1];  
composante\_principale\_2 <- data\_acp[, 2];

### Cercle de corrélation

Calcule de la corrélation entre chaque variable et les composantes principales

cor1 <- cor(composante\_principale\_1,centree\_reduite);  
cor2 <- cor(composante\_principale\_2,centree\_reduite);  
plot(cor1, cor2, xlim = c(-1, +1), ylim = c(-1, +1))  
abline(h = 0, v = 0)  
text(cor1, cor2, labels = colnames(x\_matrix))

Dans un premier temps on peut dire que la plupart des variables sont proche du cercle, elles sont ainsi bien représentées par l’ACP. On remarque que toutes les variables sont fortement corrélées négativement avec la composante principale 1.



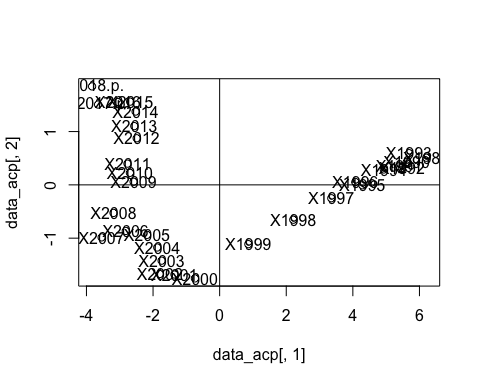
Etant donné que l’on a qu’un seul groupe de variable pour la composante principale 1 on peut l’interpréter comme le nombre de personne employées.

De plus, on a deux groupes de variables : - Le premier (*Occitanie*, *Corse*, *DOM*, *Cote d’Azure*, *Provence*, *Auvergne - Rhones Alpes*, *Nouvelle Aquitaine*, *Ile de France*, *Bretagne*, *Normandie*) positivement correlé à la composante principale 2. - Le second (*Haut de France*, *Bourgogne - Franche Comté*, *Grand-Est*, *Centre Val de Loire*) négativement corrélé avec la composante principale 2.

On peut donc interpréter la composante principale 2 comme **la population de la région**

### Graphe 2D

plot(data\_acp[, 1], data\_acp[, 2])  
text(data\_acp[, 1], data\_acp[, 2], labels = rownames(data\_acp))  
abline(h = 0, v = 0)



L’observation de ce graphique du nuage des individus, nous permet de déterminer trois groupes d’années :

* Les années 1990 : Ce groupe nous permet d’observer un grand nombre d’emploi en France et légèrement plus dans les régions peuplées.
* Les années 2000 : Ce groupe nous permet d’observer un faible nombre d’emploi en France et encore moins dans les regions peuplées.
* Les années 2010 : Ce groupe nous permet d’observer un faible nombre d’emploi en France mais plus d’emplois dans les régions peuplées.

# Étude 2 - Comparaison de l’évolution du nombre de salariés par sexe entre 1989 et 2018

## Introduction

Cette analyse en composante principale se penche sur l’employabilité des femmes et des hommes en France métropolitaine.

## Lecture des données

x\_matrix <- read.csv("CSV/generated/f&h-t-format.csv", header = T, sep = ";", row.names = 1)

### Nombre de colonnes

ncol(x\_matrix)

## [1] 2

### Nombre de lignes

nrow(x\_matrix)

## [1] 30

### Affichage des 10 premières lignes (pour uniquement 2 colonnes)

x\_matrix[1:10,1:2]

## hommes femmes  
## X2018.p. 12478958 12130012  
## X2017.p. 12376710 12082791  
## X2016 12184011 11956942  
## X2015 12059880 11887246  
## X2014 12018119 11835683  
## X2013 12053882 11796718  
## X2012 12044639 11732477  
## X2011 12095423 11750655  
## X2010 12071802 11713497  
## X2009 12042035 11699139

## Informations basiques

### Résumé (pour uniquement 2 colonnes)

summary(x\_matrix[,1:2])

## hommes femmes   
## Min. :10873025 Min. : 8825899   
## 1st Qu.:11266319 1st Qu.: 9703512   
## Median :12056881 Median :11054322   
## Mean :11855366 Mean :10779178   
## 3rd Qu.:12245559 3rd Qu.:11734159   
## Max. :12478958 Max. :12130012

### Covariance (pour uniquement 2 colonnes)

cov(x\_matrix[,1:2])

## hommes femmes  
## hommes 273807190415 5.075695e+11  
## femmes 507569518761 1.211118e+12

### Variance (pour uniquement 2 colonnes)

var(x\_matrix[,1:2]);

## hommes femmes  
## hommes 273807190415 5.075695e+11  
## femmes 507569518761 1.211118e+12

### corrélation (pour uniquement 2 colonnes)

cor(x\_matrix[,1:2])

## hommes femmes  
## hommes 1.000000 0.881414  
## femmes 0.881414 1.000000

## Données centrées réduites

centree\_reduite <- scale(x\_matrix, center = T, scale = T);

summary(centree\_reduite[,1:2])

## hommes femmes   
## Min. :-1.8773 Min. :-1.7749   
## 1st Qu.:-1.1257 1st Qu.:-0.9774   
## Median : 0.3851 Median : 0.2500   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.7457 3rd Qu.: 0.8678   
## Max. : 1.1917 Max. : 1.2275

### Covariance (pour uniquement 2 colonnes)

cov(centree\_reduite[,1:2])

## hommes femmes  
## hommes 1.000000 0.881414  
## femmes 0.881414 1.000000

### Variance (pour uniquement 2 colonnes)

var(centree\_reduite[,1:2]);

## hommes femmes  
## hommes 1.000000 0.881414  
## femmes 0.881414 1.000000

### corrélation (pour uniquement 2 colonnes)

cor(centree\_reduite[,1:2])

## hommes femmes  
## hommes 1.000000 0.881414  
## femmes 0.881414 1.000000

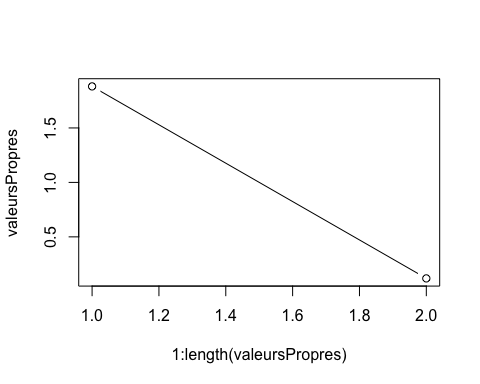
## Analyse en composante principale

### Valeurs propres

propres <- eigen(cor(centree\_reduite));  
valeursPropres <- propres$values;  
vecteursPropres <- propres$vectors;

### Graphique des valeurs propres (éboulis et coude)

plot(1:length(valeursPropres), valeursPropres, type = "b");



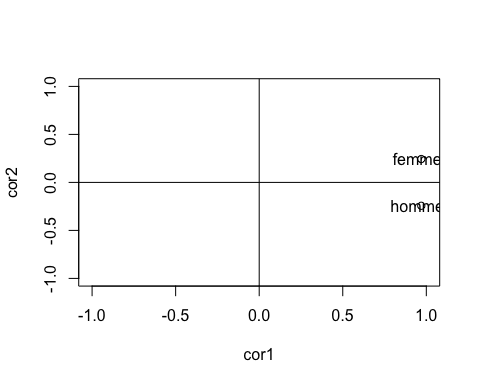
### Composantes principales

data\_acp <- centree\_reduite %\*% vecteursPropres;  
composante\_principale\_1 <- data\_acp[, 1];  
composante\_principale\_2 <- data\_acp[, 2];

### Cercle de corrélation

Calcule de la corrélation entre chaque variable et les composantes principales

cor1 <- cor(composante\_principale\_1,centree\_reduite);  
cor2 <- cor(composante\_principale\_2,centree\_reduite);  
plot(cor1, cor2, xlim = c(-1, +1), ylim = c(-1, +1))  
abline(h = 0, v = 0)  
text(cor1, cor2, labels = colnames(x\_matrix))



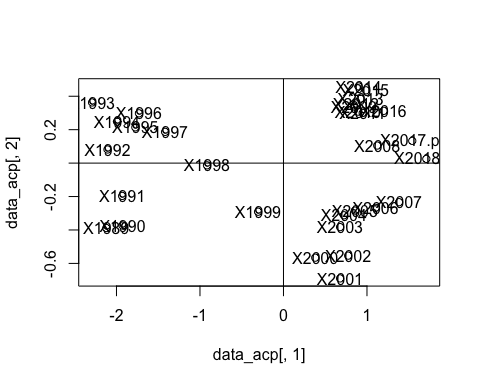
Dans un premier temps on peut dire que la plupart des variables étant proche du cercle, elles sont bien représentées par l’ACP. On observe que nos deux variables sont fortement corrélées à notre première composante principale. Cependant, on peut également constater que la variable *homme* est négativement corrélé à la seconde composante princpale, et que la variable *femme* est quant à elle légerement corrélée à cette dernière.

Etant donné que la quantité d’information portée par la première composante principale est bien supérieure à celle portée par la seconde, on peut dire que la première composante principale peut être analysée comme **La quantité de personnes employés cette année** De plus, dans une moindre mesure, la seconde composante principale peut être analysée comme **La quantité de femmes employées cette année**.

Comme on pouvait s’y attendre, étant donnée la faible quantité de variables dans cette analyse, cela ne nous a pas permis de synthétiser des variables.

### Graphe 2D

plot(data\_acp[, 1], data\_acp[, 2])  
text(data\_acp[, 1], data\_acp[, 2], labels = rownames(data\_acp))  
abline(h = 0, v = 0)



L’observation de ce graphique du nuage des individus, nous permet de déterminer quatres groupes d’années : - Le groupe 1 : Les années 1992, 1993, 1994, 1995, 1996, 1997

Ce groupe présente des valeurs relativement élevées dans la composante principale 2 et des valeurs plus faibles dans la composante principale 1. On peut donc dire que ce sont les années où ont été employés moins de personnes, et un peu plus de femmes.

* Le groupe 2 : Les années 1989, 1990, 1991

Ce groupe présente des valeurs relativement faibles dans les deux composantes principales. On peut donc dire que durant ces années peu de personnes ont été employées et également peu de femmes.

* Le groupe 3 : Les années 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007

Ce groupe présente des valeurs très faibles dans la composante principale 2 et des valeurs relativement élevées dans la composante principales 1. On peut donc dire que ces années sont celles ou ont été employés le moins de femmes mais beaucoup de personnes.

* Le groupe 4 : Les années 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018

Ce groupe présente des valeurs très élevées dans les deux composantes principales. On peut donc dire que ce sont les années où ont été employés le plus grand nombre de personnes, et dans une moindre mesure les années où ont été employées le plus de femmes.