3-Rapport-Regions

# Lecture des données

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

## Nombre de colonne

ncol(x\_matrix)

## [1] 14

## Nombre de ligne

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

## Correlation (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

## Correlation (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

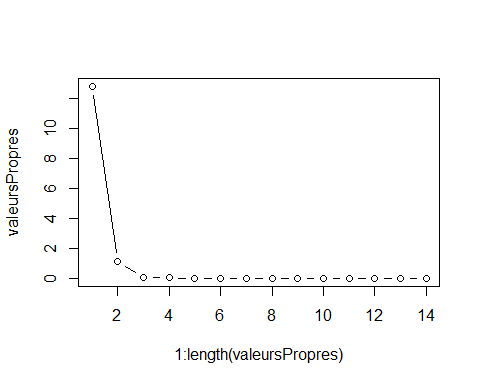
# Analyse en composante principale

## 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];  
totalInfo <- sum(valeursPropres, na.rm = FALSE);  
qte <- (valeursPropres[1] + valeursPropres[2]) / totalInfo;  
message("Quantité d'information avec deux composantes : ", toString(qte \* 100),"%");

## Quantité d'information avec deux composantes : 99.2425347937988%

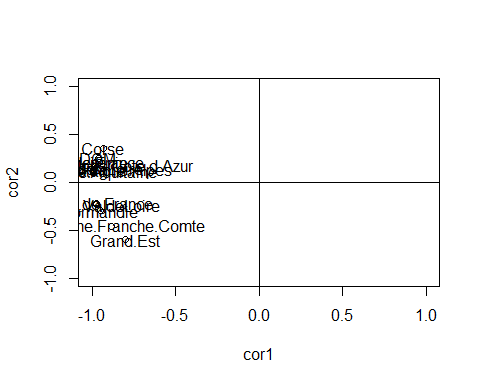
troisComposantes <- FALSE;  
if (qte < 0.8) {  
 composante\_principale\_3 <- data\_acp[, 3]  
 qte <- qte + valeursPropres[3];  
 message("Ajout d'une troisième composante pour améliorer la quantité d'information : ", toString(qte))  
 troisComposantes <- TRUE;  
} else {  
 message("On ne séléctionne que les deux première composantes principales car elles contiennent à elles seules plus de 80% des informations");  
 troisComposantes <- FALSE;  
}

## On ne séléctionne que les deux première composantes principales car elles contiennent à elles seules plus de 80% des informations

## Cercle de correlation

Calcule de la correlation entre chaque variable et les composantes principales

cor1 <- cor(composante\_principale\_1,centree\_reduite);  
cor2 <- cor(composante\_principale\_2,centree\_reduite);  
  
# Correlation 1 - 2  
plot(cor1, cor2, xlim = c(-1, +1), ylim = c(-1, +1))  
abline(h = 0, v = 0)  
text(cor1, cor2, labels = colnames(x\_matrix))



if(troisComposantes){  
 cor3 <- cor(composante\_principale\_3, centree\_reduite);  
 # Correlation 3 - 1  
 plot(cor1, cor2, xlim = c(-1, +1), ylim = c(-1, +1))  
 abline(h = 0, v = 0)  
 text(cor1, cor2, labels = colnames(x\_matrix))  
  
 # Correlation 3 - 2  
 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 variable sont proches du cercle et ainsi bien représentées par l’ACP. On remarque que toutes la variables sont fortement corrélées négativement avec la composante principale 1.

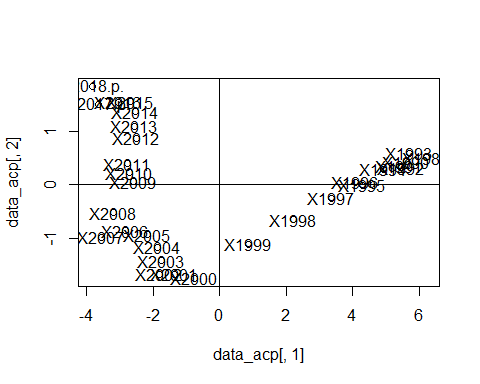
Étant donné que l’on a qu’un seul groupe de variable pour la composante principale 1 on peut l’interpréter comme le nombre de personnes employées.

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

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



if(troisComposantes){  
 # Graphe 3 - 1  
 plot(data\_acp[, 1], data\_acp[, 3])  
 text(data\_acp[, 1], data\_acp[, 3], labels = rownames(data\_acp))  
 abline(h = 0, v = 0)  
  
 # Graphe 3 - 2  
 plot(data\_acp[, 3], data\_acp[, 2])  
 text(data\_acp[, 3], 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.

Pour aller plus loin, il pourrait être de pertinent de mettre les régions sur un pied d’égalité en terme de population en exrpimant les données en entrée en tant que pourcentage de la population active. Cela pourrait éventuellement permettre d’affiner l’analyse.