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 3 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. Un script .Rmd : script. R de l'étude 3 avec des commentaires (un fichier .R est aussi données au cas où, vous ne disposez pas des extensions nécessaires).
5. Les différents fichier csv utilisés pour les études.
6. Notre script R permettant de traiter les données afin qu'elles soient lisible pour notre script de référence
7. 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 pétroliers.

Table des matières

INTRODUCTION

Étude 1 : Employabilité *des femmes et des hommes en France métropolitaine de 1989 à 2018*
Étude 2 : Employabilité entre les régions en France métropolitaine de 1989 à 2018
Étude 3 : Employabilité selon les secteurs d'activités en France métropolitaine 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-Reference.rmd
- 3---Rapport-Femme-Homme.rmd
- 3---Rapport-Regions.rmd
- 3---Rapport-Secteurs-Activite.rmd

SCRIPT R DE REFERENCE

Lecture des données

```
x_matrix <- read.csv("petrole.csv", header = T, sep = ";", row.names = 1)
```

Nombre de colonne

```
ncol(x_matrix)
```

```
## [1] 8
```

Nombre de ligne

```
nrow(x_matrix)
```

```
## [1] 16
```

Affichage des 10 premières lignes (pour uniquement 2 colonnes)

```
x_matrix[1:10,1:2]
```

```
##        NET  INT
## 1969 17.93 3.96
## 1970 16.21 3.93
## 1971 19.01 3.56
## 1972 18.05 3.33
## 1973 16.56 3.10
## 1974 13.09 2.64
## 1975 13.43 2.42
## 1976  9.83 2.46
## 1977  9.46 2.33
## 1978 10.93 2.95
```

Informations basiques

Résumé (pour uniquement 2 colonnes)

```
summary(x_matrix[,1:2])
```

```
##       NET             INT       
##  Min.   : 9.46   Min.   :2.330  
##  1st Qu.:12.38   1st Qu.:2.715  
##  Median :13.23   Median :3.075  
##  Mean   :13.85   Mean   :3.135  
##  3rd Qu.:16.30   3rd Qu.:3.570  
##  Max.   :19.01   Max.   :3.960
```

Covariance (pour uniquement 2 colonnes)

```
cov(x_matrix[,1:2])
```

```
##          NET     INT
## NET 8.423612 1.05828
## INT 1.058280 0.28244
```

Variance (pour uniquement 2 colonnes)

```
var(x_matrix[,1:2]);
```

```
##          NET     INT
## NET 8.423612 1.05828
## INT 1.058280 0.28244
```

Corrélation (pour uniquement 2 colonnes)

```
cor(x_matrix[,1:2])
```

```
##           NET       INT
## NET 1.0000000 0.6861014
## INT 0.6861014 1.0000000
```

Données centrées réduites

```
centree_reduite <- scale(x_matrix, center = T, scale = T);

summary(centree_reduite[,1:2])
```

```
##       NET                INT
##  Min.   :-1.5139   Min.   :-1.5147
##  1st Qu.:-0.5078   1st Qu.:-0.7903
##  Median :-0.2149   Median :-0.1129
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.8420   3rd Qu.: 0.8185
##  Max.   : 1.7766   Max.   : 1.5524
```

Covariance (pour uniquement 2 colonnes)

```
cov(centree_reduite[,1:2])
```

```
##           NET       INT
## NET 1.0000000 0.6861014
## INT 0.6861014 1.0000000
```

Variance (pour uniquement 2 colonnes)

```
var(centree_reduite[,1:2]);
```

```
##           NET       INT
## NET 1.0000000 0.6861014
## INT 0.6861014 1.0000000
```

Corrélation (pour uniquement 2 colonnes)

```
cor(centree_reduite[,1:2])
```

```
##           NET       INT
## NET 1.0000000 0.6861014
## INT 0.6861014 1.0000000
```

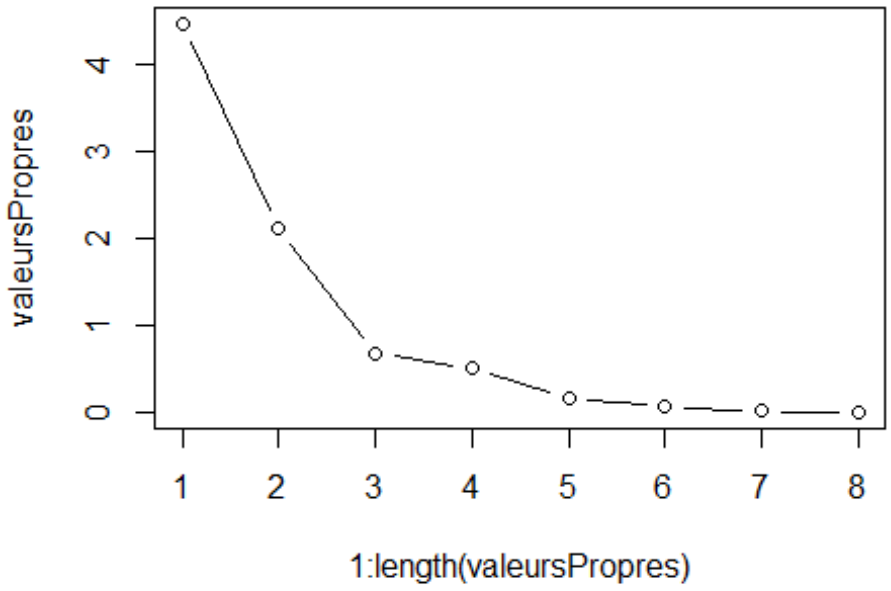
Analyse en composante principale

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

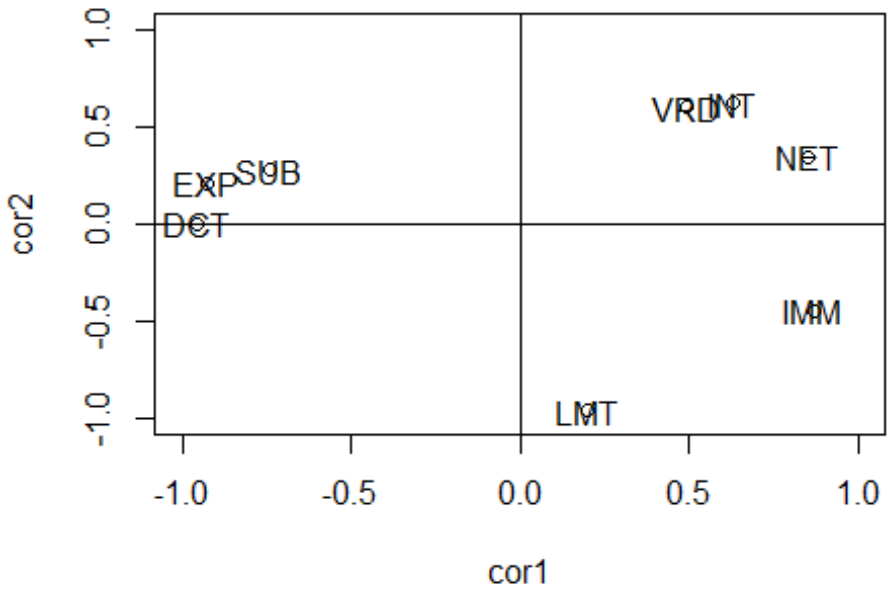
```
data_acp <- centree_reduite %*% vecteursPropres;
composante_principale_1 <- data_acp[, 1];
composante_principale_2 <- data_acp[, 2];
totalInfo <- sum(valeursPropres, na.rm = FALSE);
qte <- (valeursPropres[1] + valeursPropres[2]) / totalInfo;
message("Quantité d'information avec deux composantes : ", toString(qte * 100),"%");

## Quantité d'information avec deux composantes : 82.3152157500261%

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 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);

# Corrélation 1 - 2
plot(cor1, cor2, xlim = c(-1, +1), ylim = c(-1, +1))
```

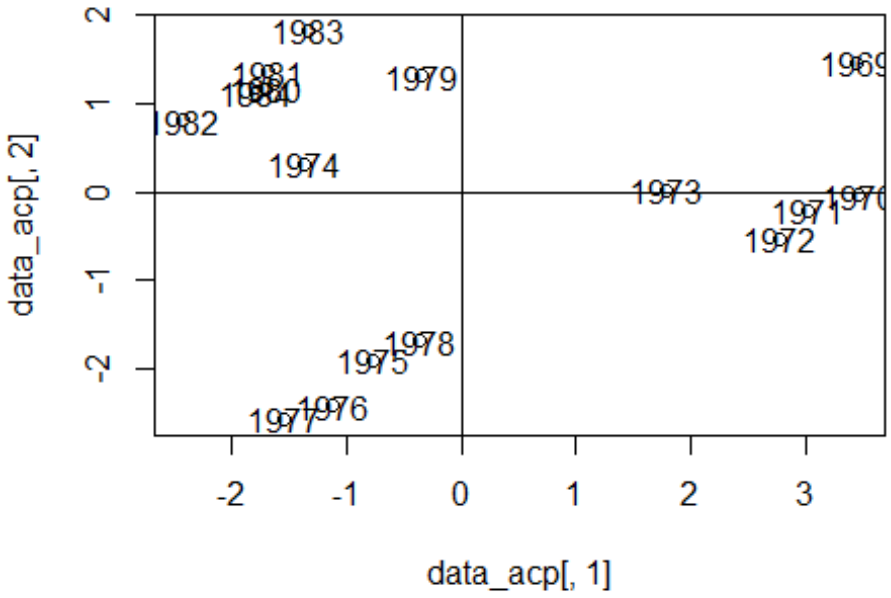
```r
abline(h = 0, v = 0)
text(cor1, cor2, labels = colnames(x_matrix))
```




```r
if(troisComposantes){
  cor3 <- cor(composante_principale_3, centree_reduite);
  # Corrélation 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))

  # Corrélation 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))
}
```

Graphe 2D

```r
# 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)
```




```r
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)

}
```

ÉTUDE 1 - EMPLOYABILITE DES FEMMES ET DES HOMMES EN FRANCE METROPOLITAINE DE 1989 A 2018

Lecture des données

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

Nombre de colonne

```
ncol(x_matrix)

## [1] 2
```

Nombre de ligne

```
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
```

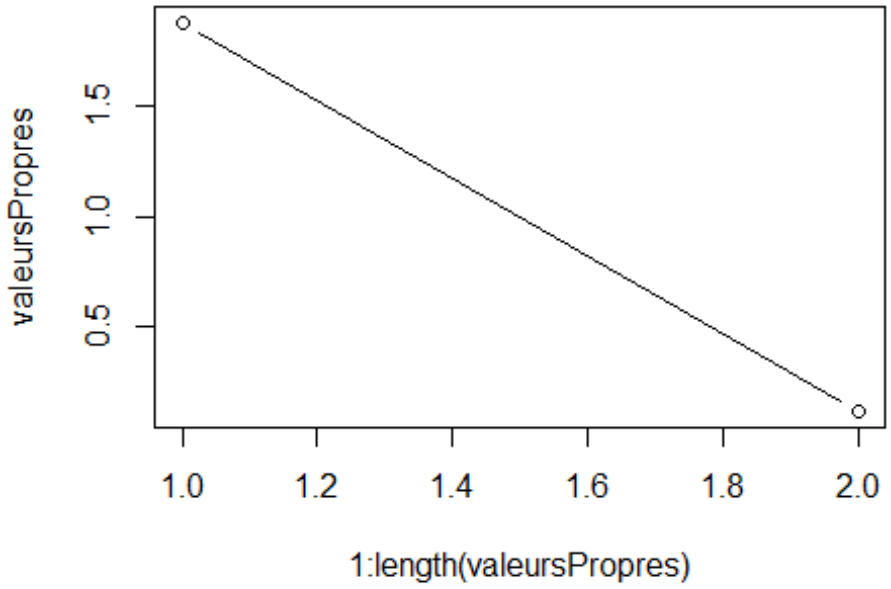
Analyse en composante principale

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

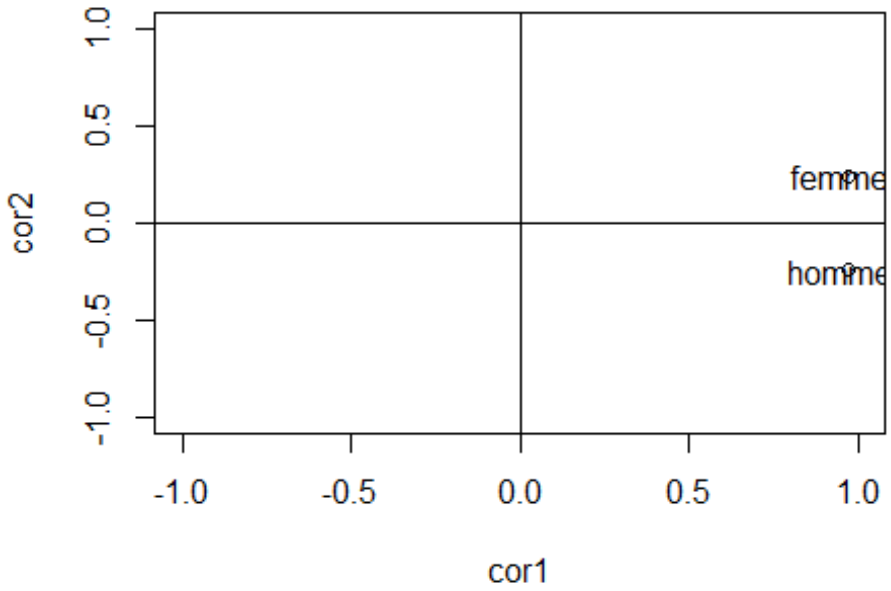
```
data_acp <- centree_reduite %*% vecteursPropres;
composante_principale_1 <- data_acp[, 1];
composante_principale_2 <- data_acp[, 2];
totalInfo <- sum(valeursPropres, na.rm = FALSE);
qte <- (valeursPropres[1] + valeursPropres[2]) / totalInfo;
message("Quantité d'information avec deux composantes : ", toString(qte * 100),"%");

## Quantité d'information avec deux composantes : 100%

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 Corrélation

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);
```

```r
# 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))
```




```r
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))
}
```

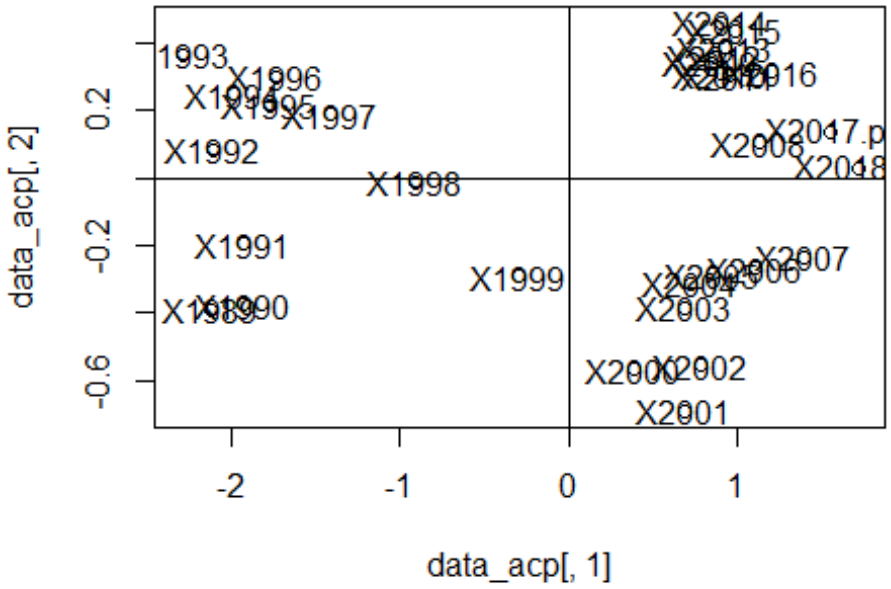
Dans un premier temps on peut dire que la plupart des variable sont proches du cercle et ainsi bien représentées par l'ACP. On observe que nos deux variables sont fortement corrélées à notre première composante principale. Cependant, on peut également constater que la variable *homme* est négativement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

Étant 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, 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

```
# 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 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.

ETUDE 2 - EMPLOYABILITE ENTRE LES REGIONS DE 1989 A 2018

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
```

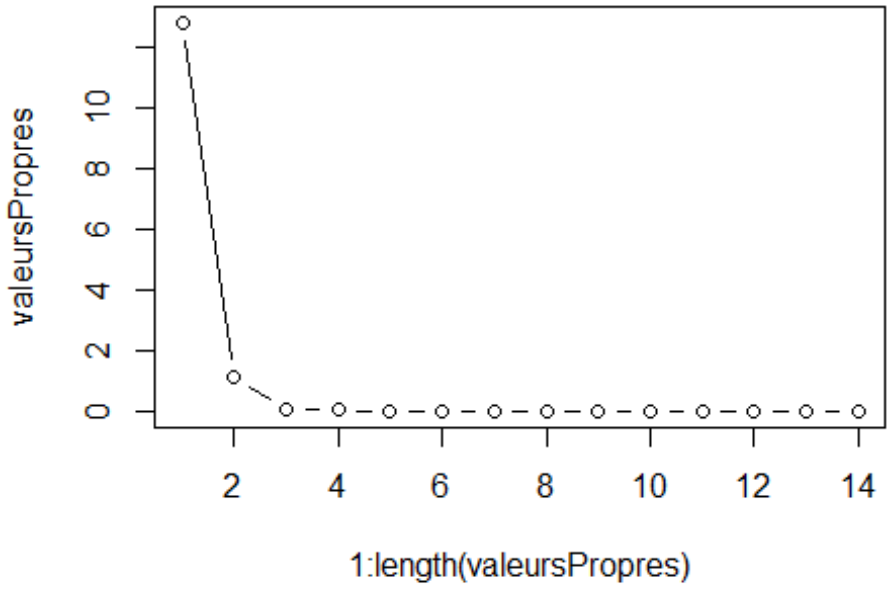
Analyse en composante principale

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

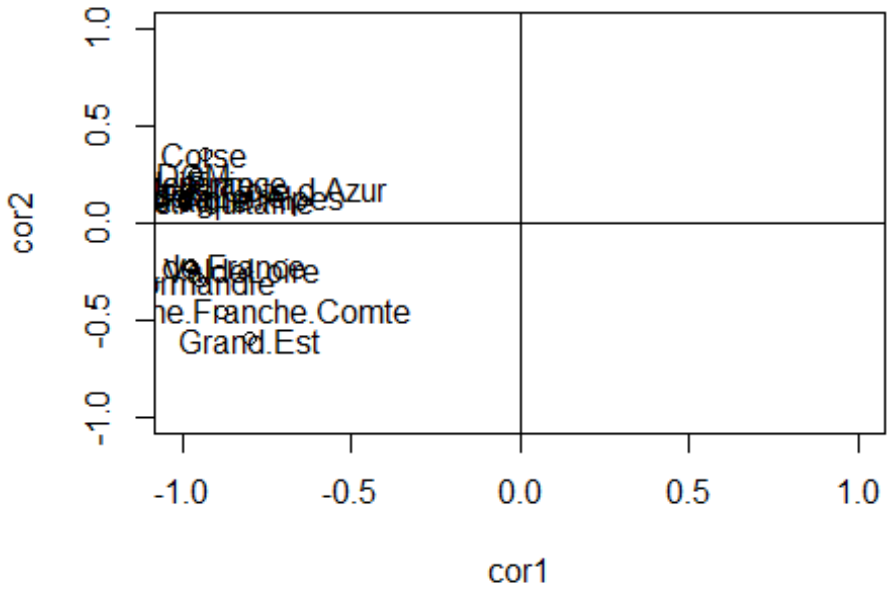
```
data_acp <- centree_reduite %*% vecteursPropres;
composante_principale_1 <- data_acp[, 1];
composante_principale_2 <- data_acp[, 2];
totalInfo <- sum(valeursPropres, na.rm = FALSE);
qte <- (valeursPropres[1] + valeursPropres[2]) / totalInfo;
message("Quantité d'information avec deux composantes : ", 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);
```

```r
# 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))
```




```r
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.

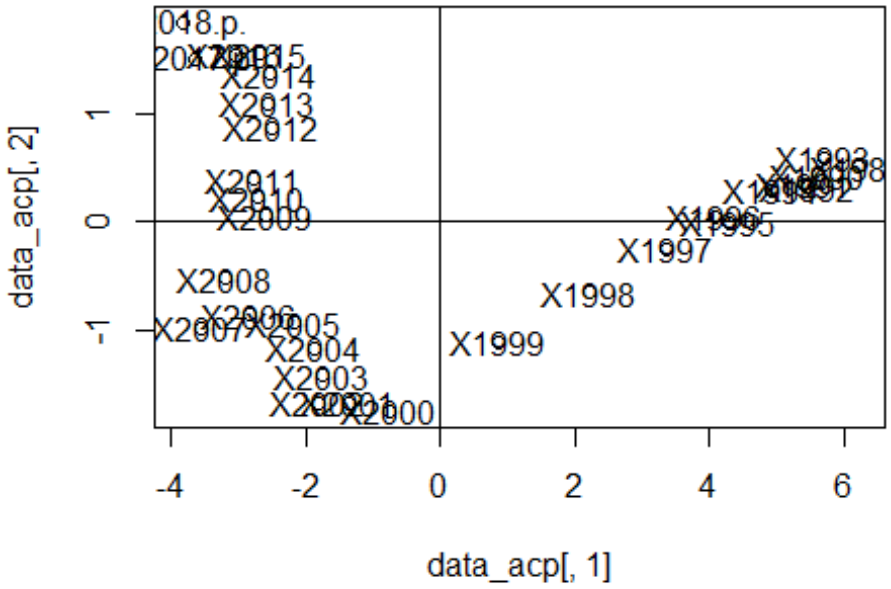
Étant donné que l'on a qu'un seul groupe de variable pour la composante principale 1 on peut l'interpréter comme le nombre de personnes employées.

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 termes de population en exprimant les données en entrée en tant que pourcentage de la population active. Cela pourrait éventuellement permettre d'affiner l'analyse.

ETUDE 3 - EMPLOYABILITE SELON LES SECTEURS D'ACTIVITES DE 1989 A 2018

Lecture des données

```
x_matrix <- read.csv("csv/generated/year-activity-format.csv", header = T, sep = ";",
row.names = 1)
```

Nombre de colonne

```
ncol(x_matrix)

## [1] 5
```

Nombre de ligne

```
nrow(x_matrix)

## [1] 30
```

Affichage des 10 premières lignes (pour uniquement 2 colonnes)

```
x_matrix[1:10,1:2]

##          NA38.TAZ.Agriculture NA38.TBE.Industrie
## X2018.p.               240762            3105002
## X2017.p.               242032            3094187
## X2016                  238996            3095842
## X2015                  238764            3123779
## X2014                  235193            3159226
## X2013                  233705            3193554
## X2012                  229178            3231683
## X2011                  225602            3252758
## X2010                  221317            3270035
## X2009                  224872            3354146
```

Informations basiques

Résumé (pour uniquement 2 colonnes)

```
summary(x_matrix[,1:2])

##  NA38.TAZ.Agriculture NA38.TBE.Industrie
##  Min.   :212438       Min.   :3094187
##  1st Qu.:228630       1st Qu.:3257077
##  Median :238880       Median :3832804
##  Mean   :243188       Mean   :3733275
##  3rd Qu.:258187       3rd Qu.:4048762
##  Max.   :284929       Max.   :4549924
```

Covariance (pour uniquement 2 colonnes)

```
cov(x_matrix[,1:2])

##                      NA38.TAZ.Agriculture NA38.TBE.Industrie
## NA38.TAZ.Agriculture            357604300         1108722275
## NA38.TBE.Industrie             1108722275       209417229113
```

Variance (pour uniquement 2 colonnes)

```
var(x_matrix[,1:2]);
```

```
##                       NA38.TAZ.Agriculture NA38.TBE.Industrie
## NA38.TAZ.Agriculture            357604300         1108722275
## NA38.TBE.Industrie             1108722275       209417229113
```

Corrélation (pour uniquement 2 colonnes)

```
cor(x_matrix[,1:2])
```

```
##                       NA38.TAZ.Agriculture NA38.TBE.Industrie
## NA38.TAZ.Agriculture             1.0000000          0.1281195
## NA38.TBE.Industrie               0.1281195          1.0000000
```

Données centrées réduites

```
centree_reduite <- scale(x_matrix, center = T, scale = T);

summary(centree_reduite[,1:2])
```

```
##  NA38.TAZ.Agriculture NA38.TBE.Industrie
##  Min.   :-1.6261      Min.   :-1.3965
##  1st Qu.:-0.7698      1st Qu.:-1.0406
##  Median :-0.2278      Median : 0.2175
##  Mean   : 0.0000      Mean   : 0.0000
##  3rd Qu.: 0.7932      3rd Qu.: 0.6894
##  Max.   : 2.2073      Max.   : 1.7846
```

Covariance (pour uniquement 2 colonnes)

```
cov(centree_reduite[,1:2])
```

```
##                       NA38.TAZ.Agriculture NA38.TBE.Industrie
## NA38.TAZ.Agriculture             1.0000000          0.1281195
## NA38.TBE.Industrie               0.1281195          1.0000000
```

Variance (pour uniquement 2 colonnes)

```
var(centree_reduite[,1:2]);
```

```
##                       NA38.TAZ.Agriculture NA38.TBE.Industrie
## NA38.TAZ.Agriculture             1.0000000          0.1281195
## NA38.TBE.Industrie               0.1281195          1.0000000
```

Corrélation (pour uniquement 2 colonnes)

```
cor(centree_reduite[,1:2])
```

```
##                       NA38.TAZ.Agriculture NA38.TBE.Industrie
## NA38.TAZ.Agriculture             1.0000000          0.1281195
## NA38.TBE.Industrie               0.1281195          1.0000000
```

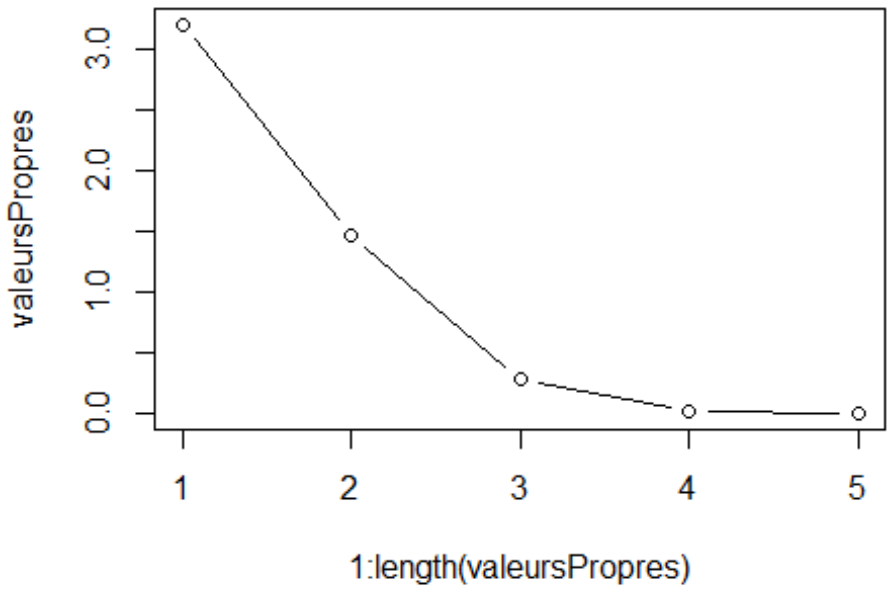
Analyse en composante principale

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

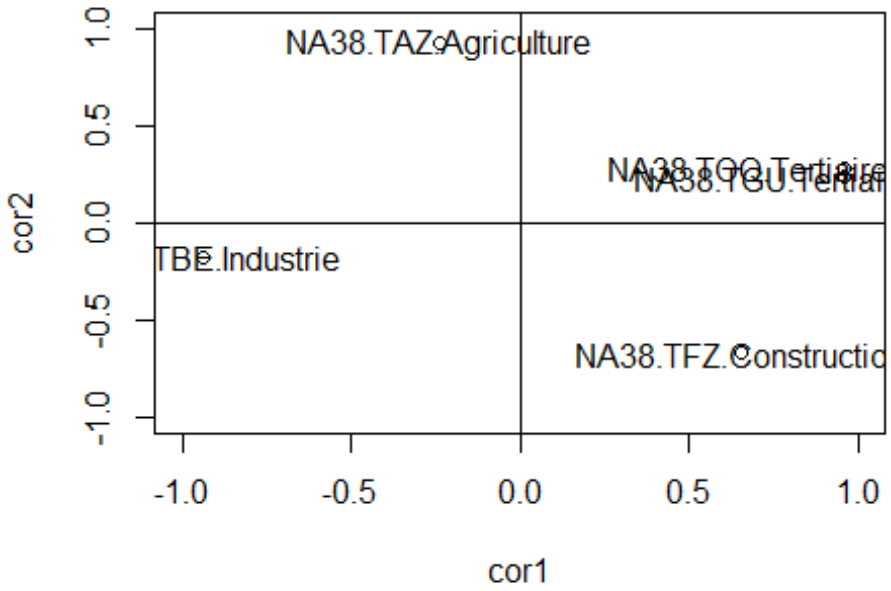
```
data_acp <- centree_reduite %*% vecteursPropres;
composante_principale_1 <- data_acp[, 1];
composante_principale_2 <- data_acp[, 2];
totalInfo <- sum(valeursPropres, na.rm = FALSE);
qte <- (valeursPropres[1] + valeursPropres[2]) / totalInfo;
message("Quantité d'information avec deux composantes : ", toString(qte * 100),"%");

## Quantité d'information avec deux composantes : 93.5145146785711%

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 corrélation

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.

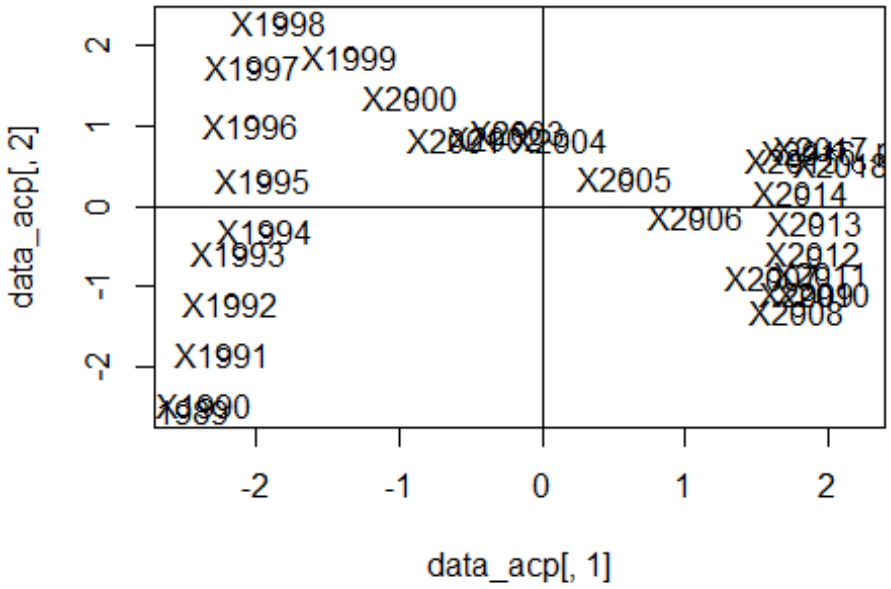
On observe que les secteur de l'industrie sont fortement négativement corrélés avec la composante principale 1 tandis que que les deux secteurs du tertiaire lui sont fortement corrélés. De plus, le secteur de la construction lui est également assez correlé, alors que le secteur de l'agriculture lui est légèrement négativement correlé.

On peut donc interpréter la composante principale 1 comme le fait d'être un secteur qui fournit un service plutôt qu'un secteur qui manufacture des produits.

Malheureusement, nous n'avons pas pu déterminer d'interprétation satisfaisante pour la composante principales 2.

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)

}
```

On observe que plus le temps passe, plus les valeurs de la composante principale 1 sont élevées. On peut donc en conclure que l'emploi dans les secteurs d'activités lié à la production industrielle a baissé entre les années 1989 et 2018 au profit des secteurs de service.