TP1 : Analyse Factorielle des Correspondences

1 ACP et SVD généralisée

1. Charger le jeu de données `USArrests` dans R avec la commande `load`. Afficher les données. Quelle est la classe de cet objet ?

```
data("USArrests")
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
dim(USArrests)
```

```
## [1] 50  4
```

```
class(USArrests)
```

```
## [1] "data.frame"
```

2. Calculer avec les fonctions `princomp` et `prcomp` les composantes principales de l'ACP (les scores des 50 états) et afficher avec la fonction `head` les résultats pour les 5 premiers états.

```
head(princomp(USArrests,cor=TRUE)$scores)
```

```
##            Comp.1 Comp.2  Comp.3   Comp.4
## Alabama    -0.986  1.133 -0.4443  0.15627
## Alaska     -1.950  1.073  2.0400 -0.43858
## Arizona    -1.763 -0.746  0.0548 -0.83465
## Arkansas    0.141  1.120  0.1146 -0.18281
## California -2.524 -1.543  0.5986 -0.34200
## Colorado   -1.515 -0.988  1.0950  0.00146
```

```
head(prcomp(USArrests,scale=TRUE)$x)
```

```
##               PC1    PC2     PC3      PC4
## Alabama    -0.976  1.122 -0.4398  0.15470
## Alaska     -1.931  1.062  2.0195 -0.43418
## Arizona    -1.745 -0.738  0.0542 -0.82626
## Arkansas    0.140  1.109  0.1134 -0.18097
## California -2.499 -1.527  0.5925 -0.33856
## Colorado   -1.499 -0.978  1.0840  0.00145
```

3. La fonction `gsvd` réalise la décomposition en valeur singulières généralisée d'une matrice réelle Z de dimension $n \times p$ avec les métriques diagonales $N = diag(\mathbf{r})$ sur $\mathbb{R}^n$ et $M = diag(\mathbf{c})$ sur $\mathbb{R}^p$. Le code de cette fonction est le suivant :

```r
# fonction SVD generalisee avec metriques diagonales
gsvd <- function(Z,r,c)
{
  #-----entree---------------
  # Z matrice numerique de dimension (n,p) et de rang k
  # r poids de la metrique des lignes N=diag(r)
  # c poids de la metrique des colonnes M=diag(c)
  #-----sortie---------------
  # d vecteur de taille k contenant les valeurs singulieres (racines carres des valeurs propres)
  # U matrice de dimension (n,k) des vecteurs propres de de ZMZ'N
  # V matrice de dimension (p,k) des vecteurs propres de de Z'NZM
  #----------------------------

  k <- qr(Z)$rank
  colnames<-colnames(Z)
  rownames<-rownames(Z)
  Z <- as.matrix(Z)
  Ztilde <- diag(sqrt(r)) %*% Z %*% diag(sqrt(c))
  e <- svd(Ztilde)
  U <-diag(1/sqrt(r))%*%e$u[,1:k] # Attention : ne s'ecrit comme cela que parceque N et M sont diagonales!
  V <-diag(1/sqrt(c))%*%e$v[,1:k]
  d <- e$d[1:k]
  rownames(U) <- rownames
  rownames(V) <- colnames
  if (length(d)>1)
    colnames(U) <- colnames (V) <- paste("dim", 1:k, sep = "")
  return(list(U=U,V=V,d=d))
}
```

(a) Standardiser les données `USArrests` avec la fonction `scale`.

```r
Z <- scale(USArrests)
```

(b) Calculer avec la fonction `gsvd` les composantes principales de l'ACP et afficher avec la fonction `head` les résultats pour les premiers états.

```r
r <- rep(1/nrow(Z),nrow(Z)) #lignes ponderees par 1/n
c <- rep(1,ncol(Z)) #colonnes ponderees par 1
U <- gsvd(Z,r,c)$U
d <- gsvd(Z,r,c)$d
Psi <- U %*% diag(d) #matrice des coordonnees factorielles des lignes
head(Psi)
```

```
##              [,1]   [,2]    [,3]     [,4]
## Alabama    -0.976  1.122 -0.4398  0.15470
## Alaska     -1.931  1.062  2.0195 -0.43418
## Arizona    -1.745 -0.738  0.0542 -0.82626
## Arkansas    0.140  1.109  0.1134 -0.18097
## California -2.499 -1.527  0.5925 -0.33856
## Colorado   -1.499 -0.978  1.0840  0.00145
```

(c) Comparer avec les résultats trouvés avec les fonctions `princomp` et `prcomp`.

(d) Comparer avec les résultats trouvés avec les fonctions `PCA` du package `FactoMineR`.

```r
library(FactoMineR)
?PCA
head(PCA(USArrests,graph=FALSE)$ind$coord)
```

```
##             Dim.1  Dim.2   Dim.3    Dim.4
## Alabama     0.986 -1.133  0.4443  0.15627
## Alaska      1.950 -1.073 -2.0400 -0.43858
## Arizona     1.763  0.746 -0.0548 -0.83465
## Arkansas   -0.141 -1.120 -0.1146 -0.18281
## California  2.524  1.543 -0.5986 -0.34200
## Colorado    1.515  0.988 -1.0950  0.00146
```

2 Données Smoke

Il s'agit d'un tableau de contingence donnant les fréquences de 4 catégories de fumeur (en colonne) pour 5 catégories de salarié (en ligne) dans une entreprise fictive. Les catégories en ligne sont SM=Senior

Managers, JM=Junior Managers, SE=Senior Employees, JE=Junior Employees, SC=Secretaries.

1. Charger le jeu de données Smoke du package package `ca` dans R avec la commande `load`. Afficher les données.

```
library(ca)
data(smoke)
smoke

##    none light medium heavy
## SM    4     2      3     2
## JM    4     3      7     4
## SE   25    10     12     4
## JE   18    24     33    13
## SC   10     6      7     2
```

2. AFC et SVD généralisée.
   (a) Construire la matrice F des fréquences, les vecteurs **r** et r **c** des distributions marginales et la matrice Z des écarts à l'indépendance.

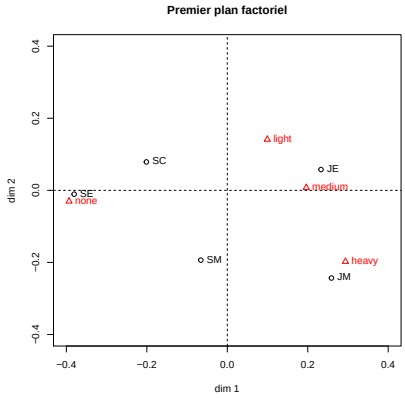
```
F <- smoke/sum(smoke)
r<-apply(F,1,sum)
r

##     SM     JM     SE     JE     SC
## 0.0570 0.0933 0.2642 0.4560 0.1295

c<-apply(F,2,sum)
c

##   none  light medium  heavy
##  0.316  0.233  0.321  0.130

Z <- (F-r%*%t(c))/r%*%t(c)
```

   (b) Calculer avec la fonction `gsvd` les matrices X et Y d des coordonnées factorielles des profil-lignes et colonnes de l'AFC.

```
U<-gsvd(Z,r,c)$U
V<-gsvd(Z,r,c)$V
d<-gsvd(Z,r,c)$d
X <- sweep(U,2,STAT=d,FUN="*") #coordonnees factorielles des profil-lignes
X

##       dim1    dim2     dim3
## SM -0.0658 -0.1937  0.07098
## JM  0.2590 -0.2433 -0.03371
## SE -0.3806 -0.0107 -0.00516
## JE  0.2330  0.0577  0.00331
## SC -0.2011  0.0789 -0.00808

Y <- sweep(V,2,STAT=d,FUN="*") #coordonnees factorielles des profild-colonne
Y

##           dim1     dim2     dim3
## none   -0.3933 -0.03049 -0.00089
## light   0.0995  0.14106  0.02200
## medium  0.1963  0.00736 -0.02566
## heavy   0.2938 -0.19777  0.02621
```

   (c) Représenter avec la fonction `plot` les profil-lignes et les profil-colonnes sur le premier plan factoriel de l'AFC.

```
plot(X[,1:2],xlab="dim 1",ylab="dim 2",xlim=c(-0.4,0.4),ylim=c(-0.4,0.4),main="Premier plan factoriel")
abline(v = 0, lty = 2)
abline(h = 0, lty = 2)
text(X[,1:2],rownames(smoke),pos=4)

points(Y[,1:2],pch=2,col=2)
text(Y[,1:2],colnames(smoke),pos=4,col=2)
```

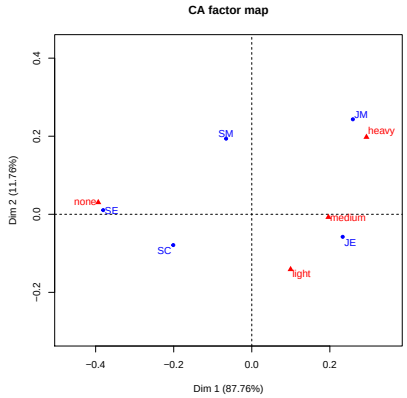


(d) Quel est le pourcentage d'inertie expliquée par le premier plan factoriel de l'AFC

```
T <- sum(d^2) #Inertie totale
d[1:2]^2/T*100 #pourcentage d'inertie des axes

## [1] 87.8 11.8

sum(d[1:2]^2/T)*100 #pourcentage d'inertie du plan

## [1] 99.5
```

3. Retrouver ces résultats avec le package `FactoMineR` et la fonction `CA`.

```
library(FactoMineR)
?CA
res <- CA(smoke,graph=FALSE)
res$eig

##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1   7.48e-02              8.78e+01                             87.8
## dim 2   1.00e-02              1.18e+01                             99.5
## dim 3   4.14e-04              4.85e-01                            100.0
## dim 4   8.15e-33              9.57e-30                            100.0

head(res$row$coord) #matrice X

##      Dim 1   Dim 2    Dim 3
## SM -0.0658  0.1937  0.07098
## JM  0.2590  0.2433 -0.03371
## SE -0.3806  0.0107 -0.00516
## JE  0.2330 -0.0577  0.00331
## SC -0.2011 -0.0789 -0.00808

head(X)

##       dim1    dim2     dim3
## SM -0.0658 -0.1937  0.07098
## JM  0.2590 -0.2433 -0.03371
## SE -0.3806 -0.0107 -0.00516
## JE  0.2330  0.0577  0.00331
## SC -0.2011  0.0789 -0.00808

head(res$col$coord) #matrice Y

##          Dim 1    Dim 2    Dim 3
## none   -0.3933  0.03049 -0.00089
## light   0.0995 -0.14106  0.02200
## medium  0.1963 -0.00736 -0.02566
## heavy   0.2938  0.19777  0.02621

head(Y)

##           dim1     dim2     dim3
## none   -0.3933 -0.03049 -0.00089
## light   0.0995  0.14106  0.02200
## medium  0.1963  0.00736 -0.02566
## heavy   0.2938 -0.19777  0.02621

?plot.CA
plot(res)
```



3 Données textuelles

Il s'agit ici de proposer une méthodologie d'analyse textuelle pour identifier les auteurs de deux fragements de texte anonymes. On connaît pour chacun de ces fragments de texte la fréquence d'apparition de certaines lettres. On suppose également que les auteurs de ces textes appartiennent à la liste suivante d'écrivains du 17ème et 18ème siècle : Charles Darwin, René Descartes, Thomas Hobbes, Mary Shelley et Mark Twain. Ainsi, 3 échantillons de 1000 caractères de textes de ces auteurs ont étés examinés. La fréquence d'apparition de 16 lettres pour chacun de ces 15 échantillons est donnée dans un tableau de contingence.

1. Récupérez les données et charger le jeu de données dans R avec la commande `read.csv`. Afficher les données.

```
library(ca)
data <- read.csv(file="writers.csv", header = TRUE,row.names = 1)
data
```

```
##         B  C  D  F  G  H   I  L  M   N  P  R   S  U  W  Y
## CD1    34 37 44 27 19 39  74 44 27  61 12 65  69 22 14 21
## CD2    18 33 47 24 14 38  66 41 36  72 15 62  63 31 12 18
## CD3    32 43 36 12 21 51  75 33 23  60 24 68  85 18 13 14
## RD1    13 31 55 29 15 62  74 43 28  73  8 59  54 32 19 20
## RD2     8 28 34 24 17 68  75 34 25  70 16 56  72 31 14 11
## RD3     9 34 43 25 18 68  84 25 32  76 14 69  64 27 11 18
## TH1    15 20 28 18 19 65  82 34 29  89 11 47  74 18 22 17
## TH2    18 14 40 25 21 60  70 15 37  80 15 65  68 21 25  9
## TH3    19 18 41 26 29 58  64 18 38  78 15 65  72 20 20 11
## MS1    13 29 49 31 16 61  73 36 29  69 13 63  58 18 20 25
## MS2    17 34 43 29 14 62  64 26 26  71 26 78  64 21 18 12
## MS3    13 22 43 16 11 70  68 46 35  57 30 71  57 19 22 20
## MT1    16 18 56 13 27 67  61 43 20  63 14 43  67 34 41 23
## MT2    15 21 66 21 19 50  62 50 24  68 14 40  58 31 36 26
## MT3    19 17 70 12 28 53  72 39 22  71 11 40  67 20 41 17
## TextX1 24 26 80 17 32 91  86 54 32  91 19 58  93 50 58 30
## TextX2 19 33 35 22 40 96 116 39 40 129 17 72 104 30 25 24
```

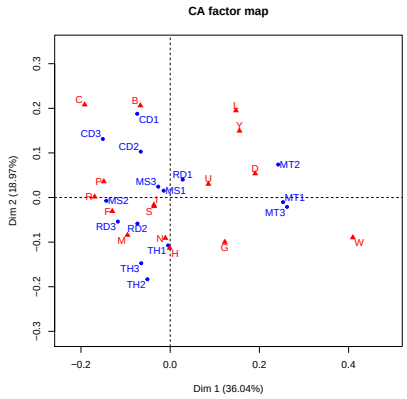
2. On considère dans un premier temps le tableau de contingence des 15 échantillons dont on connaît les auteurs. Effectuer un test du χ^2 d'indépendence pour répondre 'a la question : les distributions des lettres sont-elles significativement différentes d'un échantillon à l'autre ? Vous pouvez utiliser la fonction `chisq.test`

```
K <- data[1:15,]
chisq.test(K)
```

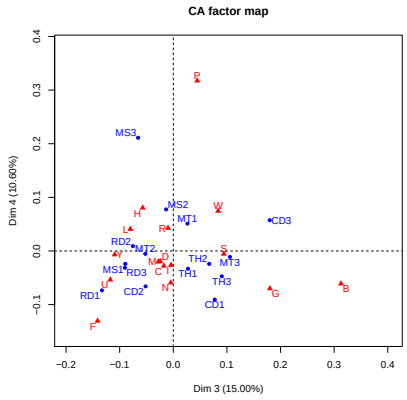
```
##
##  Pearson's Chi-squared test
##
## data:  K
## X-squared = 500, df = 200, p-value <2e-16
```

3. Effectuer une AFC avec la fonction la fonction `CA FactoMineR` et interpréter les résultats.

```
res <- CA(K)
```

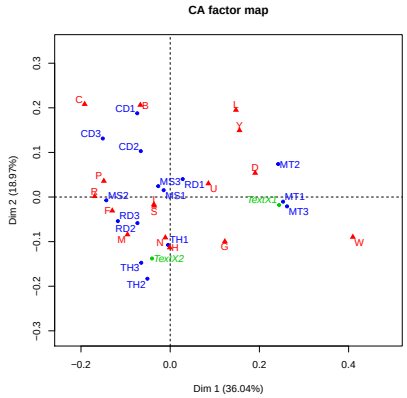



```
res$eig[1:8,]

##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1   0.018228                 36.04                               36.0
## dim 2   0.009594                 18.97                               55.0
## dim 3   0.007585                 15.00                               70.0
## dim 4   0.005363                 10.60                               80.6
## dim 5   0.003577                  7.07                               87.7
## dim 6   0.002111                  4.17                               91.8
## dim 7   0.001592                  3.15                               95.0
## dim 8   0.000917                  1.81                               96.8

plot(res,axes=c(3,4))
```



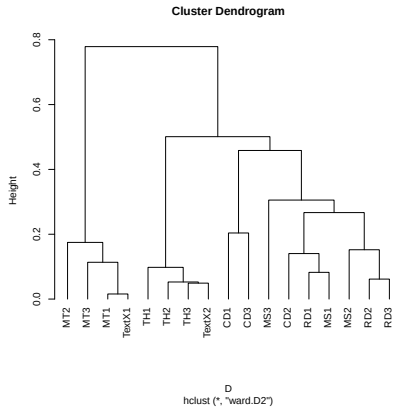
4. Effectuer une AFC avec la fonction `CA FactoMineR` en ajoutant les deux textes inconnus en lignes supplémentaires.

```
res <- CA(data,row.sup=c(16,17),graph=FALSE)
plot(res,col.row.sup=3)
```



5. Faire avec la fonction `hclust` une classification ascendante hiérarchique de Ward des 17 échantillons décrits par leurs coordonnées factorielles sur les 4 premières dimensions de l'AFC. Quelle est la partition en 4 classes ?

```
#matrice des coordonnees factorielles sur 4 dimensions
X <- rbind(res$row$coord[,1:4],res$row.sup$coord[,1:4])
#matrice de distance euclidiennes entre les 17 echantillons
D <- dist(X)
#CAH
tree <- hclust(D,method="ward.D2")
#Dendrogramme
plot(tree,hang=-1)
```




```
#partition en 4 classes
cutree(tree,k=4)

##    CD1    CD2    CD3    RD1    RD2    RD3    TH1    TH2    TH3    MS1
##      1      2      1      2      2      2      3      3      3      2
##    MS2    MS3    MT1    MT2    MT3 TextX1 TextX2
##      2      2      4      4      4      4      3
```