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 ?
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.
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 :

```
# 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`.

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

   (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`.

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.
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.
   (b) Calculer avec la fonction `gsvd` les matrices X et Y d des coordonnées factorielles des profil-lignes et colonnes de l'AFC.
   (c) Représenter avec la fonction `plot` les profil-lignes et les profil-colonnes sur le premier plan factoriel de l'AFC.
   (d) Quel est le pourcentage d'inertie expliquée par le premier plan factoriel de l'AFC
3. Retrouver ces résultats avec le package `FactoMineR` et la fonction `CA`.

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.
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`
3. Effectuer une AFC avec la fonction la fonction `CA FactoMineR` et interpréter les résultats.
4. Effectuer une AFC avec la fonction `CA FactoMineR` en ajoutant les deux textes inconnus en lignes supplémentaires.
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 ?