TP - APPRENTISSAGE NON SUPERVISÉ, ANALYSE EXPLORATOIRE - ACP, AC ACM

Dans ce TP on met en oeuvre les techniques de l'analyse en composantes principales (ACP), de l'analyse des correspondances (AC) et de l'analyse des correspondances multiples (ACM), à l'aide du package `FactoMineR`. On utilisera également un jeu de données du package `ade4`

```
library(ade4)
library(FactoMineR)
```

Exercice 1 (Analyse en composantes principales des températures de villes d'europe)**.** Téléchargez les données `temperature`

```
temperature <-
   read.table("http://factominer.free.fr/book/temperature.csv",
            header=TRUE,sep=";",dec=".",row.names=1)
```

Le jeu de données contient les températures mensuelles moyennes de différentes villes. Les moyennes et amplitudes annuelles sont indiquées en plus, ainsi que les coordonnées géographiques (latitude, longitude). Une variable qualitative indique l'aire géographique (nord, sud, est, ouest). L'objectif ici est de dresser une typologie des villes en utilisant seulement les températures mensuelles et de valider l'analyse ensuite avec les variables supplémentaires. Les commandes suivantes permettent d'inspecter les caractéristiques principales du jeu de données.

```
class(temperature)
names(temperature)
rownames(temperature)
dim(temperature)
plot(as.numeric(temperature[1,1:12]), ylim= range(temperature[,1:12]))
lines(as.numeric(temperature[2,1:12]))
lines(as.numeric(temperature[3,1:12]), col="red")
lines(as.numeric(temperature[4,1:12]), col="blue")
```

La commande suivante effectue l'ACP du jeu de données. les individus 24 à 35 sont passés en individus supplémentaires (pas des capitales). On passe en variables supplémentaires toutes les variables qui ne sont pas des moyennes mensuelles. Par défaut la fonction `PCA` centre et réduit les colonnes (`scale.unit = TRUE`)

```
res <- PCA(temperature,ind.sup=24:35,quanti.sup=13:16,quali.sup=17)
```

Pour fermer toutes les fenêtres graphiques : `graphics.off()` .

Pour passer en revue ce que fait la fonction `PCA` et inspecter le résultat :

```
?PCA
names(res)
```

N.B. pour accéder aux éléments d'une liste on utilise le symbole $: par exemple

```
res$call
```

Pour tracer le nuage des individus (capitales) :

```
plot.PCA(res,choix="ind")
```

La variable supplémentaire `Area` (colonne 17) pourrait permettre d'interpréter les axes. On peut colorier les différents individus en fonction de la valeur de cette variable :

```
plot.PCA(res,choix="ind",habillage=17)
```

1. Comment interpréter les deux premiers axes au vu de ces résultats ? Autrement dit qu'est-ce qui oppose les individus aux extrémités de chacun des axes ?

Le package permet une description 'automatique' des axes à l'aide des variables (classement en fonction de la corrélation)

```
dimdesc(res)
```

2. Quelles sont les variables les plus corrélées à la composante 1 ? à la composante 2 ?

Les valeurs propres de la matrice de variance (valeurs, pourcentage, pourcentage cumulé) sont dans l'élément `eig` de la liste renvoyée par `PCA`

3. Quelle est la part de variance expliquée par les 2 premières composantes ? est-il utile de considérer les composantes suivantes pour ce jeu de données ?

Les résultats numériques (coordonnées des projections, corrélations au carré, contributions) pour les individus actifs, c.a.d. ceux ayant participé à l'analyse, sont stockés dans l'élément `ind` de la liste. On pourra afficher le graphe des individus actifs comme ceci :

```
plot.PCA(res, choix = c("ind"),
       invisible=c("ind.sup", "quali", "quanti.sup"))
```

4. Donnez deux capitales représentatives de l'axe 1, à l'opposé sur l'axe. Considérer les contributions (ou les $\cos^2$) et repérer deux villes ayant des coordonnées de signes opposés.
5. Procédez de même avec les individus supplémentaires (qui ont le rôle de données de validation), c.a.d. les villes n'étant pas des capitales.

```
plot.PCA(res, choix = c("ind"), invisible = c("ind"))
res$ind.sup
```

Pour tracer le cercle des corrélations :

```
plot.PCA(res, choix = "var")
res$var
```

6. Quels sont les mois contribuant le plus à l'inertie sur l'axe 1 ? sur l'axe 2 ?
   N.B. Dans le package on utilise la convention $\|a\| = \sqrt{\lambda}$, les contributions sont donc les carrés des coordonnées, divisés par la valeur propre de l'axe.
7. Les résultats pour les variables quantitatives additionnelles sont stockées dans `res$quanti.sup` À quelle composante principale pouvez-vous rattacher prioritairement chacune des variables supplémentaires ?
8. Concernant les variables qualitatives supplémentaires : chaque catégorie est identifiée au barycentre des individus qui la possèdent.

```
res$quali.sup
plot.PCA(res, choix = "ind", invisible = c("ind", "ind.sup"))
```

   À quelle composante la catégorie 'East' est-elle le plus corrélée ? quel est le signe de cette corrélation ?
9. En se basant uniquement sur les deux premières composantes de l'ACP, peut-on deviner le signe de la corrélation entre 'amplitude' et 'janvier' ?
10. **Conclusion** : dressez une typologie sommaire des températures en Europe.

Exercice 2 (Analyse des correspondances)**.**
On utilise les données `JO` . Les lignes correspondent à des disciplines olympiques, les colonnes à des pays. Chaque cellule contient le nombre de médailles (or/argent/bronze) gagnées par un pays entre 92 et 2008 (5 jeux, donc 15 médailles par discipline). On a bien une table de contingence (creuse) obtenue à partir de données originales binaires où chaque ligne correspond à une médaille, avec deux « 1 » par lignes (pays concerné et discipline), les cellules restantes valant 0. Il y a 24 disciplines. Pour charger la table de contingence et afficher la description des données :

```
data(JO)
?JO
```

On vérifie qu'il y a bien 15 médailles par discipline :

```
apply(JO, 1, sum)
```

La table de contingence `JO` est très particulière : chaque profil ligne a le même poids $1/15$.

La fonction `CA` de `FactoMineR` effectue par défaut un test du χ^2 (pour voir s'il y a des dépendances à étudier). Le résultat est affiché, entre autres choses, avec la commande **`summary`**.

```
resJO <- CA(JO)
summary(resJO)
```

Le test du χ^2 rejette largement l'indépendance mais la p-valeur ne doit pas être prise au pied de la lettre : la condition $n_{ij} > 5$ pour toute paire (i, j) n'est pas satisfaite.

Pour obtenir les profils lignes et colonnes de la table :

```
## profils lignes
rowprof <- JO / apply(JO, 1, sum)
apply(rowprof,1,sum)

## profils colonnes
colprof <- t(t(JO) / apply(JO, 2, sum)) #t() : transposee
apply(colprof,2,sum)
```

1. Inspectez les valeurs propres de l'AC.

```
round(resJO$eig,1)
```

   Combien faudrait-il garder de dimensions pour expliquer 50% de la variance ?

2. Vérifiez que l'inertie totale vaut $1/n\times$ la statistique du χ^2. On accède directement à cette dernière via la fonction `chisq.eig`.
3. Les coordonnées des profils lignes sur les axes principaux sont accessibles via

```
resJO$row$coord
```

   Vérifiez que le barycentre des projections des profils lignes sur les deux premiers axes est le vecteur nul, lorsque l'on utilise comme poids les fréquences marginales, obtenues comme ceci :

```
n <- sum(JO)
rowW <- apply(JO,1,sum)/n
rowW
colW <- apply(JO,2,sum)/n
colW
```

4. Vérifiez que la variance pondérée des coordonnées des lignes sur le premier axe est égale à la première valeur propre.

5. Vérifiez que la corrélation entre les vecteurs des coordonnées des profils lignes projetées sur l'axe 1 et l'axe 2 est nulle. Justifiez théoriquement ce résultat.
6. Calculez les contributions des profils lignes à l'axe 1 en utilisant les poids `rowW`, les coordonnées des lignes `resJO$row$coord` et les valeurs propres `resJO$eig[,1]`. Vérifiez votre résultat avec

```
resJO$row$contrib
```

7. Interprétation des résultats :
   (a) Comment interprétez-vous les axes de l'ACP des lignes en vu du graphe des profils lignes ?

```
plot.CA(resJO, invisible= "col" )
```

   (b) On cherche à dresser des profils de pays en termes de points forts disciplinaires. Rajoutez les pays (c.a.d. tracez la représentation jointe), comme ceci :

```
plot.CA(resJO)
```

   (c) Calculez les contributions des pays à l'axe 1 en utilisant leurs poids, les coordonnées des colonnes et la première valeur propre (c.f. question 6). Vérifiez votre résultat en inspectant

```
resJO$col$contrib[,1]/100
```

   Rangeons les contributions par ordre décroissant :

```
contr <- resJO$col$contrib[,1]
contr[rev(order(contr))]
```

   (d) Quels sont les 5 pays contribuant le plus à l'inertie sur l'axe 1 ? De quel côté se situent les États-Unis par rapport à la direction 'endurance' ?

Exercice 3 (ACM : données bancaires)**.** On charge les données 'banque' du package `ade4`

```
data(banque)
dim(banque)
names(banque)
head(banque)
?banque
```

Ces données résultent d'une enquête auprès de 810 clients d'une banque et décrivent les clients suivant certaines caractéristiques. Nous retenons ici les variables age, sexe, interdit, credhab, credcon (4,5,6,12,13).

1. Représentez les données pour chacune des variables, grâce à la commande **`plot`**.
2. Le tableau disjonctif est donné par la commande `tab.disjonctif`.

```
ids <- c(4,5,6,12,13)
idSup <- setdiff(1:21, ids )
tabdisj <- tab.disjonctif(banque[,ids])
colnames(tabdisj)
head(tabdisj)
```

   (a) Sans calcul, donnez la somme totale du tableau et de chaque ligne.
   (b) Étudiez les effectifs de chaque modalité, et repérez les éventuelles modalités rares.

3. On effectue une ACM avec factoMineR comme ceci :

```
resMCA <- MCA(banque[,ids])
```

   (a) que représentent les 3 graphes ?
   (b) Calculez l'inertie totale, et donnez l'inertie relative de chaque axe. On utilisera `resMCA$eig`
   (c) Retrouvez l'inertie totale à partir de la statistique du χ^2 du tableau disjonctif.
   (d) On considère l'axe 1 : les contributions des modalités de chaque variable à l'axe sont accessibles ainsi :

```
ctrs <- resMCA$var$contrib[,1]
ctrs
```

   Quelles sont les trois catégories contribuant le plus à l'axe 1 ?
   (e) Même question pour l'axe 2.
   (f) **Conclusion :** Dressez une typologie des clients, c.a.d. interprétez les axes à l'aide de la carte des catégories et des variables.
   (g) Refaire l'analyse en passant en argument de MCA toutes les variables et en précisant que les variables autres que celles considérées ci-dessus sont des variables supplémentaires. Cela confirme-t-il l'interprétation des résultats ?

```
resMCA <- MCA(banque, quali.sup = idSup)
graphics.off()
plot.MCA(resMCA, choix = "ind", invisible = "ind", selectMod = "cos2 30",
       unselect = "gray50")
```