TP 2 : Analyse descriptive élémentaire multidimensionnelle et Analyse en Composantes Principales

Ce TP a pour objectif de vous montrer l'exploration élémentaire multidimensionnelle d'un jeu de données et d'approfondir sa compréhension grâce à la méthode d'Analyse en Composante Principale.

Plusieurs Packages `R` fournissent des outils permettant de réaliser une Analyse en Composantes Principales, on en retiendra deux :

- le package `stats` et les fonctions `prcomp`, `princomp`
- le package `FactoMineR` et la fonction `PCA`.

Le jeu de données "Criminalité" sera l'occasion d'utiliser le package `stats` et celui des "hôtels méditerranéens" pour le package `FactoMineR`.

1 Criminalité aux USA

1.1 Exploration élémentaire

Ce jeu de données provient de la documentation SAS, il a le mérite d'être simple à interpréter. Les 50 états américains sont décrits par 7 variables exprimant un taux de criminalité pour chaque type de délit : `Murder, Rape, Robbery, Assault, Burglary, Larceny, Auto Thieft`.

- Récupérer le jeu de données `crime.dat`.
- Charger ces données.

```
crime <- read.table(file.choose(),dec=".",row.names=1)
names(crime) <- c("murder","rape","robbery","assault","burglary",
 "larceny","auto")
```

- Comme dans le TP précédent, étudier la distribution de chacune des variables (histogrammes, boxplots, indices...) et repérer d'éventuelles données atypiques.

```
summary(crime)
hist(crime$murder)
boxplot(crime$murder)
```

- Obtenir les covariances et les corrélations entre les variables quantitatives, repérer d'éventuelles liasons non linéaires...

```
var(crime)  # la commande cov(crime) permet également d'afficher la matrice
            # de variance-covariance
cor(crime)
plot(crime)
```

1.2 ACP

1. Réaliser l'ACP de la table `crime`.

   ```
   crime.acp<-princomp(crime,cor=T)
   ```

   L'option `cor=T` conduit à une ACP réduite alors qu'elle ne l'est pas par défaut.

2. Choisir la dimension à partir des valeurs propres,

3. Tracer les représentations individus vs. variables

   ```
   plot(crime.acp)
   boxplot(as.data.frame(crime.acp$scores))
   biplot(crime.acp)
   biplot(crime.acp,choices=2:3)
   ```

4. Que vaut la variance de chacune des composantes ?

5. Confirmer ou infirmer le choix de dimension précédent.

6. Identifier les valeurs atypiques sur la troisième composante.
   Ces états sont potentiellement influents sur la définition des axes. Ce sont ceux déjà repérés lors de l'étude unidimensionnelle. La variance de chaque composante (si les options de l'ACP n'ont pas été modifiées) est la valeur propre correspondante. Aux trois premières correspondent des tailles de boîtes raisonnables tandis que la 4ème ainsi que les suivantes sont plus petites. Un choix de 2 voire 3 dimensions semble raisonnable mais deux individus atypiques apparaissent sur la 3ème composante. De façon générale, pour s'assurer de la robustesse des résultats, il est bon de vérifier que l'analyse calculée sans individus influents est identique c'est-à-dire qu'elle conduit aux mêmes axes factoriels. Dans le cas contraire, une discussion avec les commanditaires ou spécialistes du domaine concerné s'impose. S'agit-il d'une erreur de mesure, d'échantillonnage ? Faut-il ou non conserver une observation atypique dans les données ou faut-il conserver une composante avec des valeurs atypiques.

7. Identifiez ces points pour les exclure des calculs :

   ```
   crime0<-subset(crime,row.names(crime)!="New_York")
   crime0<-subset(crime0,row.names(crime0)!="Massachusetts")
   ```

   Rééexécutez les instructions précédentes. Le plan (1,2) est-il modifié ? L'axe 3 ?

8. Graphique des variables : interpréter le premier axe et identifier cet effet `taille`. Interpréter le 2ème axe.

2 Hôtels Méditerranéens

2.1 ACP avec une variable qualitative supplémentaire

1. Charger le fichier `hotels.csv` dans `R` .

   ```
   hotels <- read.csv(file.choose(),row.names=1)
   ```

2. Quelles sont les différentes variables ? Quelle est leur nature ? Qui sont les individus sur qui on va faire porter l'analyse en composantes principales ? Obtenir les statistiques descriptives, les covariances et les corrélations entre les variables quantitatives du jeu de données. Repérer d'éventuelles liaisons non linéaires entre les variables.

3. Faire l'analyse en composantes principales du tableau de données, puis construire les diagrammes des valeurs propres suivants. Par combien d'axes l'information est-elle résumée de manière satisfaisante ?

```
library(FactoMineR)
res.pca <- PCA(hotels, quali.sup = 1, scale.unit = TRUE, ncp = 8,
graph = FALSE)
barplot(res.pca$eig$per, ylab = "Inertie expliquée (%)",
xlab = "Composante")
barplot(res.pca$eig$cum, ylab = "Inertie cumulée expliquée (%)",
xlab = "Composante")
abline(h = 80, lty = 2, lwd = 2)
```

4. Représenter les individus dans le premier et le second plan factoriel en étiquetant les données de telle sorte que l'on puisse identifier à quel hôtel est associé chaque point tout en indiquant les pays où sont implantés les hôtels.

```
plot(res.pca, choix = "ind", habillage = 1)
plot(res.pca, choix = "ind", habillage = 1, axes = c(3,4))
```

   Commenter la qualité de la représentation obtenue sur les quatres premiers axes factoriels en analysant le contenu des tableaux suivants :

```
res.pca$var$coord, res.pca$var$cor, res.pca$var$cos2, res.pca$var$contrib
res.pca$ind$coord, res.pca$ind$cos2, res.pca$ind$contrib
```

   Construire les cercles des corrélations des variables avec le premier et le second axe factoriel puis avec le troisième et le quatrième axe factoriel. On obtiendra des graphiques similaires à ceux reproduits ci-dessous.

```
plot(res.pca, choix = "var")
plot(res.pca, choix = "var", axes = c(3, 4))
```

5. La fonction `dimdesc` permet d'obtenir une description automatique des axes de l'ACP. Commenter ses résultats lorsqu'elle est appliquée à `res.pca`.

```
dimdesc(res.pca)
```

6. La fonction `coord.ellipse` permet d'obtenir des régions de confiance pour les modalités d'une variable qualitative. Commenter ses résultats lorsqu'elle est appliquée à res.pca.

```
elldata = cbind.data.frame(hotels[, 1], res.pca$ind$coord)
coordell = coord.ellipse(elldata, bary = TRUE)
plot.PCA(res.pca, habillage = 1, ellipse = coordell, new.plot = F)
plot.PCA(res.pca, habillage = 1, ellipse = coordell, new.plot = F)
```

2.2 ACP avec variables qualitative et quantitative supplémentaires

Comparer les résultats que vous venez d'obtenir avec ceux reproduits ci-dessous. On commencera par chercher la différence existant entre l'analyse qui vient d'être faite et celle qui a été réalisée ci-dessous. On s'intéressera en particulier au rôle joué par la variable `Prix`. En quoi cette seconde manière d'analyser les données est-elle plus intéressante ?

```
res.pca2 <- PCA(hotels, quali.sup = 1, quanti.sup = 8, scale.unit = TRUE,
ncp = 8, graph = FALSE)
```

Reproduire l'analyse précédente sur `res.pca2` et commenter les résultats obtenus.