TP N°3----ANALYSE EN COMPOSANTES PRINCIPALES

Plusieurs fonctions, de différents packages, sont disponibles dans le logiciel R pour le calcul de l'ACP:

- prcomp() et princomp() [fonction de base, package stats],
- PCA() [package FactoMineR],
- dudi.pca() [package ade4],
- et epPCA() [package ExPosition]

Exemple:

Nous allons utiliser le jeu de données Cancer.txt, composé de 7 types de cancers sur lesquels, 10 marqueurs ont été testés.

```
# Lecture des données
> data <- read.table("cancers.txt",sep="\t",header=T)
> colnames(data)
> rownames(data) = data[,1] # On donne à chaque ligne le nom de son cancer (Cancer 1, Cancer 2)
> x <- data[1:7,-1] ;x # On exclue la colonne 1 qui contenait les noms des cancers
```

Explorer ce jeu de données :

- afficher la matrice X,
- utiliser la commande summary(X) pour visualiser quelques informations importantes,
- afficher les boîtes à moustaches associées (commande boxplot),
- que donne la commande summary(t(X)), boxplot(t(x)) ?

```
# Calcul de la matrice de corrélation :
> matCors <- cor(x)
> matCors
> symnum(matCors, abbr.colnames=FALSE)
#les coefficients de corrélation entre 0 et 0.3 sont remplacés par un espace (" “);
#les coefficients de corrélation entre 0.3 et 0.6 sont remplacés par”.“; etc …

> library(corrplot)
> corrplot(matCors, type="upper", order="hclust", tl.col="black", tl.srt=45)

# Pour procéder à l'ACP, nous commençons par la normalisation des données
> Xsc <- scale(x, scale = T)
> Xsc
> boxplot(Xsc)

# Exécuter chacune des lignes suivantes et expliquer ce qu'elles font et ce que l'on obtient :
> Sigma <- t(Xsc) %*% Xsc/nrow(Xsc)
> ACP <- eigen(Sigma)
> ACP$values
> plot(ACP$values, type = "b")
> inertie <- cumsum(ACP$values)/sum(ACP$values)
> inertie
> pourcinertie <- inertie*100
> pourcinertie
```

> plot(ACP$values/sum(ACP$values), type="b")

Utiliser les résultats obtenus pour répondre aux questions suivantes :

- combien d'axes retiendriez-vous pour l'ACP de ce jeu de données ?
- quelle est la part d'inertie expliquée par le plan principal ?
- calculer la matrice CP des composantes principales.
- calculer la qualité de la représentation de chacun des individus et stocker les résultats dans un vecteur cos2.
- calculer et commenter les contributions des individus aux variables principales.

Représentation graphique

Grâce à la matrice CP, nous pouvons représenter les données dans le plan principal. Exécuter les commandes suivantes et commenter les résultats,

> plot(CP[, 1:2], pch = 2, cex = cos2)
> text(CP[, 1:2], labels = rownames(C), pos = 3)

Nous pouvons également obtenir la matrice contenant les corrélations entre les variables initiales et les variables principales :

> Rho <- diag(1/sqrt(diag(Sigma))) %*% ACP$vectors %*% diag(sqrt(ACP$values))
> rownames(Rho) <- colnames(x)

- Vérifier que la représentation des variables se trouve bien dans le cercle unité. A l'aide de la commande CercleCor(Rho), afficher et commenter le cercle des corrélations.
- Pour finir, nous pouvons représenter simultanément les individus et les variables dans la représentation "biplot". Discuter du résultat obtenu par la commande Biplot(CP,Rho).

➢ On peut aussi utiliser le package « FactoMineR »

```
> res.acp <- PCA(X, graph = FALSE)
> print(res.acp)
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 23 individuals, described by 10 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

➢ On peut aussi utiliser le package « factoextra »

Les fonctions suivantes de factoextra:

- get_eigenvalue(res.acp): Extraction des valeurs propres / variances des composantes principales

- fviz_eig(res.acp): Visualisation des valeurs propres
- get_pca_ind(res.acp), get_pca_var(res.acp): Extraction des résultats pour les individus et les variables, respectivement.
- fviz_pca_ind(res.acp), fviz_pca_var(res.acp): visualisez les résultats des individus et des variables, respectivement.
- fviz_pca_biplot(res.acp): Création d'un biplot des individus et des variables.

```
# Tester et commenter chaque ligne de commande:
> library("factoextra")
> eig.val <- get_eigenvalue(res.acp)
> eig.val
> fviz_eig(res.acp, addlabels = TRUE, ylim = c(0, 50))
> var <- get_pca_var(res.acp)
> var
# Coordonnées
> head(var$coord)
# Cos2: qualité de répresentation
> head(var$cos2)
# Contributions aux composantes principales
> head(var$contrib)
# Coordonnées des variables
> head(var$coord, 4)
> fviz_pca_var(res.acp, col.var = "blue")
> head(var$cos2, 4)
> fviz_cos2(res.acp, choice = "var", axes = 1:2)
# Colorer en fonction du cos2: qualité de représentation
> fviz_pca_var(res.acp, col.var = "cos2",  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE # Évite le chevauchement de texte
        )
head(var$contrib, 4)
library("corrplot")
corrplot(var$contrib, is.corr=FALSE)
# Contributions des variables à PC1
fviz_contrib(res.acp, choice = "var", axes = 1, top = 10)
# Contributions des variables à PC2
fviz_contrib(res.acp, choice = "var", axes = 2, top = 10)
fviz_contrib(res.acp, choice = "var", axes = 1:2, top = 10)
fviz_pca_var(res.acp, col.var = "contrib",  gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")   )
res.desc <- dimdesc(res.acp, axes = c(1,2), proba = 0.05)
# Description de la dimension 1 et 2
res.desc$Dim.1
res.desc$Dim.2
ind <- get_pca_ind(res.acp)
ind
head(ind$coord)
head(ind$cos2)
head(ind$contrib)
fviz_pca_ind (res.acp, col.ind = "cos2",gradient.cols = c("#00AFBB",
"#E7B800", "#FC4E07"), repel = TRUE # Évite le chevauchement de texte )
fviz_contrib(res.acp, choice = "ind", axes = 1:2)
fviz_pca_biplot(res.acp, repel = TRUE, col.var = "#2E9FDF", col.ind =
"#696969"  )
```

cancer.txt

Type	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	
Cancer1		22.92307692	69.46153846	73.76923077	2.692307692	82.07692308	1.1	1.185714286	4.235714286	5.230769231	3.635714286
Cancer2		17.56521739	65.04347826	79.73913043	51.41304348	49.39130435	22.13333333	7.314893617	6.15106383	1.02826087	9.234042553
Cancer3		23.45483871	70.67741935	78.5	39.70967742	71.08064516	16.76935484	9.032234375	8.915625	2.902222222	12.99047619
Cancer4		9.285714286	73.4	77.93333333	86.93333333	14.93333333	6.230769231	37.58461538	17.47142857	8.133333333	8.88
Cancer5		39.16666667	61.64	65.4	55.4	39.68	7.708333333	34.40769231	22.33846154	8.415384615	29.44615385
Cancer6		69.11111111	46.63157895	36.55789474	40.94736842	20.94736842	32.11111111	66.565	44.6	23.49	57.425
Cancer7		26.375	50.5	77.375	28	53.75	4.75	9.555555556	12.93333333	2.3125	11.51111111
Cancer8		0	92	76	0	86	0	0.3	1.2	0.4	1

Lien recommandé : http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/