# TP3 : Analyse en composantes principales (ACP)

Ce TP a pour objectif de faire une ACP du jeu de données `eaux`, dont vous avez fait les analyses uni- et bi-dimensionnelles lors du TP2. Vous rédigerez vos réponses et observations dans un rapport grâce à R Markdown en interprétant bien les résultats.

Avant d'entamer l'ACP, mettez-vous dans le dossier de travail, vérifiez qu'il contient bien les données `eauxModif.txt` enregistrées lors du dernier TP. Commencez par charger toutes les librairies dont nous aurons besoin au cours de ce TP, à savoir `shape`, `corrplot` et `FactoMineR`.

Par ailleurs, afin de calculer la variance non-corrigée comme dans le cours, créez la fonction `mavar` suivante :

```
> mavar <- function(x)
  {
    n=dim(as.matrix(x))[1]
    return(var(x)*(n-1)/n)
  }
```

## 1 Jeu de données `eaux`

### 1.1 Lecture des données

Importez les données à l'aide de la commande suivante :

```
> eaux = read.table("eauxModif.txt",header=TRUE)
```

Vérifiez à l'aide des fonctions `head()` et `str()` que votre jeu de données contient bien les acronymes des eaux comme noms de lignes, deux variables qualitatives `Type` et `Nature` sous forme de facteurs et 8 variables quantitatives.

### 1.2 Matrice de données brutes x

i. Pour faire l'ACP, nous n'utilisons que les variables quantitatives. Stockez-les dans une matrice `X`.

ii. Représentez les boites à moustaches de toutes les variables de `X` sur un même graphe (avec la fonction `boxplot`). Selon vous, vaudrait-il mieux faire une ACP centrée, ou une ACP centrée réduite ? Pensez à justifier votre raisonnement.

iii. Comme dans le TP2, représentez graphiquement les corrélations des variables quantitatives (vous pourrez utiliser la fonction `corrplot`). Quelles sont les variables linéairement corrélées (ou non) ?

Pour étudier ces données dans leur ensemble (prise en compte simultanée des 8 variables), nous effectuons une ACP. Afin de bien comprendre les sorties du logiciel, nous commencerons par faire l'ACP "à la main" avant d'interpréter les sorties obtenues avec le package `FactoMineR`.

Par ailleurs, nous allons faire l'ACP centrée dans un premier temps, puis, l'ACP centrée-réduite dans un second temps, avec les métriques usuelles $\mathtt{M} = I_p$ et $\mathtt{W} = \frac{1}{n} I_n$.

## 2 ACP centrée

### 2.1 À la main

i. Créez la matrice de travail `trav` des données centrées à l'aide de la commande `scale` (en précisant bien que l'on souhaite centrer mais pas réduire !). Vérifiez que les variables (i.e. les colonnes) sont bien centrées :

```
> apply(trav,2,mean)
```

ii. À l'aide de la commande `diag`, créez les matrices `M` et `W` associées aux métriques usuelles dans les espaces des individus et des variables respectivement.

iii. Calculez à l'aide des matrices `trav` et `W` la matrice `covX` de variance-covariance des données (correspondant à la matrice $\Gamma$ du cours dans le cas centré). On rappelle que le produit matriciel s'écrit $\%*\%$ dans R. Comparez `covX` avec le résultat des commandes `var` de R, et de `mavar` créé en début de TP.

iv. Calculez l'inertie globale, notée `Inertie`. Représentez les pourcentages d'inertie des variables de départ dans un diagramme en barres. Qu'observez-vous ?

v. Décomposition en composantes principales :

(a) Donnez les valeurs propres de la matrice (`covX` $\times$ `M`) grâce à la commande `eigen`. À quoi correspond chaque valeur propre ?

(b) Calculez leur somme. À quoi correspond cette valeur ?

(c) Représentez les pourcentages (cumulés ou non) d'inertie portés par chaque axe. Combien de composantes principales décidez-vous de garder ? Justifiez votre raisonnement.

(d) Stockez dans la matrice `A` les vecteurs propres de (`covX` $\times$ `M`). Que représentent-ils ? Vérifiez que `A` est bien de dimension $p \times p$.

(e) Stockez dans la matrice `C` les composantes principales. Vérifiez que la matrice `C` est bien de dimension $n \times p$. Que représentent les coordonnées de la première colonne de `C` ?

## 2.2 Avec `FactoMineR`

i. Voici un script exploitant certaines fonctionnalités de la librairie `FactoMineR` pour faire de l'ACP. Interprétez les sorties obtenues.

```
> res.acp <- PCA(eaux,scale.unit=FALSE,ncp=8,
+                quali.sup=1:2,graph=FALSE)
> attributes(res.acp)
> res.acp$eig[,"eigenvalue"]
> barplot(res.acp$eig[,"percentage of variance"],
+       names.arg=paste("Dim",1:8,sep="."),
+       main="Pourcentages d'inerties")
> res.acp$eig[,"cumulative percentage of variance"]
```

ii. À quoi correspondent les matrices suivantes :

```
> res.acp$svd$V
> res.acp$ind$coord
```

iii. Étude des individus :

(a) Représentez le graphe des individus et comparez avec les composantes principales calculées dans la partie 2.1 avec les commandes :

```
> plot(res.acp,choix="ind")
> plot(C[,1],C[,2]);abline(h=0);abline(v=0)
```

Que signifie l'option `choix="ind"` ? Que remarquez-vous ?

(b) Afin de rendre le graphe plus lisible, modifiez les options suivantes dans la fonction `plot` : `cex`, `invisible="quali"`.

(c) Coloriez les individus selon leur modalité pour la variable `Nature` (voir l'option `habillage`).

(d) Enfin, avec l'option `select`, n'affichez que les 10 points ayant la plus grande contribution dans la construction des deux premiers axes principaux. Vous pourrez également changer l'option `unselect`. Vous pouvez également sélectionner les individus selon leur qualité de représentation (désignée par le `cos2`).

(e) Interprétez le graphe des individus.

iv. Étude des variables :

(a) Représentez le graphe des corrélations des variables, et comparez-le au graphe suivant :

```
> plot(res.acp,choix="varcor")
> plot(cor(X,-C[,1]),cor(X,-C[,2]),
       xlim=c(-1,1),ylim=c(-1,1),asp=1)
> abline(h=0);abline(v=0);plotcircle(r=1)
```

Que signifie l'option `choix="varcor"` ? Vous pouvez également si besoin ajuster les options d'affichage comme pour les individus.

(b) Interprétez le graphe des variables.

(c) Remarque : afin de visualiser plus que les deux premières composantes principales, représentez, à l'aide de la fonction `corrplot`, les corrélations entre les variables de départ et toutes les composantes principales calculées. Interprétez les résultats.

# 3 ACP centrée-réduite

Reprenons l'étude dans le cas centré réduit.

## 3.1 À la main (si vous avez du temps, sinon, passez directement à la partie 3.2)

i. Créez la matrice `trav2` des données centrées et réduites (à l'aide de la variance non corrigée !) :

```
> trav2 <- trav %*% diag( (diag(mavar(X)))^(-1/2) )
> apply(trav2,2,mean) ; apply(trav2,2,mavar)
```

Comparez `trav2` avec le résultat obtenu avec la commande `scale`.

ii. Calculez à l'aide des matrices `trav2` et `W` la matrice `corX` des corrélations (correspondant encore à la matrice $\Gamma$ du cours, mais dans le cas centré-réduit) des variables initiales. Comparez `corX` avec le résultat des commandes `cor` de R.

iii. Que vaut l'inertie globale, dans le cas d'une ACP centrée-réduite ?

iv. Calculez les nouveaux axes principaux, leurs inerties axiales et les nouvelles composantes principales.

## 3.2 Avec `FactoMineR`

i. Quelle option faut-il changer pour faire une ACP centrée réduite avec FactoMineR ? Stockez les résultats dans `res.acp2` et comparez-les avec les inerties et les composantes principales calculées dans la partie 3.1

ii. Combien de composantes principales gardez-vous ?

iii. Représentez le graphe des individus (en ajustant les options d'affichage).

iv. Représentez le graphe des variables.

v. Quelles différences remarquez-vous avec l'ACP centrée ? Interprétez les résultats.

vi. Représentez le graphe des individus et des variables impliquant les axes 1 et 3. Interprétez les résultats.

# 4 Pour aller plus loin

Voici un script exploitant certaines fonctionnalités de la librairie `factoextra` apportant une aide à l'interprétation.

```
> library("factoextra")
```

Interprétez les sorties obtenues.

i. Étude des inerties :

```
> eig.val <- get_eigenvalue(res.acp2) ; eig.val
> fviz_eig(res.acp2, addlabels = TRUE, ylim = c(0,50))
```

ii. Étude des individus :

```
> fviz_pca_ind(res.acp2,col.ind=eaux$Type)

> fviz_pca_ind(res.acp2, geom.ind = "point",
+              col.ind = eaux$Type,
+              addEllipses = TRUE,
+              legend.title = "Groups")

> fviz_pca_ind(res.acp2, geom.ind = "point",
+              col.ind = eaux$Nature,
+              addEllipses = TRUE,
+              legend.title = "Groups")

fviz_pca_ind(res.acp2, col.ind = "cos2", repel = TRUE,
+             gradient.cols = c("#00AFBB", "#E7B800",
+                               "#FC4E07"))

> fviz_contrib(res.acp2, choice = "ind", axes = 1:2)
```

iii. Étude des variables :

```
> var <- get_pca_var(res.acp2) ; var

> fviz_pca_var(res.acp2, col.var = "cos2",repel = TRUE,
+              gradient.cols = c("#00AFBB", "#E7B800",
+                                "#FC4E07"))
```

Remarquons que l'option `repel=TRUE` évite le chevauchement de texte, et l'option `addEllipses = TRUE` affiche des ellipses de concentration.