# TP1 : Introduction et ACP

## 1 Mise en place

Le logiciel libre R peut être utilisé de plusieurs manières. Il est possible de l'utiliser en mode interactif (commande `R --vanilla`) bien que cela ne soit pas particulièrement adapté à des séances de travaux pratiques. En effet, pour garder une trace de votre travail et pour éviter de taper plusieurs fois les mêmes commandes, il est plus pratique de travailler dans un fichier texte que l'on chargera dans R à l'aide de la commande `source`. Il existe également de nombreux outils (Rcmdr, rgedit, ...) d'interface graphique pour R.

Afin de faciliter la récupération des données et de disposer de certaines fonctions graphiques dans la suite de cette séance, nous commençons par charger le script *tp1.R*,

```
source("http://www.math.univ-toulouse.fr/~xgendre/ens/m2se/tp1.R")
```

## 2 Exemples jouets

Dans cette section, nous reprenons les deux exemples introductifs vus en cours. Les données considérées sont brièvement explorées puis nous procédons à l'analyse en composantes principales (ACP). Enfin, nous attirons l'attention sur quelques outils existants.

### 2.1 Notes

Le premier jeu de données que nous considérons est celui des notes obtenues par $n = 9$ élèves dans $p = 4$ matières. Pour récupérer ces données dans une matrice $X$, nous entrons la commande

```
X <- DataNotes()
```

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 aux notes des 4 matières (commande `boxplot`),
- que donne la commande `summary(t(X))` ?

Pour procéder à l'ACP, nous commençons par centrer les données,

```
Xbar <- scale(X, scale = F)
```

Afficher les boîtes à moustaches associées aux données centrées et rappeler pourquoi nous n'avons pas considéré utile de normaliser les données dans le cours.

Exécuter chacune des lignes suivantes et expliquer ce qu'elles font et ce que l'on obtient :

```
Sigma <- t(Xbar) %*% Xbar/nrow(Xbar)
ACP <- eigen(Sigma)
ACP$values
cumsum(ACP$values)/sum(ACP$values)
plot(ACP$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 `C` des composantes principales.
- calculer la qualité de la représentation de chacun des élèves et stocker les résultats dans un vecteur `cos2`.
- calculer et commenter les contributions des élèves aux variables principales.

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

```
plot(C[, 1:2], pch = 2, cex = cos2)
text(C[, 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(C,Rho)`.

## 2.2 Arcs et arbalètes

Le second jeu de données que nous avons utilisé en cours était celui relatif aux arcs et arbalètes du jeu vidéo "The Elder Scrolls V : Skyrim". Il contient les statistiques de $n = 14$ armes relatives à $p = 4$ variables et peut être obtenu par la commande suivante,

```
X <- DataSkyrim()
```

Comme pour l'exemple précédent, explorer le jeu de données :
- afficher la matrice $X$,
- utiliser la commande `summary(X)` pour visualiser quelques informations importantes,
- afficher les boîtes à moustaches associées aux observations des 4 variables.

Pour procéder à l'ACP, nous commençons par centrer les données,

```
Xbar <- scale(X, scale = F)
```

Afficher les boîtes à moustaches associées aux données centrées. Commenter ce graphique et exécuter les lignes suivantes en les expliquant,

```
S <- t(Xbar) %*% Xbar/nrow(Xbar)
M <- diag(1/diag(S))
Xbar <- Xbar %*% sqrt(M)
Sigma <- t(Xbar) %*% Xbar/nrow(Xbar)
```

En particulier, expliquer le rôle de la ligne `Xbar <- Xbar %*% sqrt(M)`.

Reprendre les questions de l'exemple précédent et procéder à l'ACP de ces données.

### 2.3 Autres outils

Le logiciel R, comme beaucoup d'autres logiciels, propose des outils pour réaliser directement l'ACP d'un jeu de données. Parmi les commandes de base, il faut citer `prcomp` et `princomp`. Ces deux commandes ne diffèrent que par peu de points et nous ne nous intéresserons que à `prcomp`.

Reprenons notre premier exemple et affichons certains graphiques,

```
X <- DataNotes()
ACP <- prcomp(X)
plot(ACP)
biplot(ACP)
```

Avec l'aide sur la commande `prcomp`, que contient la variable `ACP` ? Comparer la matrice de changement de base avec ce que nous avions obtenu. Qu'est-ce que cela implique sur la matrice des composantes principales ? Commenter les différences avec la section 2.1.

De même, reprenez le deuxième exemple et commenter les différences avec la méthode du cours. Comment indiquer directement à `prcomp` que l'on souhaite travailler avec les données normalisées ?

## 3 Crimes aux États-Unis

Dans cette section, nous considérons les statistiques de crimes commis aux États-Unis en 1977. Ce jeu de données contient les nombres de crimes pour 100 000 habitants pour $p = 7$ types de crimes dans les $n = 50$ états. Pour récupérer ces données,

```
X <- DataCrimes()
```

Explorer rapidement ces données et calculer l'ACP. A l'aide de l'éboulis des valeurs propres, quel nombre $d$ de variables principales utiliseriez-vous ? A quelle part d'inertie expliquée $r_d$ cela correspond-il ? Calculer la contribution des variables initiales à l'inertie des variables principales ? Discuter des résultats obtenus par rapport à votre choix de $d$.

Représenter les variables sur le cercle des corrélations à l'aide de la fonction `CercleCor`. A l'aide de ce graphiques, comment allez-vous interpréter les axes dans le plan principal ? Représenter les individus dans le plan principal avec la fonction `plot`. Cette représentation est-elle suffisante pour résumer correctement l'information des données ? Pour $d > 2$, en utilisant la fonction `pairs`, représenter les données dans les plans engendrés par toutes les paires d'axes principaux parmi les $d$ premiers.

Calculer la contribution de chaque état à l'inertie des axes. Repérer quelques états dont certaines contributions sont importantes. A l'aide de la figure 1, pouvez-vous donner une interprétation géographique aux axes de l'ACP ?

## 4 Pays de l'OCDE

Le dernier exemple traité dans cette séance est relatif à des données issues de l'observatoire de l'OCDE. Pour chaque pays membre et pour chacune des années 1975, 1977, 1979 et 1981, nous connaissons les valeurs prises par les variables suivantes :

1. NATA : taux brut de natalité,
2. CHOM : taux de chômage,
3. APRI : pourcentage d'actifs dans le secteur primaire,


FIGURE 1 – Carte des États-Unis

4. ASEC : pourcentage d'actifs dans le secteur secondaire,
5. PIB : produit intérieur brut (par habitant),
6. FBCF : formation brute de capital fixe (par habitant),
7. INFL : hausse des prix,
8. RECC : recettes courantes (par habitant),
9. MINF : mortalité infantile,
10. PROT : consommation de protéines animales (par habitant),
11. NRJ : consommation d'énergie (par habitant).

Pour récupérer ces données,

```
Data <- DataOCDE()
```

La première colonne contient les initiales du pays concerné et les colonnes suivantes contiennent les $p = 11$ variables.

Après une inspection de ces données, décider de les normaliser ou non et procéder à l'ACP. Étudier le nombre d'axes à retenir, la qualité de la représentation, les contributions aux axes, ...

Dans le plan principal, représenter les $n = 68$ points relatifs aux 17 pays considérés. Mettre en évidence les groupes de 4 points correspondants à chaque pays et les relier entre eux de façon à visualiser le "sens de variation" de chaque pays dans le plan principal. A l'aide de votre étude préliminaire, interpréter ce graphique.