M1- Mathématiques Appliquées Paris-Saclay - ENSTA
Apprentissage statistique (STA203), 2024-2025
Christine.Keribin@universite-paris-saclay.fr, Justine Lebrun (SNCF)

(Re)prise en main du logiciel R et bases de l'ACP

Ce TP conjugue deux objectifs: appréhender les bases mathématiques de l'ACP sur un exemple basique en deux dimensions tout en se (re-) familiarisant au langage R (en particulier, utilisation du package `ggplot2`)

1. Acquisition des données: lire (`read.table`) le jeu de données `seeds` à partir du site `http://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt` Lister les premières observations (`head`)
   (a) Vérifier qu'il contient 210 observations et 8 variables (`dim`): il s'agit de mesures de grandeurs caractéristiques pour 210 échantillons de grains.
   (b) Ajouter le nom des variables (`names`): `area`, `perimeter`, `compact`, `length`, `width`, `asym`, `lgroove`, `variety`.
   (c) Enregistrer (`save`) le `data.frame` obtenu sous forme d'un fichier `.RData`
   (d) Détruire le data.frame (`rm`) et le recharger par la lecture du fichier précédemment créé(`load`).
2. Commenter les résultats de la fonction `summary`. Cette commande est-elle adaptée à la dernière variable? Enregistrer la variable `width` dans le vecteur `V1` et la variable `length` dans le vecteur `V2`. Quelle est la longueur de ces vecteurs?
3. Visualisation: représenter `V2` en fonction de `V1` (`ggplot()`, `aes()`, `geom_point()`) Afficher le centre de gravité du nuage de points. Représenter les segments (`geom_segment()`) reliant les points à leur centre de gravité. Calculer l'inertie de la configuration. Faire la même représentation avec des variables centrées-réduites. L'inertie de la nouvelle configuration est-elle la même? Afficher les deux graphes dans une fenêtre partagée en deux (`cowplot::plot_grid()`). On poursuivra avec les variables centrées-réduites
4. Recherche d'une direction maximisant l'inertie projetée:
   (a) Ajouter la représentation de l'axe des abscisses (`geom_hline()`).
   (b) Représenter la projection des points sur cet axe. Calculer l'inertie de cette représentation et le carré des distances des points à leurs projections.
   (c) Faire de même avec la droite de régression (`lm`, `geom_smooth(method='lm')` de `V2` en fonction de `V1`. Superposer les segments joignant chaque point à son ajustement, puis ceux joignant chaque point à sa projection orthogonale. Expliquer la différence d'analyse.
   (d) Représenter la première bissectrice. Calculer l'inertie de cette représentation et le carré des distances des points à leurs projections. Cette droite est-elle optimale vis à vis du critère de l'inertie (`eigen`, `cor`)?
   (e) Retrouver ces résultats avec la fonction `PCA` du package `FactoMineR` que vous devrez peut-être installer (fenêtre `packages`, puis sélectionner `install`), puis charger (`library(FactoMineR)`).

Christine.Keribin@universite-paris-saclay.fr, Justine Lebrun (SNCF)

Le package `ggplot2`

C'est la nouvelle interface de plot. Puissante, elle permet en général de mieux élaborer les graphes, en particulier quand il y a des sous-groupes d'observations (ce qui n'est pas le cas dans ce TP). Elle segmente les instructions et utilise une grammaire de graphique dont les principaux éléments sont :

- Data (`ggplot()`): le jeu de données contenant les variables utilisées
- Aesthetics (`aes()`): les variables à représenter; on peut inclure des couleurs ou des tailles si ces dernière sont associées à des variables; ou définir des variables externes au jeu de données initial
- Geometrics (`geom_...()`): le type de représentation souhaitée (`...` = `point`, `line`, `abline`, `segment`, `text`, `bar`, `boxplot`, `histogram`, `barplot`, etc.
- Statistics (`stat_...()`): éventuelles transformations de données
- Scales (`scale_...()`): pour le contrôle

```
> library(ggplot2)
> D = data.frame(X=seq(-2*pi, 2*pi, by = 0.01))
> ggplot(D)+ aes(x=X, y=sin(X))+ geom_line()
```

Bibliopgraphie:

- R pour la statistique et la science des donnéées (PUR 2018), Cornillon *et al*, `https://r-stat-sc-donnees.github.io/`
- `http://fermin.perso.math.cnrs.fr/Files/ggplot_visu.html`
- le *cheat sheet* dans les documents du cours