ANNE SABOURIN anne.sabourin@u-paris.fr,

# TP3 : EVALUATION

**Exercice 1.** Pour cet exercice, nous travaillons sur le jeu de données `Ozone`. Il est disponible à l'adresse `"https://r-stat-sc-donnees.github.io/ozone.txt"`. Pour le récupérer on pourra procéder ainsi :

```
url <- "https://r-stat-sc-donnees.github.io/ozone.txt"

file_name <- "ozone.txt"
file_path <- "./"

# Call the download.file() function, passing in the URL and
# file name/location as arguments
download.file(url, paste(file_path, file_name, sep = ""), mode = "wb")

ozone <- read.table(file_name)
head(ozone)
```

Ce jeu de données contient des informations concernant la pollution de l'air, notamment :

- `maxO3` : maximum de concentration d'ozone observé sur la journée en $\mu$gr/m3
- `T9`, `T12`, `T15` : Température observée à 9, 12 et 15h,
- `Ne9`, `Ne12`, `Ne15` : Nébulosité observée à 9, 12 et 15h,
- `Vx9`, `Vx12`, `Vx15` : Composante E-O du vent à 9, 12 et 15h,
- `maxO3v` : Teneur maximum en ozone observée la veille,
- `vent` : orientation du vent à 12h,
- `pluie` : occurrence ou non de précipitations.

1. Après avoir enlevé les variables qualitatives `vent` et `pluie`, ajuster un modèle de régression linéaire multiple non pénalisée pour expliquer la concentration d'ozone `maxO3` en fonction des autres variables. Identifier les variables significatives. (On peut utiliser simplement la fonction `lm` de `R`.
2. Centrer et réduire les données puis ajuster une régression ridge en faisant varier la pénalité $\lambda$ sur une grille pertinente (par exemple celle automatiquement choisie par le package `glmet`).
3. Ajuster cette fois une régression Lasso en faisant varier $\lambda$ sur la grille puis tracer le chemin de régularisation de chacune des variables.
4. Ajuster finalement une régression Elastic Net puis tracer le chemin de régularisation de chacune des variables.
5. Pour chacune des trois régressions linéaires pénalisées, calibrer la pénalité par validation croisée puis comparer les variables du modèle retenu. On pourra utiliser la fonction `cv.glmnet`
6. On cherche dans cette question à comparer les trois approches suivantes en terme de risque de prédiction :
   (a) Sélection les variables significatives dans un modele linéaire simple puis ajuster un modele linéaire sur les variables sélectionnées
   (b) Sélectionner des variables à l'aide du LASSO pour un certain $\lambda$ choisi par cross-validation puis ajuster le modèle linéaire sur les variables sélectionnées
   (c) Effectuer une ACP du jeu de données des covariables, sélectionner une dimension à l'aide d'un screeplot puis ajuster un modèle linéaire.

   On rappel que faire l'ACP de $X$ consiste seulement à calculer les vecteurs propres et valeurs propres de $X^\top X$, puis après avoir choisi un certain nombre de valeurs propres 'plus grandes' que les autres, projeter le jeu de données sur les premiers vecteurs propres correspondant. Dans `R`, on peut effecteur une décomposition en valeurs propres et vecteurs propres par la fonction `eig`.
   Après avoir mis de côté $1/3$ des données pour servir de jeu test à l'étape d'évaluation du risque de prédiction $\mathbb{E}\left((Y-\hat{Y})^2\right)$, mettez en oeuvre les trois approches ci-dessus. Comparez les résultats. Discutez les limites de vos conclusions ;
7. Construire un exemple sur données simulées tel que la méthode LASSO sera meilleure que la méthode l'ACP suivie des moindres carrés. Expliquez d'abord pourquoi ce serait le cas puis vérifiez -le sur un jeu de données simulées (disons de taille $n = 500$)