TP 5 - Compte rendu individuel

Matthias MAZET

2025-12-08

Contents


```
# Packages statistiques
library(glmnet)

# Packages de style
library(ggplot2)
library(ggpubr)
library(GGally)
library(tidyverse)
```

0.1 *Préparation des données.*

```
data <- read.csv("day.csv")
```

Pour simplifier les méthodes, on supprime la variable “dteday” qui est au format “chr”. On supprime aussi “casual” et “registered” qui, en les sommant, donne la valeur de la variable à prédire “cnt”. On vérifie ensuite les valeurs manquantes et la dimension des données.

```
# Nettoyage
data <- data[, -c(2, 14, 15)] # Suppression de variables
paste("Nombre de valeurs manquantes :", sum(is.na(data)))
```

```
## [1] "Nombre de valeurs manquantes : 0"
```

```
paste("Dimension du dataset :", dim(data)[1], "x", dim(data)[2])
```

```
## [1] "Dimension du dataset : 731 x 13"
```

```
# Format matriciel
X <- as.matrix(data[, -13])
y <- as.matrix(data$cnt)
```

Nous sommes dans un cas où n > p et où il n’y a pas de valeurs manquantes.

0.2 *Estimation et stabilité de l’estimateur.*

On ajuste un modèle de Poisson via `glmnet`. Pour cela, on précise `family = "poisson"` dans les fonctions de `glmnet`. On conserve arbitrairement la valeur de `lambda.1se` obtenue par validation croisée autonome.

```
set.seed(5)

# Fit du modèle
lambda_lasso <- cv.glmnet(x = X, y = y, alpha = 1, family = "poisson")$lambda.1se
mod_lasso <- glmnet(x = X, y = y, alpha = 1, lambda = lambda_lasso, family = "poisson")

# Variables sélectionnées (avec suppression de l'intercept)
paste(
  "Variables sélectionnées (coef. non nuls) :",
  paste(colnames(data)[which(coef(mod_lasso) != 0)[-1] - 1], collapse = ", ")
)
```

```
## [1] "Variables sélectionnées (coef. non nuls) : instant, season, yr, holiday, weekday, weathersit, temp
```

Regardons maintenant sa stabilité sur 20 itérations. pour chaque itération, on regarde quelles variables sont sélectionnées à `lambda.1se` et on stocke l’information dans une matrice :

```
set.seed(5)

# Matrices des variables sélectionnées à chaque itérations
mat_vars <- matrix(data = 0, ncol = ncol(X), nrow = 20)
colnames(mat_vars) <- colnames(X)

# Itérations boostrap
for (i in 1:20) {
  # Individus bootstrap
  ind_al <- sample(1:nrow(data), size = nrow(data), replace = TRUE)
  X <- as.matrix(data[ind_al, -13])
```

```r
  y <- data[ind_al, 13]
  # Fit du modèle
  lambda_lasso <- cv.glmnet(x = X, y = y, alpha = 1, family = "poisson")$lambda.1se
  mod_lasso <- glmnet(x = X, y = y, alpha = 1, lambda = lambda_lasso, family = "poisson")
  # Variables sélectionnées à lambda.1se
  mat_vars[i, which(coef(mod_lasso) != 0)[-1] - 1] <- 1
}

# Visualisation
### Dataset de résultat au format long
df_lasso <- t(mat_vars) %>%
  as.data.frame() %>%
  mutate(variable = rownames(.)) %>%
  pivot_longer(cols = -variable, names_to = "iteration", values_to = "selected") %>%
  mutate(iteration = as.numeric(gsub("V", "", iteration)))
### Figure ggplot
ggplot(df_lasso, aes(x = iteration, y = variable, fill = selected)) +
  geom_tile() +
  scale_fill_gradientn(colours = c("white", "black"), limits = c(0, 1), name = "Sélection") +
  labs(x = "Itération", y = "Indices des variables", subtitle = "Lasso - lambda.1se") +
  theme_light() +
  theme(
    plot.subtitle = element_text(size = 10, face = "bold"),
    axis.title = element_text(size = 10, face = "bold"),
    axis.text = element_text(size = 8),
    axis.line = element_line(colour = "grey"),
    panel.border = element_blank(),
    panel.grid = element_blank(),
    legend.title = element_text(size = 10, face = "bold"),
    legend.position = "right"
  )
```

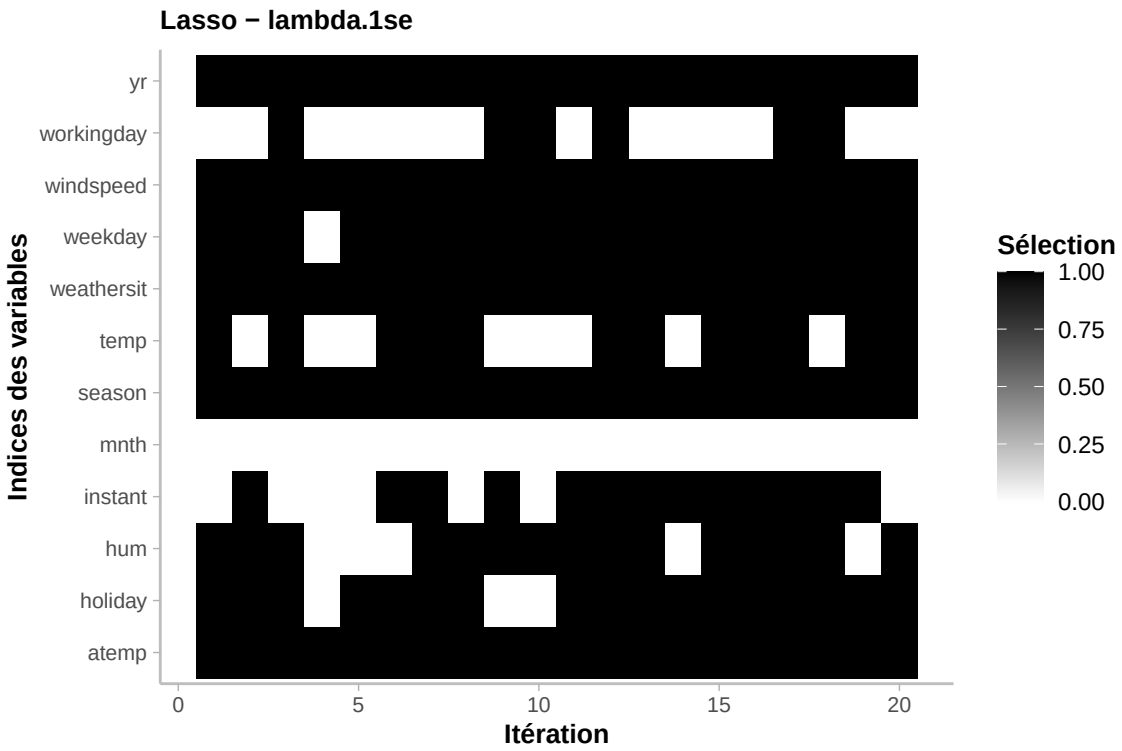

Figure 1: Stabilité à lambda.1se au fil des itérations

La figure précédente **(Fig. 1)** présente les variables sélectionnées à chaque itération, avec 1 = variable sélectionnée et 0 sinon (la légende de couleur devrait être discrète mais ggplot ne le fait pas). On peut constater que certaines variables ("yr" ou "windspeed" par exemple) sont toujours sélectionnées, tandis que d'autres le sont aléatoirement

suivant l'itération ("hum" ou "temp" par exemple). La variable "mnth" n'est à l'inverse jamais sélectionnée. La méthode présente donc une légère instabilité sur la sélection de variables, ce qui peut témoigner de multicolinéarité entre certaines variables (notamment celles sélectionnées quelquefois mais pas toutes). Cette remarque correspond à l'interprétation de la (in)stabilité dans la sélection de variables.

0.3 *Sélection post-inférence.*

1. *Construction d'un intervalle de confiance.*

```
alpha <- .05
n <- nrow(data)
# Estimation des beta par glmnet
beta_hat <- coef(mod_lasso)[-1]
sd_beta_hat <- numeric(length(beta_hat))
for (i in 1:length(beta_hat)) {
  sd_beta_hat[i] <- sd(beta_hat[i])
}

# Intervalle de confiances
borne_inf <- beta_hat - (qnorm(1 - alpha/2)*sd_beta_hat / sqrt(n))
borne_sup <- beta_hat + (qnorm(1 - alpha/2)*sd_beta_hat / sqrt(n))
```