Régression en grande dimension

Laurent Rouvière

2020-09-22

Table des matières

Présentation

Ce tutoriel présente quelques exercices d'application du cours `Modèle linéaire en grande dimension`. On pourra trouver

- les supports de cours associés à ce tutoriel ainsi que les données utilisées à l'adresse suivante https://lrouviere.github.io/stat_grand_dim/ ;
- le tutoriel sans les correction à l'url https://lrouviere.github.io/TUTO_GRANDE_DIM/
- le tutoriel avec les corrigés (à certains moment) à l'url https://lrouviere.github.io/TUTO_GRANDE_DIM/correction.

Il est recommandé d'utiliser **mozilla firefox** pour lire le tutoriel.

Des connaissances de base en R et en statistique (modèles de régression) sont nécessaires. Le tutoriel se structure en 4 parties :

- `Fléau de la dimension` : identification du problème de la dimension pour le problème de régression ;

— `Régression sur composantes` : présentation des algorithmes **PCR** et **PLS** ;
— `Régressions pénalisées` : régularisation à l'aide de pénalités de type **Ridge/Lasso**
— `Modèle additif` : conservation de la structure additive du modèle linéaire mais modélisation non paramétrique des composantes.

1 Introduction à la grande dimension

Nous proposons ici d'illustrer le problème de la grande dimension en régression. On commencera par étudier, à l'aide de simulation, ce problème pour l'estimateurs des k plus proches voisins, puis pour les estimateurs des moindres carrés dans le modèle linéaire. Quelques exercices sont ensuite proposées pour calculer les vitesses de convergence de ces estimateurs dans des modèles simples.

1.1 Fléau de la dimension pour les plus proches voisins

La fonction suivante permet de générer un échantillon d'apprentissage et un échantillon test selon le modèle

$$Y = X_1^2 + \cdots + X_p^2 + \varepsilon$$

où les X_j sont uniformes i.i.d de loi uniorme sur $[0,1]$ et le bruit ε suit une loi $\mathcal{N}(0, 0.5^2)$.

```
simu <- function(napp=300,ntest=500,p=3,graine=1234){
  set.seed(graine)
  n <- napp+ntest
  X <- matrix(runif(n*p),ncol=p)
  Y <- apply(X^2,1,sum)+rnorm(n,sd=0.5)
  Yapp <- Y[1:napp]
  Ytest <- Y[-(1:napp)]
  Xapp <- data.frame(X[1:napp,])
  Xtest <- data.frame(X[-(1:napp),])
  return(list(Xapp=Xapp,Yapp=Yapp,Xtest=Xtest,Ytest=Ytest))
}
df <- simu(napp=300,ntest=500,p=3,graine=1234)
```

La fonction **knn.reg** du package **FNN** permet de construire des estimateurs des k plus proches voisins en régression. On peut par exemple faire du 3 plus proches voisins avec

```
library(FNN)
mod3ppv <- knn.reg(train=df$Xapp,y=df$Yapp,k=3)
```

Parmi toutes les sorties proposées par cette fonction on a notamment

```
mod3ppv$PRESS
[1] 98.98178
```

qui renvoie la somme des carrés des erreurs de prévision par validation croisée Leave-One-Out (LOO). On peut ainsi obtenir l'erreur quadratique moyenne par LOO

```
mod3ppv$PRESS/max(c(nrow(df$Xapp),1))
[1] 0.3299393
```

1. Construire la fonction **sel.k** qui admet en entrée :

   — une grille de valeurs possibles de plus proches voisins (un vecteur).
   — une matrice **Xapp** de dimension $n \times p$ qui contient les valeurs variables explicatives.
   — un vecteur **Yapp** de dimension n qui contient les valeurs de la variable à expliquer

   et qui renvoie en sortie la valeur de k dans la grille qui minimise l'erreur LOO présentée ci-dessus.

Une fois la fonction créée, on peut calculer l'erreur de l'estimateur sélectionné sur un échantillon test avec

```
k.opt <- sel.k(seq(1,50,by=5),df$Xapp,df$Yapp)
prev <- knn.reg(train=df$Xapp,y=df$Yapp,test=df$Xtest,k=k.opt)$pred
mean((prev-df$Ytest)^2)
```

2. On souhaite comparer les erreurs des règles des k plus proches voisins en fonction de la dimension. On considère 4 dimensions collectées dans le vecteur `DIM` et la grille de valeurs de k suivantes :

```
DIM <- c(1,5,10,50)
K_cand <- seq(1,50,by=5)
```

Pour chaque valeur de dimension répéter $B = 100$ fois :

— simuler un échantillon d'apprentissage de taille 300 et test de taille 500
— calculer la valeur optimale de k dans **K_cand** grâce à **sel.k**
— calculer l'erreur de l'estimateur sélectionné sur un échantillon test.

On pourra stocker les résultats dans une matrice de dimension $B \times 4$.

3. A l'aide d'indicateurs numériques et de boxplots, comparer la distribution des erreurs en fonction de la dimension.
4. Conclure

1.2 Influence de la dimension dans le modèle linéaire

En vous basant sur l'exercice précédent, proposer une illustration qui peut mettre en évidence la précision d'estimation dans le modèle linéaire en fonction de la dimension. On pourra par exemple considérer le modèle linaire suivant

$$Y = X_1 + 0X_2 + \cdots + 0X_p + \varepsilon$$

et étudier la performance de l'estimateur MCO du coefficient de X_1 pour différentes valeurs de p. Par exemple avec p dans le vecteur

```
DIM <- c(0,50,100,200)
```

Les données pourront être générées avec la fonction suivante

```
n <- 250
p <- 1000
X <- matrix(runif(n*p),ncol=p)
simu.lin <- function(X,graine){
  set.seed(graine)
  Y <- X[,1]+rnorm(nrow(X),sd=0.5)
  df <- data.frame(Y,X)
  return(df)
}
```

1.3 Exercices

Exercice 1.1 (Distances entre deux points, voir Giraud (2015))**.**

Soit $X^{(1)} = (X_1^{(1)}, \ldots, X_p^{(1)})$ et $X^{(2)} = (X_1^{(2)}, \ldots, X_p^{(2)})$ deux variables aléatoires indépendantes de loi uniforme sur l'hypercube $[0,1]^p$. Montrer que

$$\mathbf{E}[\|X^{(1)} - X^{(2)}\|^2] = \frac{p}{6} \quad \text{et} \quad \sigma[\|X^{(1)} - X^{(2)}\|^2] \approx 0.2\sqrt{p}.$$

Exercice 1.2 (Vitesse de convergence pour l'estimateur à noyau)**.**

On considère le modèle de régression

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, n$$

où $x_1, \dots, x_n \in \mathbb{R}^d$ sont déterministes et $\varepsilon_1, \dots, \varepsilon_n$ sont des variables i.i.d. d'espérance nulle et de variance $\sigma^2 < +\infty$. On désigne par $\|\,.\,\|$ la norme Euclidienne dans $\mathbb{R}^d$. On définit l'estimateur localement constant de m en $x \in \mathbb{R}^d$ par :

$$\hat{m}(x) = \underset{a \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - a)^2 K\left(\frac{\|x_i - x\|}{h}\right)$$

où $h > 0$ et pour $u \in \mathbb{R}, K(u) = \mathbf{1}_{[0,1]}(u)$. On suppose que $\sum_{i=1}^n K\left(\frac{\|x_i - x\|}{h}\right) > 0$.

1. Donner la forme explicite de $\hat{m}(x)$.
2. Montrer que
$$\mathbf{V}[\hat{m}(x)] = \frac{\sigma^2}{\sum_{i=1}^n K\left(\frac{\|x_i - x\|}{h}\right)}$$
et
$$\mathbf{E}[\hat{m}(x)] - m(x) = \frac{\sum_{i=1}^n (m(x_i) - m(x)) K\left(\frac{\|x_i - x\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x_i - x\|}{h}\right)}.$$
3. On suppose maintenant que m est Lipschitzienne de constante L, c'est-à-dire que $\forall (x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$
$$|m(x_1) - m(x_2)| \leq L\|x_1 - x_2\|.$$
Montrer que
$$|\text{biais}[\hat{m}(x)]| \leq Lh.$$
4. On suppose de plus qu'il existe une constante C_1 telle que
$$C_1 \leq \frac{\sum_{i=1}^n \mathbf{1}_{B_h}(x_i - x)}{n\text{Vol}(B_h)},$$
où $B_h = \{u \in \mathbb{R}^d : \|u\| \leq h\}$ est la boule de rayon h dans $\mathbb{R}^d$ et $\text{Vol}(A)$ désigne le volume d'un ensemble $A \subset \mathbb{R}^d$. Montrer que
$$\mathbf{V}[\hat{m}(x)] \leq \frac{C_2 \sigma^2}{nh^d},$$
où C_2 est une constante dépendant de C_1 et d à préciser.
5. Déduire des questions précédentes un majorant de l'erreur quadratique moyenne de $\hat{m}(x)$.
6. Calculer h_{opt} la valeur de h qui minimise ce majorant. Que vaut ce majorant lorsque $h = h_{\text{opt}}$? Comment varie cette vitesse lorsque d augmente ? Interpréter.

2 Régression sur composantes

Les performances des estimateurs classiques (MCO) des paramètres du modèle linéaire

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon$$

peuvent se dégrader lorsque la dimension d est grande ou en présence de dépendance linéaire entre les variables explicatives. Les régressions sur composantes consistent à trouver de nouvelles composantes $Z_k, j = k, \dots, q$ avec $q \leq p$ qui s'écrivent le plus souvent comme des combinaisons linéaires des X_j dans l'idée de diminuer le nombre de paramètres du modèle ou la dépendance entre les covariables. Il existe plusieurs façons de construire ces composantes, dans cette partie nous proposons :

— la **régression sous composantes principales (PCR)** : il s'agit de faire simplement une ACP sur la matrice des variables explicatives ;
— la **régression partial least square (PLS)** qui fait intervenir la variable cible dans la construction des composantes.

Nous commençons par un bref rappel sur la sélection de variables.

2.1 Sélection de variables

On considère le jeu de données `ozone.txt` où on cherche à expliquer la concentration maximale en ozone relevée sur une journée (variable `maxO3`) par d'autres variables essentiellement météorologiques.

```
ozone <- read.table("data/ozone.txt")
head(ozone)
         maxO3   T9  T12  T15 Ne9 Ne12 Ne15     Vx9    Vx12
20010601    87 15.6 18.5 18.4   4    4    8  0.6946 -1.7101
20010602    82 17.0 18.4 17.7   5    5    7 -4.3301 -4.0000
20010603    92 15.3 17.6 19.5   2    5    4  2.9544  1.8794
20010604   114 16.2 19.7 22.5   1    1    0  0.9848  0.3473
20010605    94 17.4 20.5 20.4   8    8    7 -0.5000 -2.9544
20010606    80 17.7 19.8 18.3   6    6    7 -5.6382 -5.0000
            Vx15 maxO3v  vent pluie
20010601 -0.6946     84  Nord   Sec
20010602 -3.0000     87  Nord   Sec
20010603  0.5209     82   Est   Sec
20010604 -0.1736     92  Nord   Sec
20010605 -4.3301    114 Ouest   Sec
20010606 -6.0000     94 Ouest Pluie
```

1. Ajuster un modèle linéaire avec `lm` et analyser la pertinence des variables explicatives dans le modèle.
2. Expliquer les sorties de la commande

```
library(leaps)
mod.sel <- regsubsets(maxO3~.,data=ozone,nvmax=14)
summary(mod.sel)
Subset selection object
Call: regsubsets.formula(maxO3 ~ ., data = ozone, nvmax = 14)
14 Variables  (and intercept)
          Forced in Forced out
T9            FALSE      FALSE
T12           FALSE      FALSE
T15           FALSE      FALSE
Ne9           FALSE      FALSE
Ne12          FALSE      FALSE
Ne15          FALSE      FALSE
Vx9           FALSE      FALSE
Vx12          FALSE      FALSE
Vx15          FALSE      FALSE
maxO3v        FALSE      FALSE
ventNord      FALSE      FALSE
ventOuest     FALSE      FALSE
ventSud       FALSE      FALSE
pluieSec      FALSE      FALSE
1 subsets of each size up to 14
Selection Algorithm: exhaustive
```

```
          T9  T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v
1  ( 1 )  " " "*" " " " " " "  " "  " " " "  " "  " "
2  ( 1 )  " " "*" " " " " " "  " "  " " " "  " "  "*"
3  ( 1 )  " " "*" " " "*" " "  " "  " " " "  " "  "*"
4  ( 1 )  " " "*" " " "*" " "  " "  "*" " "  " "  "*"
5  ( 1 )  " " "*" " " "*" " "  " "  "*" " "  " "  "*"
6  ( 1 )  " " "*" " " "*" " "  " "  "*" " "  "*"  "*"
7  ( 1 )  " " "*" " " "*" " "  " "  "*" " "  "*"  "*"
8  ( 1 )  " " "*" " " "*" " "  " "  "*" " "  "*"  "*"
9  ( 1 )  " " "*" " " "*" "*"  " "  "*" " "  "*"  "*"
10  ( 1 ) " " "*" "*" "*" "*"  " "  "*" " "  "*"  "*"
11  ( 1 ) " " "*" "*" "*" "*"  " "  "*" "*"  "*"  "*"
12  ( 1 ) " " "*" "*" "*" "*"  " "  "*" "*"  "*"  "*"
13  ( 1 ) "*" "*" "*" "*" "*"  " "  "*" "*"  "*"  "*"
14  ( 1 ) "*" "*" "*" "*" "*"  "*"  "*" "*"  "*"  "*"
          ventNord ventOuest ventSud pluieSec
1  ( 1 )  " "      " "       " "     " "
2  ( 1 )  " "      " "       " "     " "
3  ( 1 )  " "      " "       " "     " "
4  ( 1 )  " "      " "       " "     " "
5  ( 1 )  " "      " "       " "     "*"
6  ( 1 )  " "      " "       " "     "*"
7  ( 1 )  "*"      " "       " "     "*"
8  ( 1 )  " "      "*"       "*"     "*"
9  ( 1 )  " "      "*"       "*"     "*"
10  ( 1 ) " "      "*"       "*"     "*"
11  ( 1 ) " "      "*"       "*"     "*"
12  ( 1 ) "*"      "*"       "*"     "*"
13  ( 1 ) "*"      "*"       "*"     "*"
14  ( 1 ) "*"      "*"       "*"     "*"
```

3. Sélectionner le meilleur modèle au sens du R^2. Que remarquez-vous ?
4. Faire de même pour le C_p et le BIC. Que remarquez-vous pour les variables explicatives qualitatives ?
5. Comparer cette méthode avec des modèles sélectionnées par la fonction `step` ou la fonction `bestglm` du package `bestglm`.

2.2 Régression sur composantes principales (méthodo)

On considère le jeu de données **Hitters** dans lequel on souhaite expliquer la variable **Salary** par les autres variables du jeu de données. Pour simplifier le problème, on supprime les individus qui possèdent des données manquantes (il ne faut pas faire ça normalement !).

```
library(ISLR)
Hitters <- na.omit(Hitters)
```

1. Parmi les variables explicatives, certaines sont qualitatives. Expliquer comment, à l'aide de la fonction **model.matrix** on peut utiliser ces variables dans un modèle linéaire. On appellera **X** la matrice des variables explicatives construites avec cette variable.
2. Calculer la matrice **Xcr** qui correspond à la matrice **X** centrée réduite. On pourra utiliser la fonction `scale`.
3. A l'aide de la fonction `PCA` du package **FactoMineR**, effectuer l'ACP du tableau **Xcr** avec l'option `scale.unit=FALSE`.
4. Récupérer les coordonnées des individus sur les 5 premiers axes de l'ACP (variables Z dans le cours).

5. Effectuer la régression linéaire sur les 5 premières composantes principales et calculer les estimateurs des MCO ($\widehat{\theta}_k, k = 1, \ldots, 5$ dans le cours).
6. En déduire les estimateurs dans l'espace des données initiales pour les données centrées réduites, puis pour les données brutes. On pourra récupérer les vecteurs propres de l'ACP (u_k dans le cours) dans la sortie **svd** de la fonction **PCA**
7. Retrouver les estimateurs dans l'espace des données initiales pour les données centrées réduites à l'aide de la fonction `pcr` du package **pls**.
8. On considère les individus suivants

```
df.new <- Hitters[c(1,100,80),]
```

Calculer de 3 façons différentes les valeurs de salaire prédites par la régression sur 5 composantes principales.

2.3 Régression PLS : méthodo

On considère les mêmes données que précédemment.

1. A l'aide du vecteur Y (*Salary*) et de la matrice des X centrées réduites calculées dans l'exercice précédent, calculer la première composante **PLS** Z_1.
2. En déduire le coefficient associé à cette première composante en considérant le modèle

$$Y = \alpha_1 Z_1 + \varepsilon.$$

3. En déduire les coefficients en fonction des variables initiales (centrées réduites) de la régression PLS à une composante

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

4. Retrouver ces coefficients en utilisant la fonction `plsr`.

2.4 Comparaison : PCR vs PLS.

1. Séparer le jeu de données (`Hitters` toujours) en un échantillon d'apprentissage de taille 200 et un échantillon test de taille 63.
2. Avec les données d'apprentissage uniquement construire les régressions PCR et PLS. On choisira les nombres de composantes par validation croisée.
3. Comparer les deux méthodes en utilisant l'échantillon de validation. On pourra également utiliser un modèle linéaire classique.
4. Comparer ces méthodes à l'aide d'une validation croisée 10 blocs.

3 Régressions pénalisées (ou sous contraintes)

Nous considérons toujours le modèle linéaire

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \varepsilon$$

Lorsque d est grand ou que les variables sont linéairement dépendantes, les estimateurs des moindres carrées peuvent être mis en défaut. Les méthodes pénalisées ou sous contraintes consistent alors à restreindre l'espace sur lequel on minimise ce critère. On va alors chercher le vecteur β qui minimise

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 \quad \text{sous la contrainte} \quad \sum_{j=1}^{d} \beta_j^2 \leq t$$

ou de façon équivalente (dans le sens où il existe une équivalence entre t et λ)

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{d}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{d}\beta_j^2.$$

Les estimateurs obtenus sont les estimateurs **ridge**. Les estimateurs **lasso** s'obtiennent en remplaçant la contrainte ou la pénalité par une norme 1 ($\sum_{j=1}^{d}|\beta_j|$). Nous présentons dans cette partie les étapes principales qui permettent de faire ce type de régression avec **R**. Le package le plus souvent utilisé est `glmnet`.

3.1 Ridge et lasso avec glmnet

On considère le jeu de données `ozone.txt` où on cherche à expliquer la concentration maximale en ozone relevée sur une journée (variable `maxO3`) par d'autres variables essentiellement météorologiques.

```
ozone <- read.table("data/ozone.txt")
head(ozone)
         maxO3   T9  T12  T15 Ne9 Ne12 Ne15     Vx9    Vx12
20010601    87 15.6 18.5 18.4   4    4    8  0.6946 -1.7101
20010602    82 17.0 18.4 17.7   5    5    7 -4.3301 -4.0000
20010603    92 15.3 17.6 19.5   2    5    4  2.9544  1.8794
20010604   114 16.2 19.7 22.5   1    1    0  0.9848  0.3473
20010605    94 17.4 20.5 20.4   8    8    7 -0.5000 -2.9544
20010606    80 17.7 19.8 18.3   6    6    7 -5.6382 -5.0000
            Vx15 maxO3v  vent pluie
20010601 -0.6946     84  Nord   Sec
20010602 -3.0000     87  Nord   Sec
20010603  0.5209     82   Est   Sec
20010604 -0.1736     92  Nord   Sec
20010605 -4.3301    114 Ouest   Sec
20010606 -6.0000     94 Ouest Pluie
```

Contrairement à la plupart des autres package **R** qui permettent de faire de l'apprentissage, le package `glmnet` n'autorise pas l'utilisation de `formules` : il faut spécifier explicitement la matrice des X et le vecteur des Y. On peut obtenir la matrice des X et notamment le codage des variables qualitatives avec la fonction `model.matrix` :

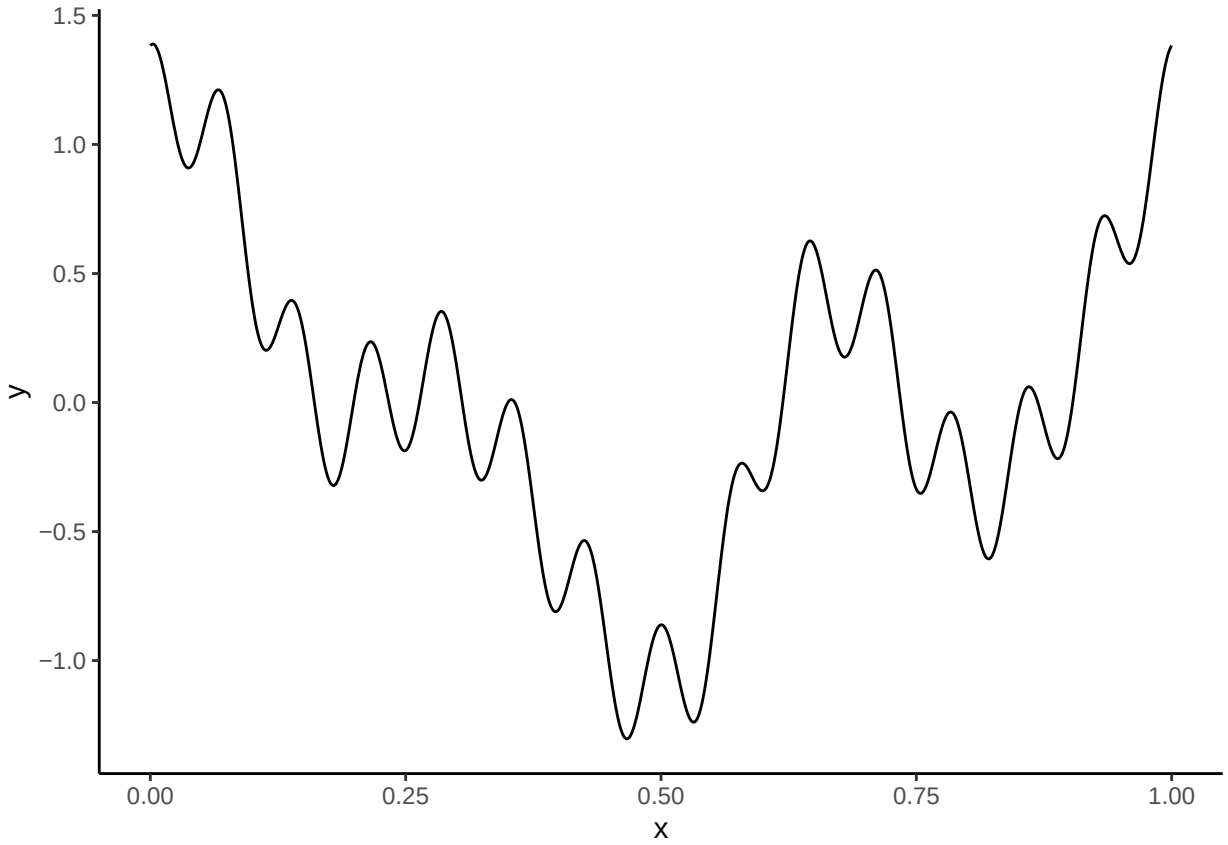
```
ozone.X <- model.matrix(maxO3~.,data=ozone)[,-1]
ozone.Y <- ozone$maxO3
```

1. Charger le package `glmnet` et à l'aide de la fonction `glmnet` calculer les estimateurs `ridge` et `lasso`.
2. Analyser les sorties qui se trouvent dans les arguments `lambda` et `beta` de `glmnet`.
3. Visualiser les chemins de régularisation des estimateurs `ridge` et `lasso`. On pourra utiliser la fonction `plot`.
4. Sélectionner les paramètres de régularisation à l'aide de la fonction `cv.glmnet`. On pourra notamment faire un `plot` de l'objet et expliquer le graphe obtenu.
5. On souhaite prédire la variable cible pour de nouveaux individus, par exemple les 25ème et 50ème individus du jeu de données. Calculer les valeurs prédites pour ces deux individus.
6. A l'aide d'une validation croisée, comparer les performances des estimateurs **MCO**, **ridge** et **lasso**. On pourra utiliser les données `ozone_complet.txt` qui contiennent plus d'individus et de variables.
   ```
   ozone1 <- read.table("data/ozone_complet.txt",sep=";") %>% na.omit()
   ozone1.X <- model.matrix(maxO3~.,data=ozone1)[,-1]
   ozone1.Y <- ozone1$maxO3
   ```
7. Refaire la question précédente en considérant toutes les interactions d'ordre 2.

3.2 Reconstruction d'un signal

Le fichier `signal.csv` contient un signal que l'on peut représenter par une fonction $m : \mathbb{R} \to \mathbb{R}$. On le visualise

```
signal <- read_csv("data/signal.csv")
ggplot(signal)+aes(x=x,y=y)+geom_line()
```

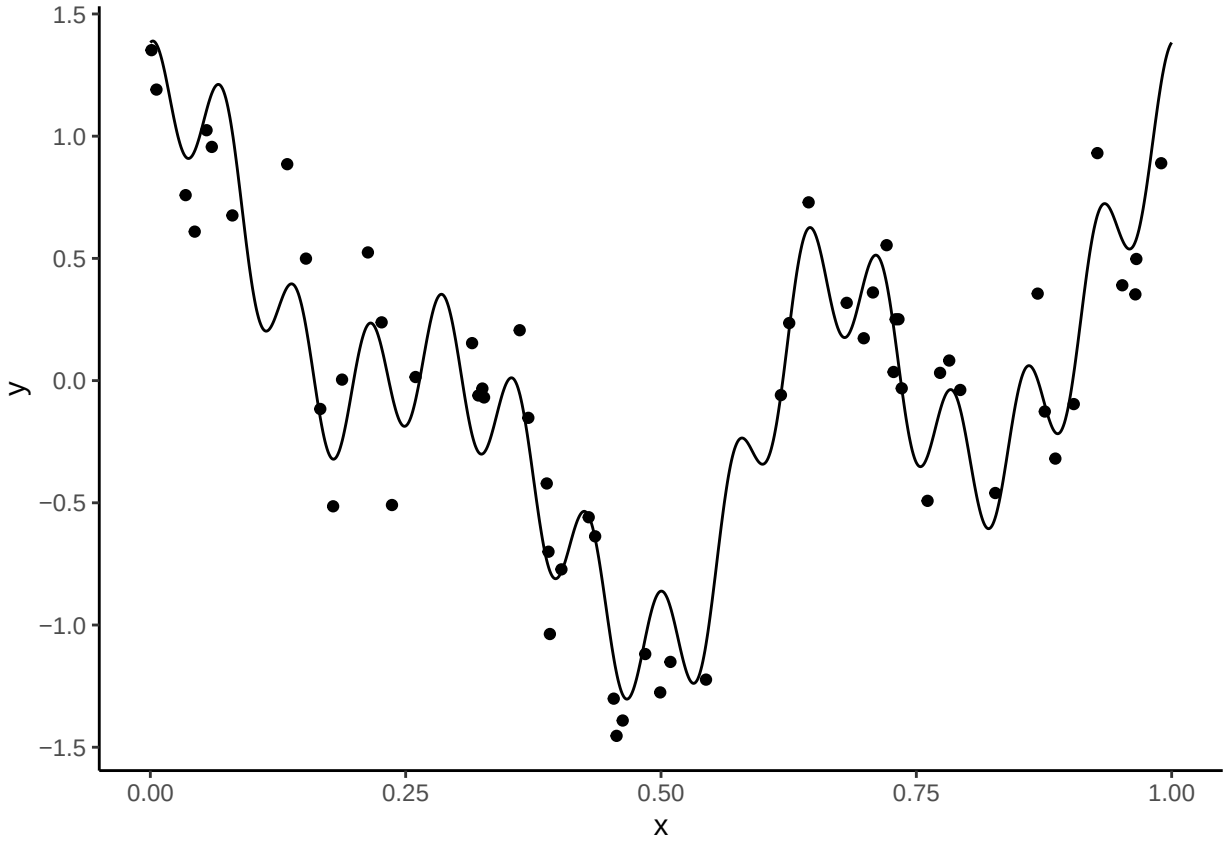


Plaçons nous dans le cas où on ne dispose que d'une version bruitée de ce signal. La courbe n'est pas observée mais on dispose d'un échantillon $(x_i, y_i), i = 1, \ldots, n$ généré selon le modèle

$$y_i = m(x_i) + \varepsilon_i.$$

Le fichier `ech_signal.csv` contient $n = 60$ observations issues de ce modèle. On représente les données et la courbe

```
donnees <- read_csv("data/ech_signal.csv")
ggplot(signal)+aes(x=x,y=y)+geom_line()+
  geom_point(data=donnees,aes(x=X,y=Y))
```



Nous cherchons dans cette partie à reconstruire le signal à partir de l'échantillon. Bien entendu, vu la forme du signal, un modèle linéaire de la forme

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

n'est pas approprié. De nombreuses approches en **traitement du signal** proposent d'utiliser une `base` ou `dictionnaire` représentée par une collection de fonctions $\{\psi_j(x)\}_{j=1,\dots,K}$ et de décomposer le signal dans cette base :

$$m(x) \approx \sum_{j=1}^{K} \beta_j \psi_j(x).$$

Pour un dictionnaire donné, on peut alors considérer un **modèle linéaire**

$$y_i = \sum_{j=1}^{K} \beta_j \psi_j(x) + \varepsilon_i. \tag{1}$$

Le problème est toujours d'estimer les paramètres β_j mais les variables sont maintenant définies par les élements du dictionnaire. Il existe différents types de dictionnaire, dans cet exercice nous proposons de considérer la base de Fourier définie par

$$\psi_0(x) = 1, \quad \psi_{2j-1}(x) = \cos(2j\pi x) \quad \text{et} \quad \psi_{2j}(x) = \sin(2j\pi x), \quad j = 1, \dots, K.$$

1. Écrire une fonction **R** qui admet en entrée :
   — une grille de valeurs de `x` (un vecteur)

— une valeur de `K` (un entier positif)

et qui renvoie en sortie une matrice qui contiennent les valeurs du dictionnaire pour chaque valeur de `x`. Cette matrice devra donc contenir `2K` colonnes et le nombre de lignes sera égal à la longueur du vecteur `x`.

2. On fixe `K=25`. Calculer les estimateurs des moindres carrés du modèle (1).
3. Représenter le signal estimé. Commenter le graphe.
4. Calculer les estimateurs **lasso** et représenter le signal issu de ces estimateurs.
5. Identifier les coefficients lasso sélectionnés qui ne sont pas nuls.
6. Ajouter les signaux ajustés par les algorithme PCR et PLS.

3.3 Régression logistique pénalisée

On considère le jeu de données sur la détection d'images publicitaires disponible ici https://archive.ics.uci.edu/ml/datasets/internet+advertisements.

```
ad.data <- read.table("data/ad_data.txt",header=FALSE,sep=",",dec=".",na.strings = "?",strip.white = TRU
names(ad.data)[ncol(ad.data)] <- "Y"
ad.data$Y <- as.factor(ad.data$Y)
```

La variable à expliquer est

```
summary(ad.data$Y)
   ad. nonad.
   459   2820
```

Cette variable est binaire. On considère une régression logistique pour expliquer cette variable. Le nombre de variables explicatives étant important, comparer les algorithmes du maximum de vraisemblance aux algorithmes de type **ridge/lasso** en faisant une validation croisée 10 blocs. On pourra utiliser comme critère de comparaison l'`erreur de classification`, la `courbe ROC` et l'`AUC`. Il faudra également prendre des décisions pertinentes vis-à-vis des données manquantes...

3.4 Exercices

Exercice 3.1 (Estimateurs ridge pour le modèle linéaire)**.**

On considère le modèle de régression

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

où les ε_i sont i.i.d de loi $\mathcal{N}(0, \sigma^2)$. Pour $\lambda \geq 0$, on note $\hat{\beta}_R(\lambda)$ l'estimateur ridge défini par

$$\hat{\beta}_R(\lambda) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

1. Exprimer $\hat{\beta}_R(\lambda)$ en fonction de $\mathbb{X}$, $\mathbb{Y}$ et λ.
2. Étudier le biais et la variance de $\hat{\beta}_R(\lambda)$ en fonction de λ. On pourra également faire la comparaison avec l'estimateur des MCO.
3. On suppose que la matrice $\mathbb{X}$ est orthogonale. Exprimer les estimateurs $\hat{\beta}_{R,j}(\lambda)$ en fonction des estimateurs des MCO $\hat{\beta}_j, j = 1, \ldots, p$. Interpréter.

Exercice 3.2 (Estimateurs lasso dans le cas orthogonal, voir Giraud (2015))**.**

On rappelle qu'une fonction $F : \mathbb{R}^n \to \mathbb{R}$ est convexe si $\forall x, y \in \mathbb{R}^n$, $\forall \lambda \in [0,1]$ on a

$$F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y).$$

On définit la sous-différentielle d'une fonction convexe F par

$$\partial F(x) = \{w \in \mathbb{R}^n : F(y) \geq F(x) + \langle w, y - x\rangle \text{ pour tout } y \in \mathbb{R}^n\}.$$

On admettra que les minima d'une fonction convexe $F : \mathbb{R}^n \to \mathbb{R}$ sont caractérisés par

$$x^\star \in \operatorname*{argmin}_{x \in \mathbb{R}^n} F(x) \Longleftrightarrow 0 \in \partial F(x^\star)$$

et que $\partial F(x) = \{\nabla F(x)\}$ lorsque F est différentiable en x.

1. Montrer que pour $x \in \mathbb{R}$

$$\partial |x| = \begin{cases} \text{signe}(x) & \text{si } x \neq 0 \\ [-1;1] & \text{sinon,} \end{cases}$$

où $\text{signe}(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x \leq 0}$.

2. Soit $x \in \mathbb{R}^n$.

   a. Montrer que

$$\partial \|x\|_1 = \{w \in \mathbb{R}^n : \langle w, x\rangle = \|x\|_1 \text{ et } \|w\|_\infty \leq 1\}.$$

   On pourra utiliser que pour tout p, q tels que $1/p + 1/q = 1$ on a

$$\|x\|_p = \sup\left\{\langle w, x\rangle : \|w\|_q \leq 1\right\}.$$

   b. En déduire

$$\partial \|x\|_1 = \{w \in \mathbb{R}^n : w_j = \text{signe}(x_j) \text{ si } x_j \neq 0, w_j \in [-1,1] \text{ si } x_j = 0\}.$$

3. Étant données n observations $(x_i, y_i), i = 1, \ldots, n$ telles que $x_i \in \mathbb{R}^p$ et $y_i \in \mathbb{R}$ on rappelle que l'estimateur lasso $\hat{\beta}(\lambda)$ est construit en minimisant

$$\mathcal{L}(\beta) = \|Y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_1. \tag{2}$$

   On admettra que la sous-différentielle $\partial \mathcal{L}(\beta)$ est donnée par

$$\partial \mathcal{L}(\beta) = \left\{-2\mathbb{X}^t(Y - \mathbb{X}\beta) + \lambda z : z \in \partial \|\beta\|_1\right\}.$$

   Montrer que $\hat{\beta}(\lambda)$ vérifie

$$\mathbb{X}^t \mathbb{X} \hat{\beta}(\lambda) = \mathbb{X}^t Y - \frac{\lambda}{2} \hat{z}$$

   où $\hat{z} \in \mathbb{R}^p$ vérifie

$$\hat{z}_j \begin{cases} = \text{signe}(\hat{\beta}_j(\lambda)) & \text{si } \hat{\beta}_j(\lambda) \neq 0 \\ \in [-1;1] & \text{sinon.} \end{cases}$$

4. On suppose maintenant que la matrice $\mathbb{X}$ est orthogonale.

   a. Montrer que

$$\text{signe}(\hat{\beta}_j(\lambda)) = \text{signe}(\mathbb{X}_j^t Y) \quad \text{lorsque } \hat{\beta}_j(\lambda) \neq 0$$

   et $\hat{\beta}_j(\lambda) = 0$ si et seulement si $|\mathbb{X}_j^t Y| \leq \lambda/2$.

   b. En déduire

$$\hat{\beta}_j(\lambda) = \mathbb{X}_j^t Y \left(1 - \frac{\lambda}{2|\mathbb{X}_j^t Y|}\right)_+, \quad j = 1, \ldots, p$$

   où $(x)_+ = \max(x, 0)$. Interpréter ce résultat.

Exercice 3.3 (Unicité de l'estimateur lasso, voir Giraud (2015))**.**

Soit $\hat{\beta}^1(\lambda)$ et $\hat{\beta}^2(\lambda)$ deux solutions qui minimisent (2). Soit $\hat{\beta} = (\hat{\beta}^1(\lambda) + \hat{\beta}^2(\lambda))/2$.

1. Montrer que si $\mathbb{X}\hat{\beta}^1(\lambda) \neq \mathbb{X}\hat{\beta}^2(\lambda)$ alors

$$\|\mathbb{Y} - \mathbb{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 < \frac{1}{2}\left(\|\mathbb{Y} - \mathbb{X}\hat{\beta}^1(\lambda)\|_2^2 + \lambda\|\hat{\beta}^1(\lambda)\|_1 + \|\mathbb{Y} - \mathbb{X}\hat{\beta}^2(\lambda)\|_2^2 + \lambda\|\hat{\beta}^2(\lambda)\|_1\right).$$

   On pourra utiliser la convexité (forte) de $x \mapsto \|x\|_2^2$.
2. En déduire que $\mathbb{X}\hat{\beta}^1(\lambda) = \mathbb{X}\hat{\beta}^2(\lambda)$.

4 Modèle additif

Le modèle additif (modèle GAM) peut être vu comme un compromis entre une modélisation linéaire et non paramétrique de la fonction de régression. Il suppose que cette fonction s'écrit

$$m(x) = m(x_1, \ldots, x_d) = \alpha + g_1(x_1) + \cdots + g_d(x_d).$$

4.1 Pseudo backfitting

L'algorithme du backfitting est souvent utilisé pour estimer les composantes du modèle additif. Etant donné un échantillon $(x_i, y_i), i = 1, \ldots, n$ on note $\bar{\mathbb{Y}}$ le vecteur des y_i et $\mathbb{X}_k$ le vecteur contenant les observations de la variable k pour $k = 1, \ldots, d$. L'algorithme se résume ainsi

1. Initialisation : $\widehat{\alpha} = \bar{\mathbb{Y}}$, $\widehat{g}_k(x_k) = \bar{\mathbb{X}}_k$.
2. Pour $k = 1, \ldots, d$:
   - $\mathbb{Y}^{(k)} = \mathbb{Y} - \widehat{\alpha} - \sum_{j \neq k} \widehat{g}_j(\mathbb{X}_j)$ (résidus partiels)
   - $\widehat{g}_k$: lissage non paramétrique de $\mathbb{Y}^{(k)}$ sur $\mathbb{X}_k$.
3. Répéter l'étape précédente tant que les $\widehat{g}_k$ changent.

On propose dans cette partie d'utiliser cet algorithme pour estimer les paramètres du modèle linéaire en remplaçant le lissage non paramétrique par un estimateur MCO. On considère le modèle de régression linéaire

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

avec X_1 et X_2 de lois uniformes sur $[0, 1]$ et ε de loi $\mathcal{N}(0, 1)$ (ε est indépendante de $(X_1, X_2)'$).

1. Générer un échantillon (x_i, y_i) de taille $n = 300$ selon le modèle ci-dessus pour $\beta_0 = 1, \beta_1 = 3, \beta_2 = 5$.
2. Créer une fonction **R** qui admet en entrée un jeu de données et qui fournit en sortie les estimateurs par la méthode du backfitting.
3. En déduire les estimateurs backfitting pour le problème considéré.
4. Comparer aux estimateurs **MCO**.

4.2 Modèle GAM

On considère les données générées selon

```
n <- 1000
set.seed(1465)
X1 <- 2*runif(n)
X2 <- 2*runif(n)
bruit <- rnorm(n)
Y <- 2*X1+sin(8*pi*X2)+bruit
donnees<-data.frame(Y,X1,X2)
```

1. Écrire le modèle
2. A l'aide du package **gam** visualiser les estimateurs des composantes additives du modèle. On utilisera tout d'abord un lissage par **spline** avec 1 ddl pour la première composante et 24.579 ddl pour la seconde.
3. Faire varier les degrés de liberté, interpréter.
4. Faire le même travail avec le lisseur **loess**. On commencera avec `degree=2` et `span=0.15` puis on fera varier le paramètre `span`.
5. Estimer le degrés de liberté avec la fonction `gam` du package **mgcv** (Il n'est pas nécessaire de charger le package pour éviter les conflits).

4.3 Régression logistique additive

On considère le jeu de données **panne.txt** qui recense des pannes de machine (`etat=1`) en fonction de leur âge et de leur marque.

1. Faire une régression logistique permettant d'expliquer la variable `etat` par la variable `age` uniquement. Critiquer le modèle.
2. Ajuster un modèle additif, toujours avec uniquement la variable `age`.
3. En utilisant le modèle additif, proposer un nouveau modèle logistique plus pertinent.

Références

Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. CRC Press.