Rapport du projet 3
Statistique en grande dimension

Nango Fofana
Ismail OKIEH OMAR
Université Gustave Eiffeil

25 Novembre 2020

Sujet : Modèle de suite gaussienne

Exercice 1: Différentes méthodes de seuillage.

L'objectif ici est d'étudier les différentes méthodes de seuillages avec les modèles de suite gaussienne. Pour cela supposons qu'on observe $\{y_1, ..., y_M\}$ vérifiant le modèle de suite gaussienne suivante:

$$y_j = a\beta_j + \eta_j.$$

avec $a \in \mathbf{R}$, $\beta_j \in \{0, 1\}$ qui sont des paramètres inconnus tels que:
$\sum_j^M \beta_j = M^{1-\alpha}$ pour un $\alpha \in]0, 1[$ fixé, et ou les variables aleatoires η_j sont i.i.d. de loi $\mathcal{N}(0, 1)$ M=50, a varie entre 1 et 10 et $\alpha = 0.3$.

1. Visualisons les données pour quelques valeurs différentes de a.

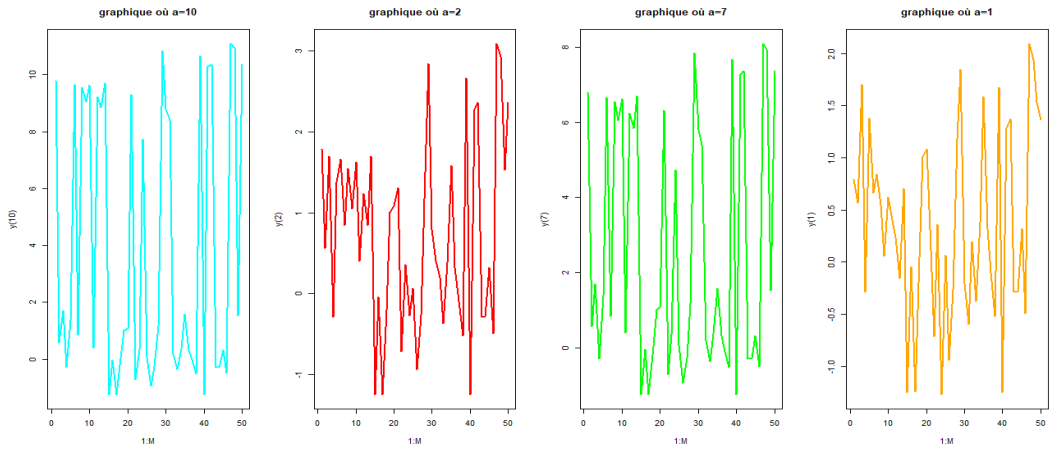
Pour la visualisation nous avons les le programme suivant:

```
M=50 #Données
alpha=0.3
s=sqrt(2*log(M)) ### tau
beta=rbinom(M,1,M^(-alpha))
eta=rnorm(M)
y=function(a){
  r=a*beta+eta
```

```
  return(r)
}
```

Pour les graphique nous avons utiliser quelque valeurs de a, le cas où $a = 10$, $a = 2$, $a = 7$,$a = 1$ On a:

```
dev.new()
par(mfrow=c(1,4))
plot(1:M,y(10),type = 'l',col="cyan",lwd=2,main = 'graphique où a=10')## pour a=10
plot(1:M,y(2),type = 'l',col="RED",lwd=2,main = 'graphique où a=2') ## pour a= 2
plot(1:M,y(7),type = 'l',col="green",lwd=2,main = 'graphique où a=7') ## pour a= 7
plot(1:M,y(1),type = 'l',col="ORANGE",lwd=2,main = 'graphique où a=1') ## pour a=1
```



Dans la suite nous voulons estimer $\theta^* = a.(\theta_1^*, ..., \theta_M^*)$. On pose $seuil = \sqrt{2\ln(M)}$ et on considère l'estimateur par seuillage fort $\hat{\theta}^H$, l'estimateur par seuillage faible $\hat{\theta}^S$ et l'estimateur dit 'non-negative garrotte' $\hat{\theta}^{NG}$ définis respectivement par:

$$\hat{\theta}_j^H = y_j \mathbf{1}_{\{|y_j|>\tau\}},\ \hat{\theta}_j^S = y_j.(1 - \frac{\tau}{|y_j|})_+,\ \hat{\theta}_j^{NG} = y_j.(1 - \frac{\tau^2}{y_j^2})_+$$

pour tout $j \in \{1, ..., M\}$, $(x)_+ = \max(0, x)$ et $\mathbf{1}_A$ désignant la fonction indicatrice de A

2- Traçons sur le même graphique les fonctions $\hat{\theta}_j^H$, $\hat{\theta}_j^S$, $\hat{\theta}_j^{NG}$ (comme fonction de y_j).

Pour pouvoir faire ces graphique nous allons utiliser le programme suivant:

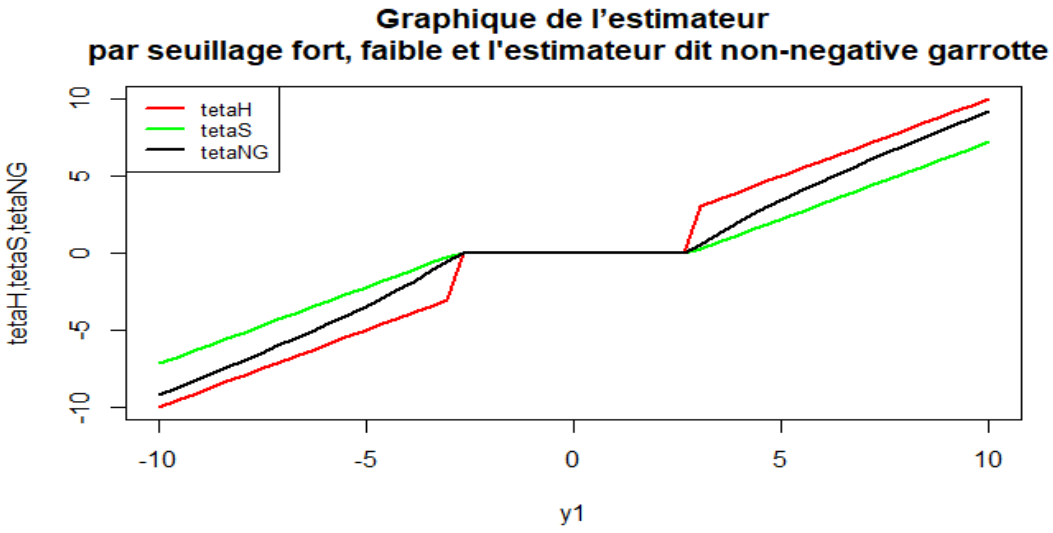
```
## theta chapeau H de l'estimateur à seuillage fort
s=sqrt(2*log(M)) ###le seuil tau
```

```
    y1=seq(-10,10,length.out = M) ## choix de la grille
    tetaH=function(y1){
      r=y1*(abs(y1)>s)
      return(r)
    }
    plot(y1,tetaH(y1),ylab= 'tetaH,tetaS,tetaNG', type = 'l',col='red',lwd=2,
    main="Graphique de l'estimateur par seuillage fort,
    faible et l'estimateur dit non-negative garrotte ") #graphique

## theta chapeau S de l'estimateur à seuillage faible
    tetaS=function(y1){
      r=y1*(1-(s/abs(y1)))*(abs(y1)>s)
      return(r)
    }
    points(y1,tetaS(y1),type = 'l',lwd=2, col='green',main="Graphique de l'estimateur
    par seuillage faible") #graphique

## theta chapeau NG de l'estimateur dit non-negative garrotte
    tetaNG=function(y1){
      r=y1*(1-(s^2/y1^2))*(y1^2>s^2)
      return(r)
    }
    points(y1,tetaNG(y1),type = 'l',lwd=2,
       main="Graphique de l'estimateur dit 'non-negative garrotte'") #graphique
    legend("topleft", legend=c("tetaH", "tetaS","tetaNG"),
       col=c("red","green", "black"), lwd=2, cex=0.8)
```

On a les graphiques suivant:

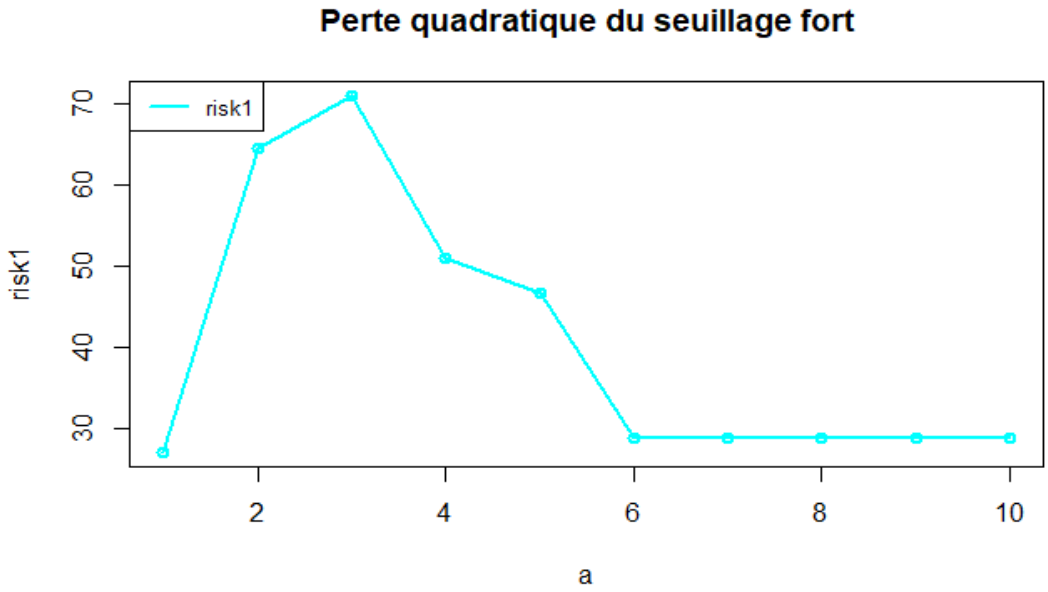


Pour la suite nous voulons étudier et représenter la perte quadratique de l'estimateur $\hat{\theta}$ définir par $R(\hat{\theta}, a) = |\hat{\theta} - \theta^*|_2^2 = \sum_{j=1}^{M} (\hat{\theta}_j - \theta_j^*)^2$

3- Représentons graphiquement, pour chacun des estimateurs ci-dessus, les quantités $R(\hat{\theta}, a)$ pour a qui varie de 1 à 10.

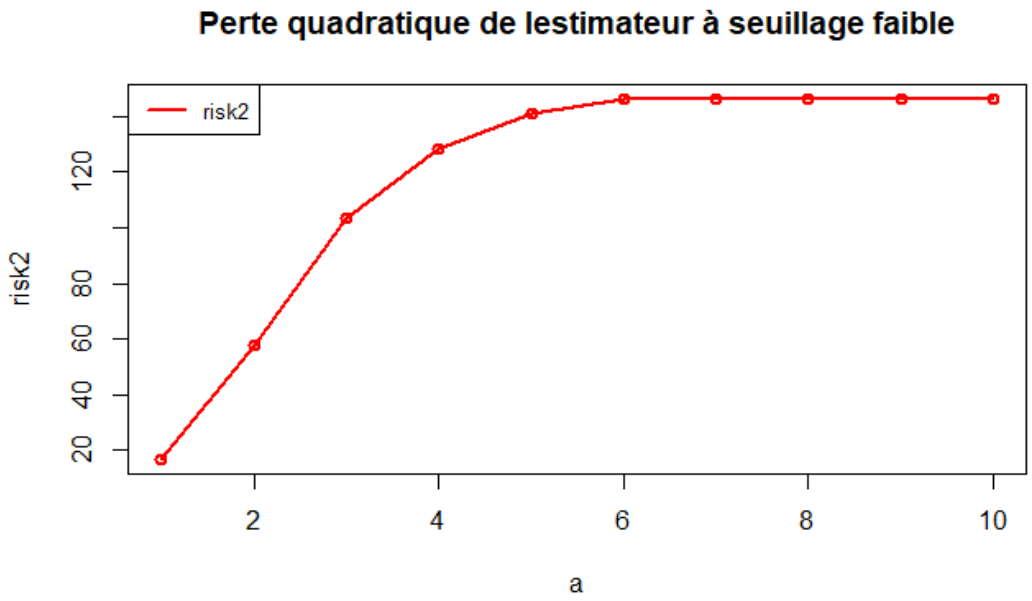
Pour l'étude de la perte quadratique nous allons étudier pour chaque estimateurs c'est à dire celui de l'estimateur à seuillage fort $\hat{\theta}^H$, lestimateur à seuillage faible $\hat{\theta}^S$ et enfin l'estimateur dit non-négative garrotte $\hat{\theta}^{NG}$ et ensuite les représenter. Pour le seuillage fort variant de 1 à 10 nous avons le programme suivant:

```
    #Perte quadratique de l'estimateur à seuillage fort
    R1=function(y,a){
      r=(y*(abs(y)>s)-a*(beta))^2
      return(sum(r))
    }
    a=1:10
    risk1=rep(0,10)
    for(i in a){
      risk1[i]=R1(y(i),i)
    }
## Graphique de l'estimateur à seuillage fort
    plot(a,risk1,type='o',lwd=2,col='cyan',
     main='Perte quadratique du seuillage fort')
    legend("topleft", legend=c("risk1"),
       col=c("cyan"), lwd=2, cex=0.8)
```



Ensuite pour l'estimateur à seuillage faible variant de 1 à 10 nous avons le programme suivant:

```
    ## Perte quadratique de l'estimateur à Seuillage faible
    R2=function(y,a){
      r=(y*(1-(s/abs(y)))*(abs(y)>s)-a*(beta))^2
      return(sum(r))
    }
    risk2=rep(0,10)
    for(i in a){
      risk2[i]=R2(y(i),i)
    }
## Graphique de l'estimateur à seuillage faible
    plot(a,risk2,type='o',col='red',lwd=2,
     main='Perte quadratique de lestimateur à seuillage faible')
    legend("topleft", legend=c("risk2"),
       col=c("red"), lwd=2, cex=0.8)
```

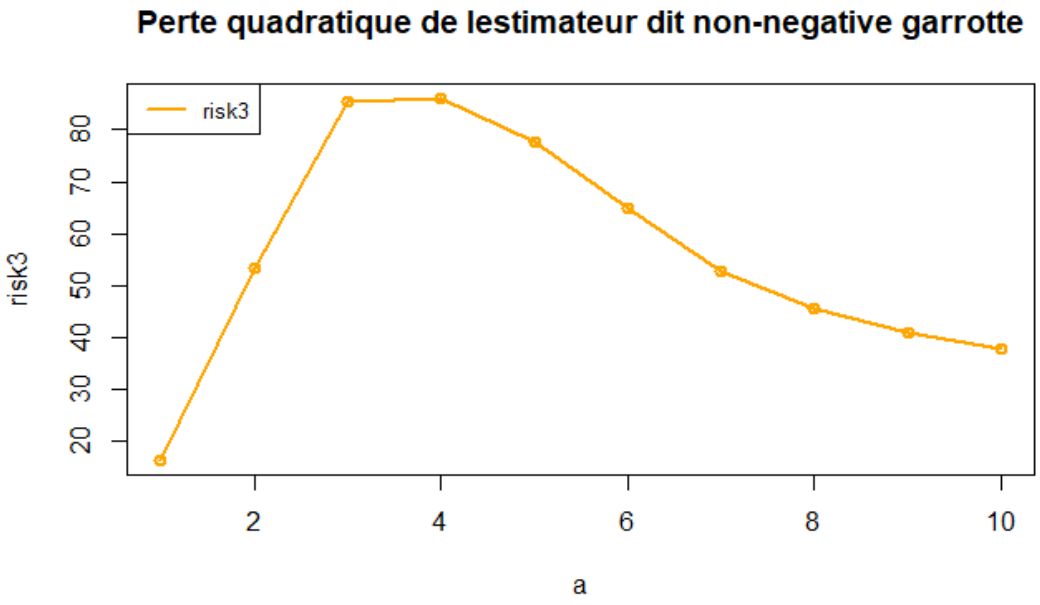


Pour l'estimateur dit non-négative garrotte nous avons le programme suivant:

```
    #### Perte quadratique de l'estimateur dit non-negative garrotte
    R3=function(y,a){
      r=(y*(1-(s^2/y^2))*(y^2>s^2)-a*(beta))^2
      return(sum(r))
    }
    risk3=rep(0,10)
```

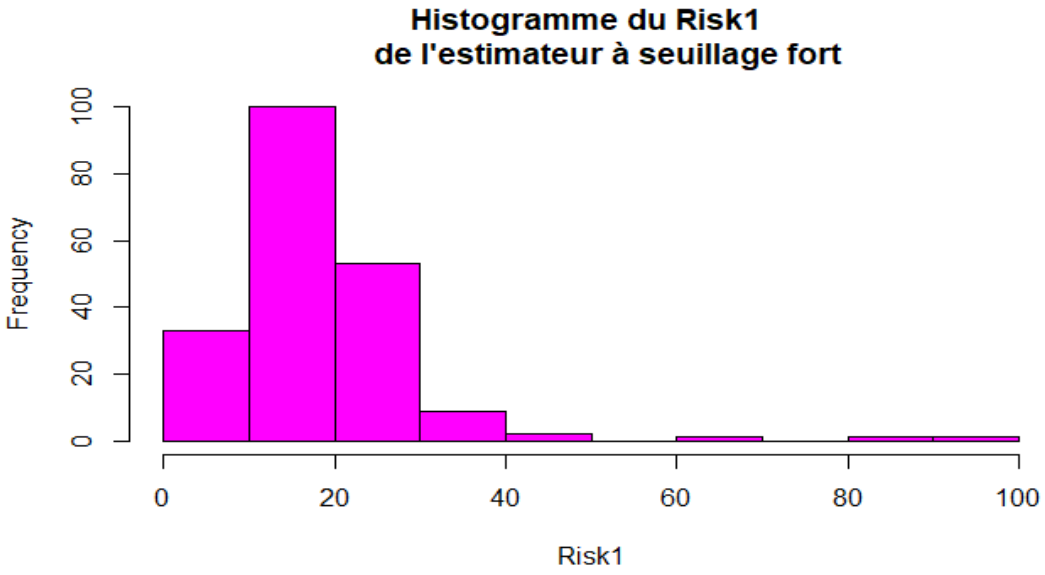
```
for(i in a){
  risk3[i]=R3(y(i),i)
}
plot(a,risk3,type='o',lwd=2,col='orange',
 main='Perte quadratique de lestimateur dit non-negative garrotte')

legend("topleft", legend=c("risk3"),
   col=c("orange"), lwd=2, cex=0.8)
```



Pour avoir une idée des vrais risques, utiliser pour les illustrations de le question 3 des moyennes sur plusieurs répétitions c'est à dire 200 répetitions pour chaque estimateur que nous allons représenté pars des histogrammes. Pour l'estimateur à seuillage fort nous avons le programme suivant:

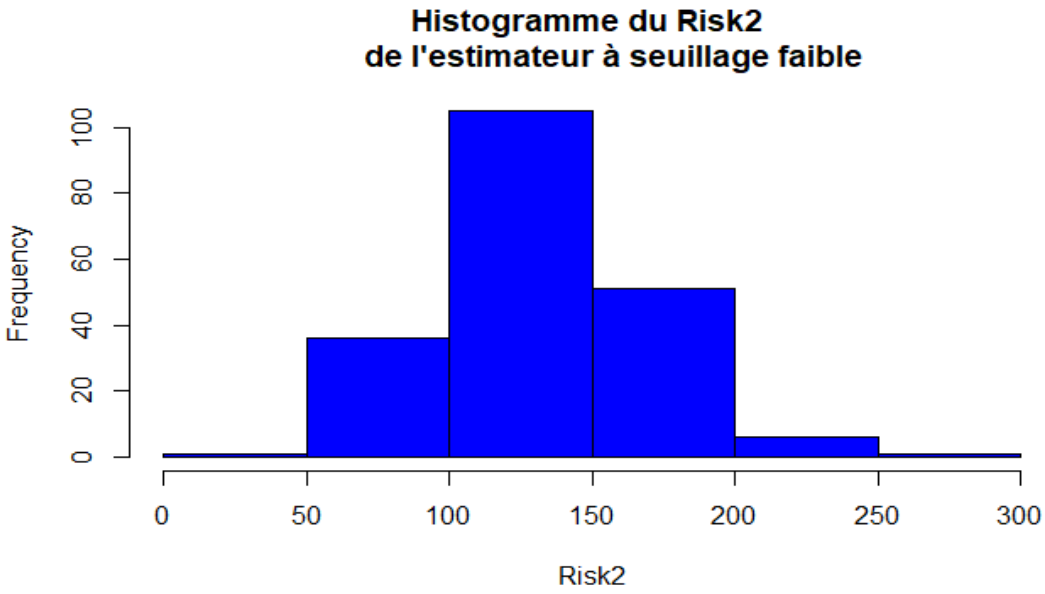
```
## Pour l'estimateur à seuillage fort
Risk1=rep(0,200)
a=1:10
for (i in(1:200)) {
  beta=rbinom(M,1,M^(-alpha))
  eta=rnorm(M)
  Risk1[i]=R1(y(i),i)
}
hist(Risk1,col='magenta',
 main = "Histogramme du Risk1
 de l'estimateur à seuillage fort ")  ## histogramme
```



Nous constatons au départ une croissance du risque jusqu'à atteindre un pique puis décroît au fur et à mesure que le nombre de répétition augmente et tend vers 0 voir même égale à 0 pour l'estimateur à seuillage fort.

Étudions maintenant pour l'esimateur à seuillage faible. On à le programme suivant:

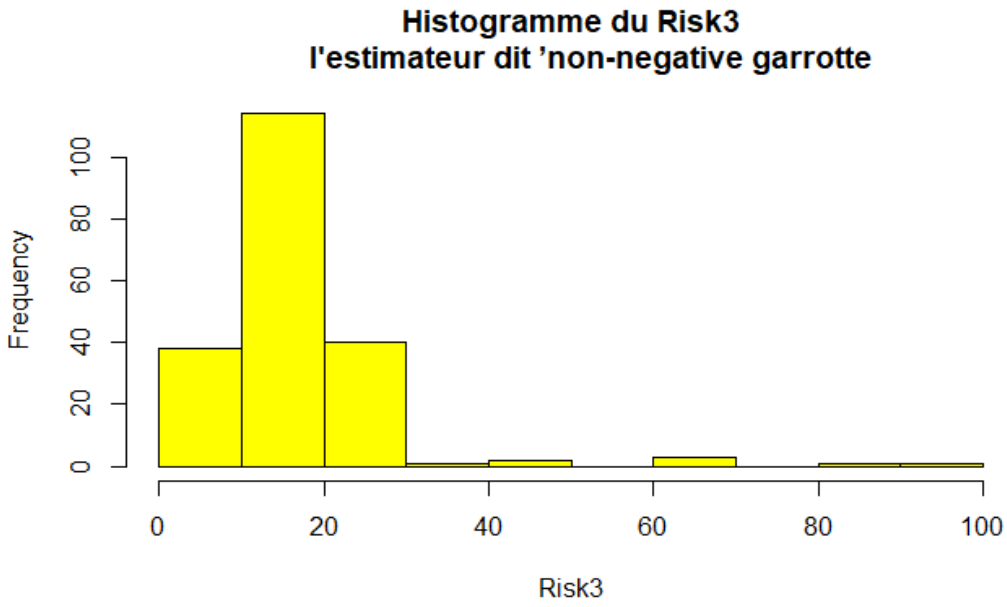
```
### Pour l'estimateur à seuillage faible répété 200 fois
Risk2=rep(0,200)
a=1:10
for (i in(1:200)) {
  beta=rbinom(M,1,M^(-alpha))
  eta=rnorm(M)
  Risk2[i]=R2(y(i),i)
}
hist(Risk2,col='blue',main = "Histogramme du
Risk2 de l'estimateur à seuillage faible") ## histogramme
```



Pour l'estimateur a seuillage faiblle ont constate que le risque est proche de 0 puis au fur et a mesure croît et atteint son pique puis décroît pour tendre vers 0.

Etudions pour le cas de l'estimateur dit non negative garotte. On à le programme suivant:

```
Risk3=rep(0,200)
a=1:10
for (i in(1:200)) {
  beta=rbinom(M,1,M^(-alpha))
  eta=rnorm(M)
  Risk3[i]=R3(y(i),i)
}
hist(Risk3,col='yellow',
main = "Histogramme du Risk3
l'estimateur dit 'non-negative garrotte") ## histogramme
```



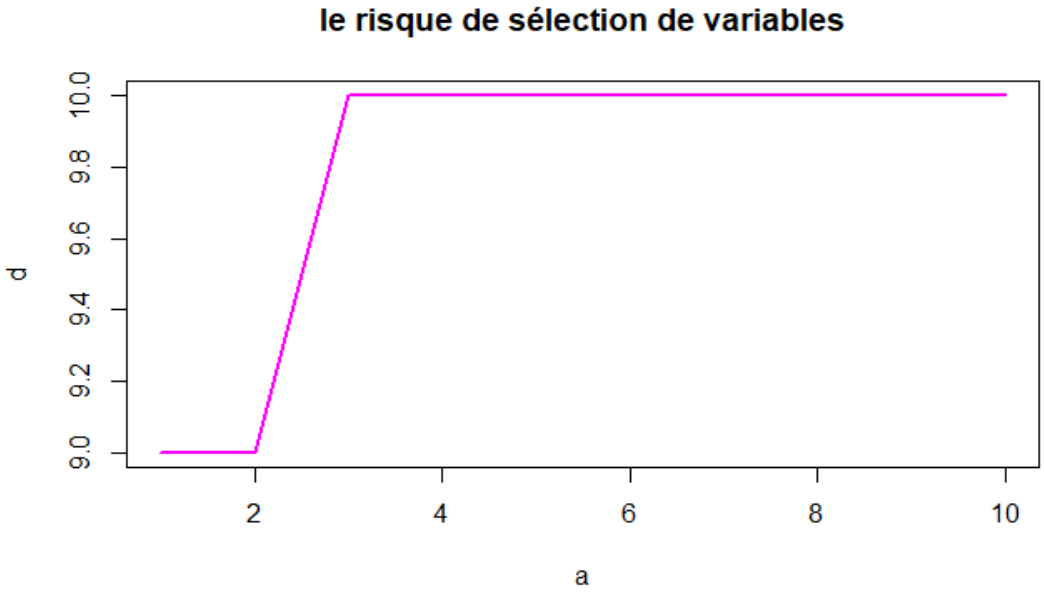
Au debut nous avons une croissance puis une fois le pique atteint nous avons une décroissance jusqu'à tendre vers 0

4- Sélectionnons les coordonnées non nulles de θ^* (par seuillage dur, c'est-à-dire, $\hat{\beta}_j = \mathbf{1}_{\{|y_j|>\tau\}}$) et étudions le risque de sélection de variables $\sum_{j=1}^{M} |\beta_j - \hat{\beta}_j|$ pour $a \in [0,1]$

On à le programme suivant:

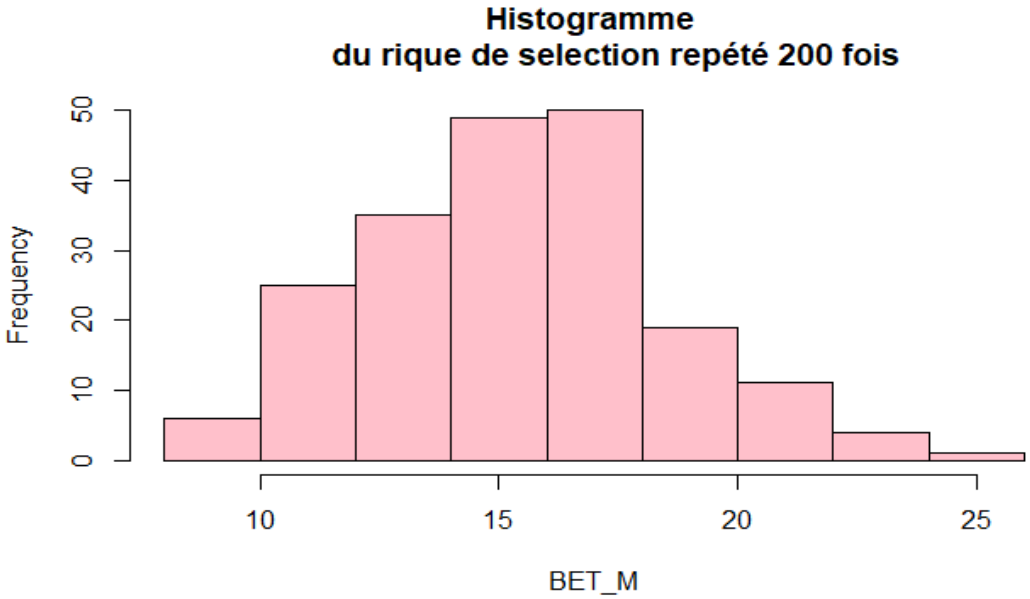
```
beta_chap=function(y){
L=rep(0,50)
for (i in 50){
  if(abs(y)>=s){L[i]=1}
  else
    L[i]=0
 }
 return(L)
}
bet=function(y){
r=abs(beta-beta_chap(y))
return(sum(r))
}
a=1:10
d=rep(0,10)
for (i in a) {
  d[i]=bet(i)
```

```
}
plot(a,d,type = 'l',lwd=2,col='magenta',
   main="le risque de sélection de variables")
```



Pour avoir une idée des vrais risques, utiliser pour les illustrations de la question 4 des moyennes sur plusieurs répétitions On a le programme suivant:

```
## Faisons une répétition de 200 FOIS
BET_M=rep(0,200)
a=1:10
for (i in(1:200)) {
  beta=rbinom(M,1,M^(-alpha))
  eta=rnorm(M)
  BET_M[i]=bet(y(i))
}
hist(BET_M, col="pink",main = "Histogramme
 du risque de sélection repété 200 fois")
```



Nous constatons nous avons une croissance de notre risque de sélection de variable qui une fois atteint si=on pique décroît tend vers zéro voir même nul.

Exercice 2: Détection de ruptures

Dans cette partie nous cherchons a connaître le point de de rupture c'est à dire a partie de quelle niveau où l'estimateur $\hat{\Delta}$ ne s'annule pas. Supposons que l'on observe $y_1, ..., y_M$ vérifiant le modèle de suite gaussienne.

$$y_j = \theta_j^* - \epsilon\eta_j$$

avec $\theta^* = (\theta_1^*, ..., \theta_M^*)^T$ est tel que:

$$\begin{cases} \theta_j^* = 3 & \text{pour } \ j \in \{1, .., 10\} \\ \theta_j^* = 7 & \text{pour } \ j \in \{11, ..., 30\} \\ \theta_j^* = 1.5 & \text{pour } \ j \in \{31, ..., 40\} \\ \theta_j^* = 2 & \text{pour } \ j \in \{41, ..., 50\} \end{cases}$$

et les variables aléatoires η_j sont i.i.d. de loi $\mathcal{N}(0, 1)$.

Pour M=50 et $\epsilon = 0.15$ proposons une méthode permettant, à partir des observations y_j , de détecter les instants de ruptures dans le vecteur θ^* inconnu, c'est-à-dire, une méthode permettant d'estimer l'ensemble:
$J^* = \{j \in \{2, ..., M\} : \theta_j^* \neq \theta_{j-1}^*\}$.

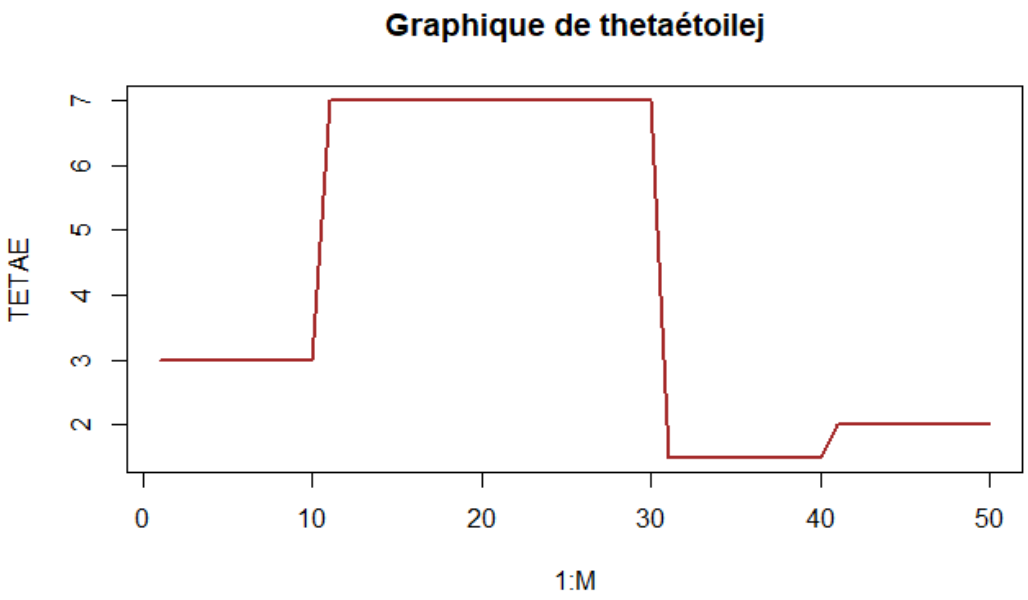
Nous allons considérer le vecteur des différences $\Delta^* = (\Delta_1^*, ..., \Delta_M^*)^T$ dont la j-ième composante Δ_j^* est donnée par : $\theta_{j+1}^* - \theta_j^*$ et observer qu'il est parcimonieux.

Essayons de Construire θ_j^* pour $j \in \{1, ..., M\}$. On a le programme suivant:

```
M=50
eta=rnorm(M)
eps=0.15 ## espsilon
SEUIL=sqrt(2*log(M-1))*sqrt(2*eps^2)

############Teta etoile de l'énoncer
TETAE=rep(0,M) ## initialisation de theta étoile
for (j in (1:M)) {
  if(j<=10 ){TETAE[j]=3}
  if(j>10 & j<=30 ){TETAE[j]=7}
  if(j>30 & j<=40 ){TETAE[j]=1.5}
  if(j>40 & j<=M  ){TETAE[j]=2}
}
plot(1:M, TETAE, type = 'l',col='brown',lwd=2,
 main = "Graphique de thetaétoilej")
```

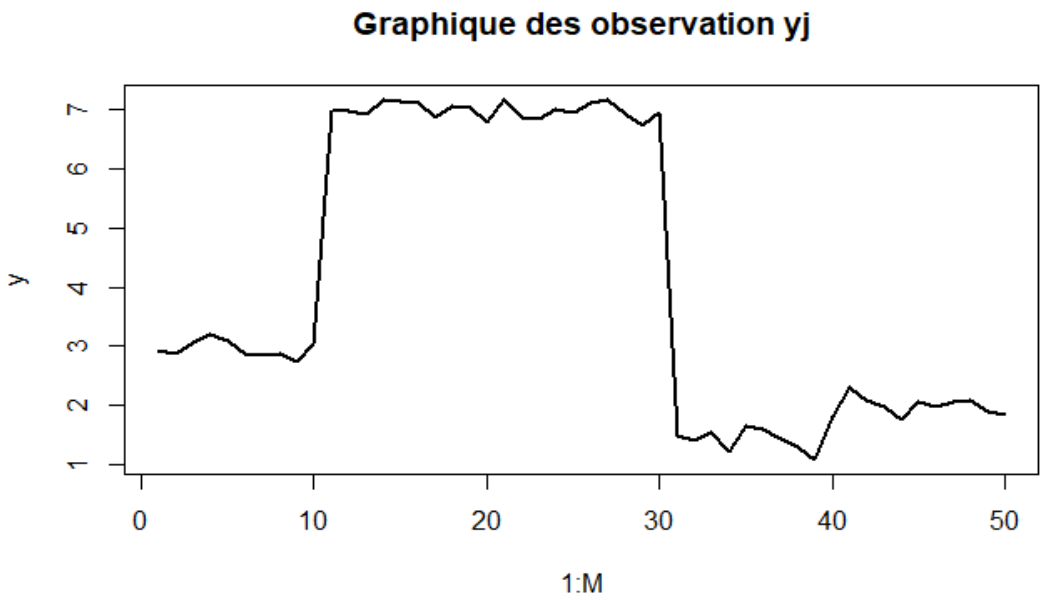
On a le graphique suivant:



Représentons les observations de $y_1, ..., y_M$ vérifiant le modèle de suite gaussienne de l'énonce on a :

```
y=TETAE+eps*eta
```

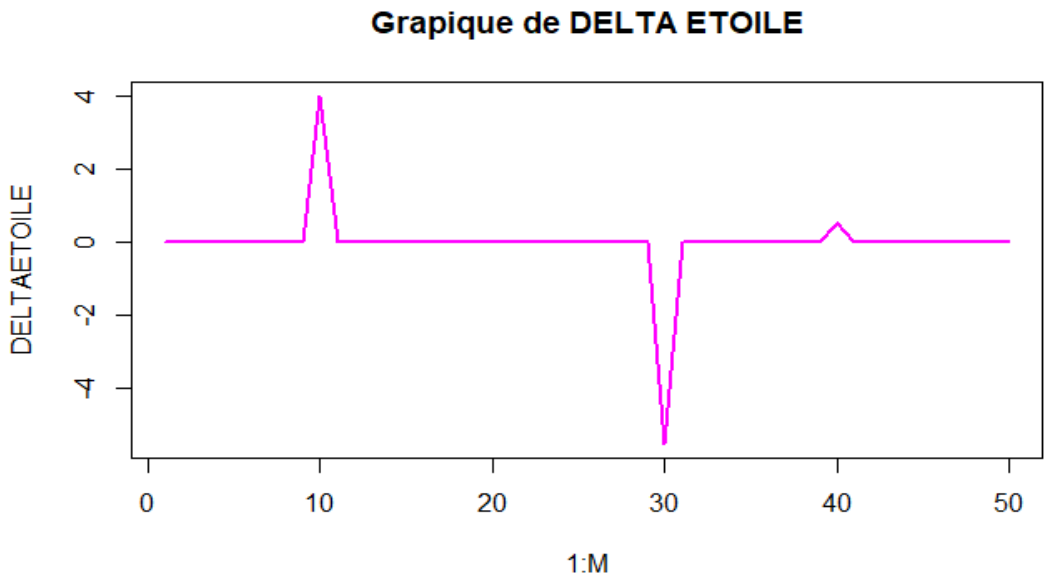
```
y
plot(1:M, y, type='l',lwd=2, main = "Graphique des observation yj")
```



Cherchons le vecteur des différences $\Delta_j^* = \theta_{j+1} - \theta_j^*$. On a le programme suivant:

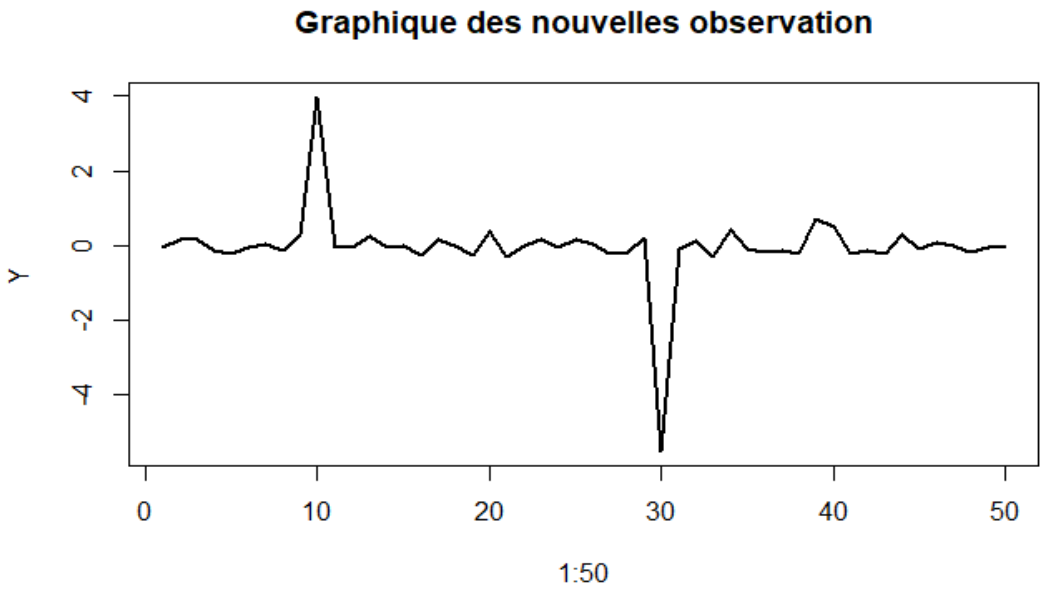
```
## Différence entre (eta(j+1)-eta(j))
Z=rep(0,M)
for (j in(1:M-1)) {
  Z[j]=eta[j+1]-eta[j]
}
Z
##Delta étoile (TETAE(j+1)-TETAE(j))
DELTAETOILE=rep(0,M)
for (j in(1:M-1)) {
  DELTAETOILE[j]=TETAE[j+1]-TETAE[j]
}
DELTAETOILE
plot(1:M, DELTAETOILE, type='l',lwd=2,col='magenta',
 main = "Grapique de DELTA ETOILE")
```

Graphique de Δ^* on a le programme suivant:



Cherchons les nouvelles observation qui sont $Y_j = y_{j+1} - y_j$ Pour les nouvelles observations nous avons le programme suivant:

```
### Les nouvelles observations
Y=DELTAETOILE+eps*Z
Y
plot(1:50,Y,type='l',lwd=2,
 main='Graphique des nouvelles observation')
```



Utilisons le programme suivant pour definir $\hat{\Delta}$ et d'obtenir les instants de rupture des différents seuillages.

Pour l'estimateur à seuillage fort on a:

```
##Delta chapeau pour l'estimateur à seuillage fort
DELTACHAP_H=function(y){
   r=y*(abs(y)>SEUIL)
   return(r)
 }
## Les instants de rupture pour l'estimateur à seuillage fort
ins_rupture1=which(DELTACHAP_H(Y)!=0)
ins_rupture1
```

Pour l'estimateur à seuillage fort les instant de rupture sont $\{10, 30, 40\}$.

Pour l'estimateur à seuillage faible on a:

```
##Delta chapeau seuillage faible
DELTACHAP_S=function(y){
  r=y*(1-(SEUIL/abs(y)))*(abs(y)>SEUIL)
  return(r)
}
## Les instants de rupture pour l'estimateur à seuillage faible
ins_rupture2=which(DELTACHAP_S(Y)!=0)
ins_rupture2
```

Pour l'estimateur à seuillage faible les instant de rupture sont de même $\{10, 30, 40\}$

Pour l'estimateur dit non-négative Garrotte on a:

```
##Delta chapeau de l'estimateur dit 'non-négative garrotte
DELTACHAP_NG=function(y){
  r=y*(1-((SEUIL)^2/y^2))*(y^2>(SEUIL)^2)
  return(r)
}
## Les instants de rupture pour l'estimateur dit non-negative garrotte
ins_rupture3=which(DELTACHAP_NG(Y)!=0)
ins_rupture3
```

Pour l'estimateur dit non-négative garrotte les instant de rupture sont de même $\{10, 30, 40\}$

Exercice 3: Données réelles pour la détection de ruptures

Dans cette partie nous voulons appliqué la méthodes sur des jeux de données c'est à dire des données réelles. Pour ce faire nous avons utilisé les jeux de données de l'entreprise de construction et de réhabilitation des logement sociaux "CDC habitat" où l'étude est consacré sur la construction la réhabilitation et le financement des logement sociaux des logement sociaux. Pour notre observation que je vais appelé B est le prix de réhabilitation de chaque logement sociaux. On a le programme suivant qui nous permettra d'importer nos jeux de donné et d'appliquer la méthodes de l'exercice 2.

```
##Nous étudions les jeux de données de CDC Habitat concernant les logement sociaux
## importations et lecture des jeux de données
setwd("C:/Users/33758/Downloads") ## fonction permettant de localiser le jeux de donnée
getwd()
CDC=read.csv(file = "constructionrehabilitation_logementsocial_surface_prix.csv",
         header = TRUE, sep=";")
attach(CDC)
names(CDC)
str(CDC)
B=rehabilitation_prixderevient_logement ### Notre nouvelle observation
 N=length(B) ##taille des observations
```

Essayons d'appliquer la méthode de l'exercice 2 pour cela nous avons le programme suivant: Pour l'estimateur à seuillage fort on a:

```
C=var(B) ##variance de B
tt=sqrt(2*log(N))*sqrt(2*C) ## le nouveau seuil
#Nous cherchons les instants de Rupture de notre nouvelle observation

## Seuillage dur
H=function(X){
  r=X*(abs(X)>tt)
  return(r)
}
```

```
##Instants de rupture pour l'estimateur à seuillage fort
c=which(H(B)!=0)
c
```

Pour l'estimateur à seuillage fort nous avons un seul instant de rupture qui est 33.

Pour l'estimateur à seuillage faible on a:

```
## Seuillage doux
S=function(X){
  r=X*(1-(tt/abs(X)))*(abs(X)>tt)
  return(r)
}
##Instants de rupture du seuillage doux
cc=which(S(B)!=0)
cc
```

Pour l'estimateur à seuillage faible nous avons un seul instant de rupture qui est également 33.

Pour l'estimateur dit non-négative Garrotte on a:

```
## Estimateur dit non-negative garrotte
NG=function(X){
  r=X*(1-((tt)^2/X^2))*(X^2>(tt)^2)
  return(r)
}
##Instants de rupture de l'estimateur dit non-negative garrotte
ccc=which(NG(B)!=0)
ccc
```

Pour l'estimateur dit non-négative Garrotte nous avons un seul instant de rupture qui est également 33.

En conclusion nous constatons que quelque soit l'estimateur choisir nos instants de rupture ne change pas ils restent les mêmees