République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en "**Mathématiques Appliquées**"

Option :**Statistique**

Par

GHARBI Abla

Titre :

Sur l'estimation à noyau de la fonction de régression

Membres du Comité d'Examen :

Pr.	YAHIA Djabrane	Proffesseur	UMKB	Président
Dr.	KHEIREDDINE Souraya	M.C.B.	UMKB	Encadreur
Dr.	ZOUAOUI Nour El-Houda	M.C.B.	UMKB	Examinateur

Juin 2022

Dédicace

Je dédie ce modeste travail accompagné d'un profondeur amour à mes parents ma mère "Bouderhem Amel", mon père "Mohamed"

À mes fréres : Khaled et Omar, mes soeurs "Merzaka", "Meriem" "Salsabil", et ma grande mère "Bouderhem Fatna",

À mes copines "Hadjer", "Nadjoua", "Djihad", "Amel" et à chaque personne de ma famille

Remerciement

Je remercie Dieu le tout puissant de mon avoir donné la santé et la volonté de l'entamer et de terminer ce mémoire.

Je tiens à remercier mes parents qui m'ont toujours soutenu et encouragé ainsi qu'à tous les membres de ma famille.

Je remercie particulièrement mon encadreur Dr. Kheireddine Souraya, je la remercie pour ces efforts et sa patience, sa rigueur et sa disponibilités durant la réalisation de ce mémoire.

Mes vifs remerciements aux membres du Jury : Proffesseur Yahia Djabrane et Dr. Zouaoui Nour El-Houda pour l'intérêt qu'ils ont porté à mon mémoire en acceptant de l'examiner.

Je tiens à remercier toutes mes copines et mes collègues que j'ai rencontré durant mes années d'étude.

Mes remerciement vont égalemment a tous les membres du Département de Mathématiques : Enseignants, Etudiants et Administrateurs.

Résumé

Ce mémoire porte sur l'estimateur non paramétrique de la fonction de régression on utilisant la méthode à noyau. Nous rappelons les propriétés asymptotiques de l'estimateur, aussi du choix de noyau K et de paramètre de lissage h. Finalement, nous donnons des explications graphiques des résultats théoriques appliqués sur des exemples de régression linéaire et non linéaire à l'aide du logiciel **R.**

Notations et symbols

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$(X_1, ..., X_{n)}$	:	échantillon de taille n de v.a's.
$h = h_n$	:	Parmètre de lissage ou Fenêtre
h_{opt}	:	Fenêtre optimale
$K(.)$	:	Noyau
$\mathcal{A}$	:	Tribu
$\mathbb{E}$	:	Espace des observations
$\mathcal{B}$	:	Tribu borélienne
P	:	une famille de probabilité
Θ	:	Ensemble des paramètres
L^1	:	Espace des fonctions intégrables distribuées
iid	:	Indépendantes et identiquement distribuées
E	:	Espérance de probabilité
$Biais$	:	Biais d'un estimateur
var	:	Variance d'un estimateur

f_X	:	Densité de X
F	:	Fonction de répartition
F^{-1}	:	Fonction des quantiles
$f_{n,X}$	:	Estimateur de f
$v.a$	:	Variable aléatoire
m	:	Fonction de régression
m_n	:	Estimateur de m
$1_{(.)}$	:	Fonction indécatrice
MSE	:	L'erreur quadratique moyenne (mean squar error).
$AMSE$	:	L'erreur quadratique moyenne asymptotique (asymptotic mean squar error).
$MISE$	:	L'erreurquadratiqueintégréemoyenne(mean integrated squared error)
$AMISE$	:	L'erreur quadratique integrée moyenne asympotique (asymptotic mean integrated squared error)

Table des matières

Table des figures

Liste des tableaux

Introduction

La statistique non paramétrique est un domaine de la statistique qui ne repose pas sur des familles de loi de probabilité paramétriques. Les méthodes non paramétriques pour la régression comprennent les histogrammes, les méthodes d'estimation par noyau

Depuis les travaux de Rosenblatt (1956) et Parzen (1962) puis de Nadaraya (1964) et Watson (1964) portant respectivement sur les estimateurs non paramétrique des fonctions de la densité et de la régression. La méthode du noyau a été largement utilisée dans de nombreux travaux. L'estimation de la fonction de régression est un problème important dans l'analyse des données avec un large gamme d'applications en filtrage et la prévision dans les communications et le contrôle des systèmes, la reconnaissance de formes et de classification. L'objet de cette mémoire est l'étude d'estimateur non paramétriques de fonction de régression par la méthode de noyau.

L'estimateur de type noyau de la régression à été largement étudié dans la littérature. Les résultats originaux par Nadaraya (1964) et Watson (1964) ont été étendues dans plusieurs journaux, et elles sont résumées par exemple dans Bosq (1998), Devroye et Györfi (1985), et Rao (1983). Citons aussi le cas de données censurées à droites : Carbonez et al. (1995), Kohler et al. (2002) et autres et le cas de données tranquées à gauche : Lemdani et Ould-Saïd (2006),...

Ce mémoire est composé de trois chapitres qui sont :

Premier chapitre :(Estimation fonctionnelle) dans ce chapitre, nous présenterons la défnition de l'estimation fonctionnelle et l'estimateur à noyau de la densité (estimateur de Parzen (1962)-Rosenblatt (1956))

Deuxième chapitre : (Estimation non paramétrique de la fonction de régression). Nous présentons, l'estimateur non paramétrique de la fonction de régression (estimateur de Nadaraya (1964)-Watson (1964)). Nous présontons aussi, les propriétés de l'estimateur et l'étude asymtotique du biais et de la variance de l'estimateur, le choix de paramètre de lissage et le noyau

Troisième chapitre : (Application sous R) dans cette partie de ce mémoire nous allons vérifiés les résultats théoriques des chapitres précédents par simulation (avec le logiciel de traitement statistique R) sur des exemples, qui expriment l'importance du paramètre de lissage h, la taille de l'échantillon utilisée et le choix du noyau K dans l'estimation à noyau de la fonction de régression.

Chapitre 1

Estimation fonctionnelle

Dans ce chapitre, nous donnerons quelques notions définitions dans l'estimation paramétrique et non paramétrique

1.1 Estimation paramétrique et Estimation non paramétrique

Premièrement.nous appelons modéle statistique, le triplet $(\mathbb{E}, \mathcal{A}, \mathbb{P})$ où $\mathbb{E}$ est l'éspace des observations (par exemple des réels) $\mathcal{A}$ une tribu sur $\mathbb{E}$ et $\mathbb{P}$ une famille de probabilité sur $(\mathbb{E}, \mathcal{A})$:

soit $X : \Omega \to \mathbb{E}$ une application mesurable on peut toujours écrire $\mathbb{P}$ par $(\mathbb{P}_\theta, \theta \in \Theta)$ soit h une application de $\mathbb{P}$ dans Θ'. Estimer $h(p)$ c'est essayer de l'évaluer au vu de l'observation d'un échantillon de la variable aléatoire X qui est à valeurs dans $\mathbb{E}$. Donc, le paramétre à estimer est l'application

$$h : P \to \Theta' où\ \Theta \to \Theta'$$

$$\theta \to h(P_\theta)$$

une éstimateur de h est une fonction $h_n : x \to h_n(X_1, ..., X_n)$ mesurable par rapport à l'observation $(X_1, ..., X_n)$.

Définition 1.1.1 ***Estimation paramétrique :*** *si l'on sait à priori que h appartient à une famille paramétrée $\{h(x, \theta), \theta \in \Theta\} où\ \theta \in \mathbb{R}^s$ et $h(.,.)$ est une fonction connue on parle, alors d'estimation paramétrique ,car estimer h revient à estimer le paramétr fini dimensionnel Θ.*

Définition 1.1.2 ***Estimation non paramétrique :*** *par contre, si l'on sait seulement que h appartient à P ensemble des lois de probabilités qui est un éspace de dimension infinie, alors on dit que l'on fait de l'éstimation non paramétrique ou l'estimation fonctionnelle.*

Dans ce qui suit,on suppose que l'on a observé un échantillon $X_1, X_2, ..., X_n$à valeurs dans $\mathbb{R}^s$ muni de sa tribu borélienne $\mathcal{B}$ de plus on suppose que les $\{X_i, i = 1, ..., n\}$ sont indépendant et identiquement distribuées(i.i.d) $\mu \in P_0$une famille de loi sur $(\mathbb{R}^s; \mathcal{B})$.

i) **La densité de probabilité** :si P_0 est une famille de loi dominée par une λ, donc elle admet (théorme de Radon Nykadim) une densité $f = \frac{d\mu}{d\lambda}$ c'est un paramétre dans $L^1.si\ \frac{d\mu}{d\lambda}$ admet une version bornée (respectivement continue et bornée), alors on peut la considerer comme un paramétre dans L^2(respectivement dans $C_b(\mathbb{R}^s)$)., enfin si $f_\mu$est différentiable on défini de nouveaux paramétres fonctionnels :les dérivées partielles de f_μ.

ii) **La fonction de répartition** : c'est la fonction définie par

$$F_\mu(x_1,...,x_s)=\mu(\prod_{i=1}^{s}]-\infty;x_i]),(x_1,...,x_s)\in\mathbb{R}^s$$

iii) **La fonction des quantiles :** Pour $s=1$,la fonction quantile d'ordre p définie par

$$F_\mu^{-1}(p)=Q(p)=\inf\{t\in\mathbb{R};F_\mu(t)\geq p\},0<p<1$$

F_μ^{-1} est un paramètre à valeur dans l'éspace de fonctions réelles définies sur $]0;1[$ monotones non décroissantes et continues à gauche.

v) **La fonction caractéristique :** Elle est défini par

$$\widehat{\mu}(t)=E_\mu[\exp\{i\left\langle t,x\right\rangle\}]o\grave{u}\quad t,x\in\mathbb{R}^s$$

est un paramétre dans $C_b(\mathbb{R}^s)$.

iv) **Le paramétre de régression :** supposons que l'on observe un échantillon$\{(X_i,Y_i),i=1,...,n\}$ d'un couple (X,Y) à valeurs dans $\mathbb{R}^{s_1}\times\mathbb{R}^{s_2}$ est soit μ_y^x, $x\in\mathbb{R}^s$ une famille de verions des lois conditionnelles de Y sachant $X=x$: toute fonction de la forme $r:x\longrightarrow r\left(\mu_y^x\right)$ est une paramètre de régression les plus usuals sont :

1) l'espérance conditionnelle(qui est la fonction de régression),

2) la densité conditionnelle,

3) le mode conditionnel,

4) la fonction de répartition conditionnelle,

5) le quantile conditionnel.

1.2 Estimation non parmétrique d'une densité

Soit $X_1, ..., X_n$ i.i.d de densité f et de fonction de répartition F

si f continue en X (ce qui va être le cas pour les classes de fonctions qu'on va considérer), alors

$$f(x) = F'(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}$$

l'idée est donc d'utiliser l'approximation suivante pour h petit

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

pour estimer la densité f on peut donc passer par un estimateur F_n de la cdf F.voyons ce qui se passe si on choisit comme estimateur la fonction de répartition empirique F_n.(on rappelle que $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x}$)on choisit un $h > 0$ petit pourque l'approximation ci-dessus soit valable,et on pose

$$f_{n,X}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1_{Xi\in]x-h,x+h]}$$

$$\begin{aligned} f_{n,X}(x) &= \frac{1}{2nh}\sum_{i=1}^{n} \{1_{\{X_i \leq x+h\}} - 1_{\{Xi \leq x-h\}}\} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} 1\{x-h \leq X_i \leq x+h\} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} 1_{\{-1 \leq \frac{X_i - x}{h} \leq 1\}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0(\frac{X_i - x}{h}) \end{aligned}$$

si on pose $K_0(x) = \frac{1}{2}1_{]-1,1]}(u)$, alors on a K_0 est est appelé le noyau de Rosenblatton(1956)

Définition 1.2.1 *soit $K : \mathbb{R} \to \mathbb{R}$ intégrable et tel que $\int K(y)dy = 1$,alors K est appelé noyau (kernel).*

Définition 1.2.2 *Un estimateur à noyau Parzen et Rosenblatt de la densité f est une fonction $\hat{f}$ définie par :*

$$f_{n,X}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Où $(h)_{n\geq 1}$ est une suite de réels positifs appelés paramètres de lissage ou largeur de la fenêtre, qui tend vers 0 quand n tend vers l'infini, et K est une densité de probabilité appelée noyau.

En littérature, beaucoup de noyaux sont cités, notamment :

Noyau rectangulaire	:	$K_1(u) = \frac{1}{2}\uparrow_{[-1,1]}(u)$
Noyau triangulaire	:	$K_2(u) = (1- \mid u \mid)\uparrow_{[-1,1]}(u)$
Noyau d'Epanechnikov	:	$K_3(u) = \frac{3}{4}(1-u^2)\uparrow_{[-1,1]}(u)$
Noyau Biweight	:	$K_4(u) = \frac{15}{16}(1-u^2)^2\uparrow_{[-1,1]}(u)$
Noyau Gaussien	:	$K_5(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$

TAB. 1.1 – Tableau des noyaux

La figure (Fig.1.1) si-après présente l'allure des cinq noyaux cités

Code R.

```
K1=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
K2=function(t){(15/16)*((1-t^2)^2)*ifelse(abs(t)<=1,1,0)}
K3=function(t){dnorm(t)}
K4=function(t){ifelse(abs(t)<1,(3/4)*(1-t^2),0)}
op=par(mfrow=c(2,2))
```

```
curve(K1(x),-1,1,ylab="K(x)",main="Triangulaire")
curve(K2(x),-1,1,ylab="K(x)",main="Biweight")
curve(K3(x),-4,4,ylab="K(x)",main="gaussien")
curve(K4(x),-1,1,ylab="K(x)",main="Epanechikov")
par(op)
```

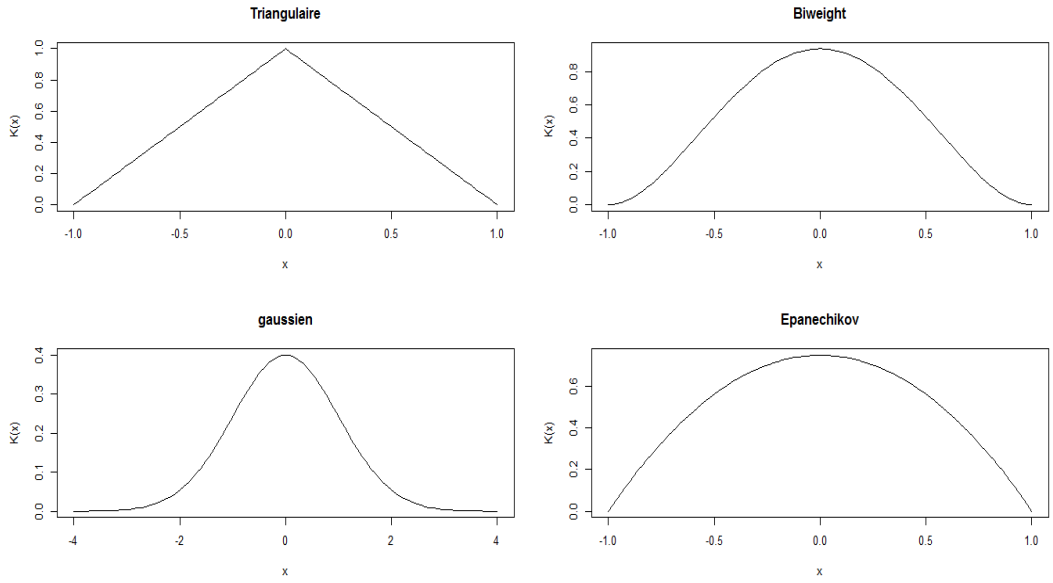

FIG. 1.1 – Allures des noyaux : Triangulaire, Biweight, Gaussien et Epanechnikov.

Remarque 1.2.1 $\int f_{n,X}(x) = 1$,*donc, si* $K(x) \geq 0$, $\forall x \in \mathbb{R}$, *alors* $\hat{f}_n$*est une densité*

Théoréme 1.2.1 *soit* $K : (\mathbb{R}^m, \mathcal{B}^m) \to (\mathbb{R}, \mathcal{B})$ *une fonction mesurable, où* $\mathcal{B}^p$ *est la tribu borélienne de* $\mathbb{R}^p$*, vérifiant :* $\exists M$*(constante) telle-que,*

$$\forall z \in \mathbb{R}^m, |K(z)| \leq M$$

$$\int_{\mathbb{R}^m} |K(z)|\, dz < \infty$$

Et

$$\|z\|^m |K(z)| \to 0 \text{ :quand } \|z\| \to \infty$$

Par ailleurs, soit $g : (\mathbb{R}^m, \mathcal{B}^m) \to (\mathbb{R}, \mathcal{B})$ *une fonction tq*

$$\int_{\mathbb{R}^m} |g(z)| dz < \infty$$

Si g *est continue, et si* $0 < h \to 0$, *quand* $n \to \infty$ *alors :*

$$\lim_{n\to\infty} \frac{1}{h^m} \int_{\mathbb{R}^m} K\left(\frac{z}{h}\right) g(x-z) dz = g(x) \int_{\mathbb{R}^m} |K(z)| \, dz$$

Si g *est uniformément continue, alors la convergence ci dessus est uniforme.*

Lemme 1.2.1 ***(Bochner)*** *1) Soit* K *un noyau de Parzen -Rosenblatt et* $f \in L^1$ *alors en tout point* x *de continuité de* f *on a*

$$\lim_{h\to 0} (f * K_h)(x) = f(x)$$

2) Soit maintenant K *un noyau quelconque ; si* $f \in L^1$ *est uniformément continue, alors*

$$\lim_{h\to 0} \sup_{x\in \mathrm{IR}} |f * K_h(x) - f(x)| = 0.$$

1.3 Les lois de probabilités usuelles

1.3.1 La loi binomiale

Définition 1.3.1 *on dit que la v.a.discréte X suit une loi binomiale $\beta(n;p)$ si sa fonction de probabilité est :*

$$p(x) = (x)^n p^x (1-p)^{n-x}, x = 0, 1, 2, ..., n$$

Proposition 1.3.1 *soit $X \backsim \beta(n;p)$, alors :*

$$E(X) = np$$
$$Var(X) = npq$$
$$\Psi_X(t) = [p\exp(t) + (1-p)]^n$$

1.3.2 Loi continue uniforme $U[a,b]$

on dit que X suit loi uniforme sur l'intervalle fini $[a,b]$ si sa densité est constane sur $[a,b]$,soit :

$$f(x) = \begin{cases} \frac{1}{b-a} & si \in \ a \leq X \leq b \\ 0 & si \qquad non \end{cases}$$

sa fonction de répartition est :

$$F(x) = \begin{cases} 0 & si \in \quad X < a \\ \frac{x-a}{b-a} & si \in \quad a \leq X \leq b \\ 1 & si \in \quad X > b \end{cases}$$

on peut aisément vérifier que :

$$E(X) = \frac{a+b}{2}$$
$$Var(X) = \frac{(b-a)^2}{12}$$

La quelle définit loi $\Gamma(r, \lambda)$ .des propriétés des sommes de v.a i.i.d on déduit immédiatement

1.3.3 Loi de weibull $w(\lambda, \alpha)$

la fonction de répartition et la fonction de densité de cette loi ,notée $w(\lambda, \alpha)$ où λ et α sont deux paramétres strictement positifs, sont :

$$F(x) = \begin{cases} 1 - \exp(-\lambda x^{\alpha}) & si \in \quad x \geq 0 \\ 0 & si \in \quad x < 0 \end{cases}$$

et

$$f(x) = \begin{cases} \alpha\lambda x^{\alpha-1} \exp(-\lambda x^{\alpha}) & si \in \quad x \geq 0 \\ 0 & si \in \quad x < 0 \end{cases}$$

$$E(X) = \frac{\Gamma\left(1+\frac{1}{\alpha}\right)}{\lambda^{\frac{1}{\alpha}}}$$
$$Var(X) = \frac{\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma^2\left(1+\frac{1}{\alpha}\right)}{\lambda^{\frac{2}{\alpha}}}$$

où Γ est la fonction gamma d'Euler.

1.3.4 La loi exponentielle$\xi(\lambda)$

la fonction de répartition $P(x \leq t)$:

$$F(t) = \begin{cases} 1 - \exp(-\lambda t) & si \in \quad t \geq 0 \\ 0 & si \in \quad t < 0 \end{cases}$$

puis de la densité ,par dérivation :

$$f(t) = \begin{cases} \lambda \exp(-\lambda t) & si \in \quad t \geq 0 \\ 0 & si \in \quad t < 0 \end{cases}$$

et

$$E(X) = \frac{1}{\lambda}$$
$$Var(X) = \frac{1}{\lambda^2}$$

1.4 Théorèmes de convergences de variables aléatoires

Dans ce qui suit, nous présentons certaines modes de convergence de variable aléatoire :

1. Convergence en probabilité : On dit que a suite de v.a. (X_n) converge en probabilité vers une v.a. X si, pour tout $\varepsilon > 0$:

$$\mathbb{P}\left(|X_n - X| < \varepsilon\right) \longrightarrow 1 \quad \text{quand } n \longrightarrow \infty, \quad on \text{ écrit alors,} \quad X_n \xrightarrow{\mathbb{P}} X.$$

2.Convergence en loi : On dit que la suite de v.a. (X_n), de fonction de répar-

tition F_n, converge en loi vers une v.a. X de fonction de répartition F, si la suite $(F_n(x))$ converge vers $F(x)$ en tout point x où F est continue : $X_n \xrightarrow{\mathcal{L}} X$, quand $n \longrightarrow \infty$.

3. Convergence en moyenne quadratique : On dit que la suite de va (X_n) converge en moyenne quadratique vers une $v.a.$ X si :

$$\mathbb{E}|X_n - X|^2 \longrightarrow 0, \quad \text{quand } n \longrightarrow \infty, \quad \text{on écrit alors, } X_n \xrightarrow{mq} X.$$

4.Théorème de la limite centrale : Si $X_1, X_2, ..., X_n$ est une suit de $v.a.$ $i.i.d.$ d'espérance $\mu < \infty$ et de variance $\sigma^2 < \infty$, alors

$$\sqrt{n}\left(\bar{X} - \mu\right)/\sigma \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad \text{quand} \quad n \longrightarrow \infty, \qquad \text{où } \bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i.$$

5.Théorème (Lois des grands nombres) : Si $(X_1, X_2, ..., X_n)$ est un échantillon provenant d'une v.a. X telle que $\mathbb{E}|X| < \infty$, alors :

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}(X) \quad \text{quand } n \longrightarrow \infty, \qquad \text{(loi faible)}$$

$$\bar{X}_n \xrightarrow{\mathbb{P}s.} \mathbb{E}(X) \quad \text{quand } n \longrightarrow \infty, \qquad \text{(loi forte).}$$

6. Définition (Biais d'un estimateur) : Un estimateur $\hat{\theta}_n$ de θ est dite sans biais si pour tout $\theta \in \Theta$ et tout entier positif n : $\mathbb{E}\left(\hat{\theta}_n\right) = \theta$.
De même, $\hat{\theta}_n$ est dite asymptotiquement sans biais si pour tout $\theta \in \Theta$:

$$\mathbb{E}\left(\hat{\theta}_n\right) \longrightarrow \theta \text{ quand } n \longrightarrow \infty.$$

La quantité : $\mathbb{E}\left(\hat{\theta}_n\right) = \theta$, est appelée le biais de l'estimateur $\hat{\theta}_n$.

7. Définition ($o_p(1)$ *et* $O_p(1)$) : La notation $o_p(1)$ segnifie qu'une suite de v.a's convergent vers 0 en probabilité. La notation $O_p(1)$ désigne une séquence qui est bornée en probabilité. Plus généralement, pour une suite donnée de v.a. R_n :

$$X_n = o_p(1) \Longleftrightarrow X_n = R_n Y_n \quad \text{et} \quad Y_n \xrightarrow{\mathbb{P}} 0,$$
$$X_n = O_p(1) \Longleftrightarrow X_n = R_n Y_n \quad \text{et} \quad Y_n = O_p(1).$$

Ces quantités vérifies les assertions :

$$o_p(1) + O_p(1) = O_p(1), \qquad o_p(1) O_p(1) = o_p(1),$$
$$(1 + o_p(1))^{-1} = O_p(1), \qquad o_p(O_p(1)) = o_p(1),$$
$$o_p(R_n) = R_n o_p(1), \qquad O_p(R_n) = R_n O_p(1).$$

Chapitre 2

Estimation non paramétrique de la fonction de régression

Dans ce chapitre, nous donnerons la définition de l'estimateur non paramétrique de la régression par la méthode du noyau. Nous étudions, les propriétés asymptotiques de l'estimateur et le paramètre de lissage h et le choix du noyau K

2.1 Introudction

Un des modèles le plus fréquemment rencontré en statistique paramétrique ou non paramétrique est le modéle de régression. On dispose d'un échantillon, composée de n couples indépendants de variables aléatoires $(X_1, Y_1), ..., (X_n, Y_n)$, et on dénote par (X, Y) un élément générique de cet échantillon. Dans le modéle de régression non paramétrique, on suppose typiquement l'existence d'une fonction $m(.)$ qui exprime la valeur moyenne de la variable réponse Y en fonction de la

variable d'entrée X :

$$Y_i = m(X_i) + \varepsilon_i, \text{ pour } 1 \leq i \leq n, \text{ avec } \varepsilon_i = \varepsilon \rightsquigarrow N\left(\mu, \sigma^2\right)$$

Définition 2.1.1 *Soit (X, Y) un couple de variables aléatoires rélles admettant une densité jointe sur $\mathbb{R}^2$ notée $f_{X,Y}$,et une densité marginale f_X. La variable Y est supposée intégrable,ie $E|Y| < \infty$.Nous pouvons, alors définir proprement la fonction de régression ou espérance conditionnelle de Y sachant $X = x$, par*

$$m(x) = E\left[Y/X = x\right] = \frac{\int_{\mathbb{R}} y f_{X,Y}(x, y)dy}{\int_{\mathbb{R}} f_{X,Y}(x)dy} = \frac{\Phi(x)}{f_X(x)} \tag{2.1}$$

lorsque la densité $f_X(x)$ est différente de zéro. Le probléme de l'estimation de $m(.)$ est du type non-paramétrique, i.e. la fonction de régression appartient à un ensemble non paramétrique (infini-dimensionnel)

2.2 L'estimateur de Nadaraya-Watson

Supposons que l'on dispose d'un néchantillon $(X_1, Y_1), ..., (X_n, Y_n)$, de variables aléatoires à valeurs réelles, de même loi que le couple(X, Y) On se propose de construire un estimateur $m_n(x)$ de la fonction de régression à partir des couples d'observations $(x_1, y_1), ..., (x_n, y_n)$.Le premier estimateur rencontré dans la littérature est l'estimateur à noyau de **Nadaraya(1964)Watson(1964)**

Définition 2.2.1 *L'estimateur de régression se présente sous la forme d'une moyenne locale pondérée des valeurs Y_i et est défini par*

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} \times 1\left\{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0\right\}$$

De manière similaire, nous pouvons définir l'estimateur par,

$$m_n(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} & si \quad \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0 \\ \frac{1}{n}\sum_{i=1}^n Y_i & \text{sinon} \end{cases} \tag{2.2}$$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} = \frac{\Phi_{n,X}(x)}{f_{n,X}(x)} \tag{2.3}$$

Définition 2.2.2 *On appelle estimateur à noyau de la densité conjointe du couple* (X, Y)

$$\hat{f}_{X,Y}(x) = \frac{1}{nh^2}\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right)$$

2.3 Propriétés de l'estimateur

D'une manière analogue aux propriétés asymptotiques de l'estimateur de Parzen(1962) Rosenblatt(1956)

nous supposons que K est un noyau vérifant les conditions suivantes

$(H1)$ K est bornée,c'est à dire $\underset{x\in\mathbb{R}}{sup}\,|K(x)| < \infty$

$(H2)$ $\lim |x|\, K(x) \to 0$,quand $|x| \to +\infty$

$(H3)$ $K \in L_1(\mathbb{R})$,c'est à dire $\int_{\mathbb{R}} |K(x)|\, dx < +\infty$

$(H4)$ $\int_{\mathbb{R}} K(x)dx = 1$

$(H5)$ $\int_{\mathbb{R}} xK(x)dx = 0$

$(H6)$ $\int_{\mathbb{R}} x^2K(x)dx < +\infty$

$(H7)$ K est bornée, intégrable et à support compact,

L'étude asymtotique du biais et de la variance de l'estimateur de Nadaraya(1964)-

Watson(1964) détermine les conditions suffisantes à la consistance de cet estimateur

2.3.1 Etude asymptotique du biais

Le traitement du biais est purement analytique et repose essentiellement sur le développement de Taylor. Il nous faut supposer certaines conditions de régularités sur les fonctions $m(.)$ et $f_X(.)$ qui détermineront l'ordre du biais asymptotique en fonction du paramètre de lissage h

L'étude asymtotique du biais basée sur la proposition suivante

Proposition 2.3.1 *lorsque Y est bornée et $nh \to \infty$*

$$E\left[m_n(x)\right] = \frac{E\left[\Phi_{n,X}(x)\right]}{E\left[f_{n,X}(x)\right]} + o((nh)^{-1})$$

lorsque$E(Y^2) < \infty$, et $nh^2 \to \infty$

$$E\left[m_n(x)\right] = \frac{E\left[\Phi_{n,X}(x)\right]}{E\left[f_{n,X}(x)\right]} + o(\left(n^{1/2}h\right)^{-1})$$

Maintenant nous sommes en mesure d'énoncer le resultat suivant

Proposition 2.3.2 *supposons que $f_X(.)$et $m(.)$ sont le classe $C^2(\mathbb{R})$ et que le noyau K est d'ordre 2, i.e. tel que*

$\int_{\mathbb{R}} K(x)dx = 1, \int_{\mathbb{R}} xK(x)dx = 0, \int_{\mathbb{R}} x^2K(x)dx < +\infty$

Nous avons alors, lorsque et si $h \to 0$ et $nh \to \infty$

$$E\left[m_n(x)\right]-\left[m(x)\right]=\frac{h^2}{2}\left(\left(m''(x)+2m'(x)\frac{f'_X(x)}{f_X(x)}\right)\int_{\mathbb{R}}u^2k(u)du\right)(1+o(1)) \tag{2.4}$$

Preuve.

$$\begin{aligned}E\left[m_n(x)\right]-\left[m(x)\right]&=\left[EK\left(\frac{x-X}{h}\right)\right]^{-1}\left\{\int_{\mathbb{R}}\frac{1}{h}k\left(\frac{x-t}{h}\right)\Phi(t)dt-m(x)\int_{\mathbb{R}}\frac{1}{h}k\left(\frac{x-t}{h}\right)f(t)dt\right\}\\&=\frac{h^2}{2}\times\left\{\{f(x)\}^{-1}\times\{m''(x)-m(x)f''(x)\}\times\int_{\mathbb{R}}u^2k(u)du\right\}(1+o(1))\end{aligned}$$

comme $\Phi(x)=m(x)f(x)$ L'équation précédent peut s'écrire : ■

$$E\left[m_n(x)-m(x)\right]=\frac{h^2}{2}\left\{\left\{m''(x)+2m'(x)\frac{f'(x)}{f_X(x)}\right\}\int_{\mathbb{R}}u^2k(u)du\right\}(1+o(1))$$

D'où

$$\lim_{n\to\infty}E[m_n(x)]=m(x)$$

2.3.2 Etude asymptotique de la variance

Proposition 2.3.3 *On suppose $E[Y^2]<\infty$, alors en chaque point de continuité des fonctions $m(x)$,$f_X(x)$ et $\sigma^2(x)=var(Y/X=x)$*

$$var\left[m_n(x)\right]=\frac{1}{nh}\left\{\frac{\sigma^2(x)}{f_X(x)}\int K^2(u)du\right\}(1+o(1)) \tag{2.5}$$

telque $f_X(x)>0$,où le terme $o(1)$ tend vers 0 lorsue $h\to 0$

Preuve. En utilisant le lemme de **Bochner,**nous obtenons

$$\begin{aligned} var\left[f_{n,X}(x)\right] &= \frac{1}{nh^2}\left\{E\left[K^2\left(\frac{x-X}{h}\right)\right] - E\left[K\left(\frac{x-X}{h}\right)\right]^2\right\} \\ &= \frac{1}{nh}\left\{\int_{\mathbb{R}} K^2(u) f_X(x-hu)du - h\left(\int_{\mathbb{R}} K(u) f_X(x-hu)\right)^2\right\} \\ &= \frac{1}{nh}\left\{(f_X(x)\int K^2(u)du\right\}(1+o(1)) \end{aligned}$$

lorsque $h \to 0$,soit la fonction $s(x) = \int y^2 f\,(x,y)\,dy$, nous avons

$$\begin{aligned} var\left[\Phi_n(x)\right] &= \frac{1}{nh}\left\{E\left[Y^2K^2\left(\frac{x-X}{h}\right)\right] - E\left[YK\left(\frac{x-X}{h}\right)\right]^2\right\} \\ &= \frac{1}{nh}\left\{\int_{\mathbb{R}} K^2(u) s(x-hu)du - h\left(\int_{\mathbb{R}} K(u)\Phi(x-hu)\right)^2\right\} \\ &= \frac{1}{nh}\left\{s(x)\int K^2(u)du\right\}(1+o(1)) \end{aligned}$$

et

$$E\left[\{f_{n,X}(x) - E(f_{n,X}(x))\}\{\Phi_{n,X}(x) - E(\Phi_{n,X}(x))\}\right] = \frac{1}{nh}\Phi(x)\int_{\mathbb{R}} K^2(u)du(1+o(1))$$

$$var\left[f_{n,X}(x)\right] = \frac{1}{nh}f_X(x)\int_{\mathbb{R}} K^2(u)du(1+o(1))$$

posons

$$A_n(x) = \begin{pmatrix} f_{n,X}(x) \\ \Phi_{n,X}(x) \end{pmatrix}$$

La matrice de variance covriance de $\sum\left[A_n(x)\right]$ est, alors donnée par l'éxpression

suivante

$$\sum [A_n(x)] = \frac{1}{nh} \begin{pmatrix} f_X(x) & \Phi(x) \\ \Phi(x) & s(x) \end{pmatrix} \int_{\mathbb{R}} K^2(u)du(1+O(1))$$

En remarquant que

$$\begin{pmatrix} -\frac{\Phi(x)}{\{f_X(x)\}^2} & \frac{1}{f_X(x)} \end{pmatrix} \begin{pmatrix} f_X(x) & \Phi(x) \\ \Phi(x) & s(x) \end{pmatrix} \begin{pmatrix} -\frac{\Phi(x)}{\{f_X(x)\}^2} \\ \frac{1}{f_X(x)} \end{pmatrix} = \frac{s(x)}{\{f_X(x)\}^2} - \frac{\{\Phi(x)\}^2}{\{f_X(x)\}^3}$$

En obtient , alors

$$\begin{aligned} var\,[m_n(x)] &= \frac{1}{nh} \left(\frac{s(x)}{\{f_X(x)\}^2} - \frac{\{\Phi(x)\}^2}{\{f_X(x)\}^3} \right) \int_{\mathbb{R}} K^2(u)du(1+O(1)) \\ &= \frac{1}{nh} \times \left\{ \frac{\sigma^2(x)}{f_X(x)} \int_{\mathbb{R}} K^2(u)du \right\} (1+O(1)) \end{aligned}$$

■

2.4 Erreur Quadratique Moyenne de $m_n(x)$

L'erreur quadratique moyenne MSE (mean square error) est une mesure permettant d'évaluer la similarité de m_n par rapport à la fonction de régression inconnue m , au point x donné de $\mathbb{R}$.

$$MSE\,(m_n(x)) = E\,[m_n(x) - m(x)]^2$$

Le développement de cette expression faite précédement , nous donne

$$MSE\left(m_n(x)\right) = var\left[m_n(x)\right] + \left[biais(m_n(x))\right]^2 \tag{2.6}$$

Nous constatons d'une part que les expression du $m_n(x)$ et de la variance de $m_n(x)$ (voir les proposition 2.3.2, 2.3.3) permettent de conclure qu'une grande valeur de h donne une augmentation du biais et une diminution de la variance (estimation fortement biaisée) et qu' un faible paramétre h , donne une diminution du biais et une augmentation de la variance (phénomène de sous lissage)

D'autre part, sous les hypothéses de ces mémes proposition , nous obtenons

$$MSE\left(m_n(x)\right) = \frac{h^4}{4}\left[(m''(x) + 2m'(x)\,\frac{f'_X(x)}{f_X(x)})(u^2K(u)) + o(1)\right]^2 \tag{2.7}$$

$$+ \frac{1}{nh}\left(\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]\right)(1 + o(1) \tag{2.8}$$

où $u^p k^q(u) = \int t^p k^q dt$ .pour trouver donc compromis entre le biais et la variance nous minimisons par rapport à h l'expression dr l'erreur quadratique moyenne asymtotique$AMSE$ (asymptotique mean square error) donne par

$$AMSE[m_n(x)] = \frac{h^4}{4}\left\{m''(x) + 2m'(x) + \frac{f'_X(x)}{f_X(x)}\right\}^2\left[u^2K(u)\right]^2 + \frac{1}{nh} \times \frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right] \tag{2.9}$$

2.5 Choix du paramètre de lissage et le noyau

1.choix du paramètre de lissage h

Comme $AMSE$ est une fonction convexe. La fenétre optimale $h^{MSE}_{opt(m_n(x))}$ obtenue

par la méthode de validation croisée non biaisée est donné par

$$h^{MSE}_{opt(m_n(x))} = argmin_h[AMSE(m_n(x))]$$

est solution de l'équation suivante

$$\frac{\partial}{\partial h_n}\left[\frac{h^4}{4}\left\{m''(x)+2m'(x)+\frac{f'_X(x)}{f_X(x)}\right\}^2\left[u^2K(u)\right]^2+\frac{1}{nh}\times\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]\right]=0$$

Lorsque$\left[m''(x)+2m'(x)+\frac{f'_X(x)}{f_X(x)}\right]^2\left[u^2K(u)\right]^2\neq 0$, on a :

$$h^{MSE}_{opt(m_n(x))} = n^{-\frac{1}{5}}\left(\frac{\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]}{\left[m''(x)+2m'(x)+\frac{f'_X(x)}{f_X(x)}\right]^2\left[u^2K(u)\right]^2}\right)^{\frac{1}{5}}$$

Définition 2.5.1 *On pose*

$$Biais(x)^2=\left[m''(x)+2m'(x)+\frac{f'_X(x)}{f_X(x)}\right]^2\left[u^2K(u)\right]^2$$

et

$$var(x)=\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]$$

Donc

$$AMSE=\frac{1}{4}Biais(x)^2h^4+\frac{var(x)}{nh}$$

On calcule la dérivée de l'$AMSE$:

$$\frac{\partial}{\partial h}AMSE=\frac{\partial}{\partial h}\left[\frac{1}{4}Biais(x)^2h^4+\frac{var(x)}{nh}\right]=Biais(x)^2h^3-\frac{var(x)}{nh^2}$$

Il est clair que la deuxième dérivée de l'$AMSE$ est positive, alors on a une certitude qu'elle admet un minimum :

$$\begin{aligned} Biais(x)^2h^3 - \frac{var(x)}{nh^2} &= Biais(x)^2h^5 - \frac{var(x)}{n} \\ &= 0 \end{aligned}$$

On trouve :

$$\begin{aligned} h^{MSE}_{opt(m_n(x))} &= \left(\frac{var(x)}{nBiais(x)^2}\right)^{\frac{1}{5}} \\ &= \left(\frac{\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]}{n\left[m''(x) + 2m'(x) + \frac{f'_X(x)}{f_X(x)}\right]^2\left[u^2K(u)\right]^2}\right)^{\frac{1}{5}} \end{aligned} \tag{2.10}$$

On s'intéresse maintenant à l'approche globale pour la selection du paramétre h, pour cela on introduit le critére d'erreur quadratique intégré $MISE$ (mean integrated squared error) de $m_n(x)$

$$MISE[m_n(x)] = E\left[\int_{\mathbb{R}} \left(m_n(x) - m(x)\right)^2 dx\right]$$

En appliquant le théorème de Fubini, on a

$$MISE[m_n(x)] = \int_{\mathbb{R}} \left[E\left(m_n(x) - m(x)\right)^2 dx\right] \tag{2.11}$$

Sous les même hypothèses que les proposition 2.3.2et2.3.3 ,on a

$$AMISE[m_n(x)] = \frac{h^4}{4}\int_{\mathbb{R}}\left\{m''(x) + 2m'(x) + \frac{f'_X(x)}{f_X(x)}\right\}^2 dx\left[u^2K(u)\right]^2 + \int_{\mathbb{R}}\frac{1}{nh}\times\frac{\sigma^2(x)}{f_X(x)}dx\left[K^2(u)\right] \tag{2.12}$$

La fenêtre $h^{MISE}_{opt(m_n(x))}$ minimisant l'$AMISE$ obtenue par la méthode de validation croisée biaisée est donnée par

$$h^{MISE}_{opt(m_n(x))} = argmin_h[AMSIE(m_n(x))]$$

$$h^{MISE}_{opt(m_n(x))} = n^{-1/5}\left\{\frac{\int_{\mathbb{R}}\frac{\sigma^2(x)}{f_X(x)}\left[K^2\right]dx}{\int_{\mathbb{R}}\left\{m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}\right\}^2 dx\left[u^2K\right]^2}\right\}^{1/5} \tag{2.13}$$

Un travail similaire se fait pour le choix optimum du paramètre de lissage dans le cas de l'estimateur de Parzen(1962)-Rosemblatt (1956), nous obtenons

$$h^{MSE}_{opt(f_{n,X}(x))} = n^{-1/5}\left\{\frac{f_X(x)\left[K^2\right]}{(f''_X(x))^2\left[u^2K\right]^2}\right\}^{1/5} \tag{2.14}$$

$$h^{MISE}_{opt(f_{n,X}(x))} = n^{-1/5}\left\{\frac{\left[K^2\right]}{\int_{\mathbb{R}}(f''_X(x))^2dx\left[u^2K\right]^2}\right\}^{1/5} \tag{2.15}$$

quand $f''_X(x) \neq 0$

Nous notons que l'expression de h optimal, minimisant asymptotiquement les quatre critères d'erreurs à la forme $Cn^{-1/5}$, alors

$$h_{opt} = Cn^{-1/5} \tag{2.16}$$

Où la constante C est en fonction de la distribution et de termes aléatoires inconnues.

2. Choix du noyau K :

Le choix du noyau semble être réglé lorsque on choisit le $MISE$ comme critère d'optimisation. En réalité, on peut être amené à envisager d'autre noyaux pour des raisons de lissage ou de facilité de calcul.

Le tableau2.1donne efficacité relative de ces noyaux les plus utilisés. Comme ces efficacités sont trés rapprochées, le choix du noyau influe peu sur la valeur du $MISE$ asymptotique. En pratique on choisira K en tenant compte de la facilité des calculs plutôt que de l'efficacité relative.

$$Eff(K) = \frac{AMISE(Kopt)}{AMISE(K)} \tag{2.17}$$

noyau	Efficacité
Uniforme	0.9295
Triangulaire	0.9859
Gaussien	0.9512
biweight	0.9939
Epanechnikov	1

TAB. 2.1 – Efficacité du noyau

Chapitre 3

Application sous R

On termine ce mémoire par une étudé de simmulation,utilisant le logicilel R pour calculer et représenter graphiquement la fonction de régression et son estimateur en vue de les comparer dans des situations simulées. Il s'agit de l'estimateur proposé par Nadaraya-Watson)(1964) et présenté au chapitre 2. Nous donnons des exemples sur cet estimateur qui expriment l'importance de paramètre de lissage h, du noyau K.

Ensuite, nous présentons les résultats obtenus pour les différents jeux de données ainsi que pour les différentes noyaux K (noyau Gaussien à support non compact et noyau triangulaire à support compact), différents valeurs de h strictement positif (h fixé ou h varié), régression linéaire et non linéaire.

3.1 Présentation des données

Supposons qu'on a observé un échantillon de taille n d'un couple de v.a (X, Y), la relation entre X_i et Y_i (les valeurs de X et Y respectivement) ; est définie dans

le cadre

du modèle de régression standard suivant :

$$Y_i = m(X_i) + \varepsilon_i, i = 1, ..., n$$

où :ε_i des erreurs centrées et indépendantes de X_i et m la fonction de regression

$$m(x) = E\left[Y/X = x\right] = \frac{\int_{\mathbb{R}} y f_{X,Y}(x, y)dy}{f_X(x)} \tag{3.1}$$

où $f_X(x)$ est la densité de la variable X.

L'estimation non paramétrique d'une fonction de régression par la méthode du noyau [NW] est définie par la forme suivante :

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} = \frac{\Phi_{n,X}(x)}{f_{n,X}(x)} \tag{3.2}$$

Il dépend de la taille de l'échantillon n et aussi du noyau K et de la fenêtre h, qu'il faut choisir pour calculer $m_n(x)$: avec $\Phi_{n,x}(x)$ est l'estimateur naturel de $\Phi(x)$:

$$\Phi_{n,X}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)$$

et $f_{n,X}(x)$ l'estimateur à noyau étudié au chapitre 1 de la densité :

$$f_{n,X}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

Dans la suite de ce chapitre, nous supposons que notre modèle à la forme

$$y = m(x) + \varepsilon;\ \text{où } \varepsilon \to \mathcal{N}\left(0.\sigma^2\right)$$

et nous étudions les deux cas suivant :

1) Régression linéaire : $m(x) = 2 + 0.7x + \varepsilon$.

2) Régression non linéaire : $m(x) = sin(x) + \varepsilon$.

on supposons que : X est de loi normale centré de variance $\sigma^2 = 2$ et ε un terme d'erreur de loi $\mathcal{N}(0; 1)$.

Nous allons donc étudier les cas suivants dans chaque modèle :

1. Paramètre de lissage ou fenêtre h fixée, K noyau gaussien (noyau à support non compact) et n varié.
2. Paramètre de lissage ou fenêtre h fixée, K noyau Triangulaire (noyau à support compact) et n varié.
3. Paramètre de lissage ou fenêtre h fixée, K noyau d'Epanechnikov (noyau à support compact) et n varié.
4. n fixée et la fenêtre h variée (noyau K gaussien).
5. n fixée et la fenêtre h variée (noyau K Triangulaire).
6. n fixée et la fenêtre h variée (noyau K d'Epanechnikov).

3.2 Etude du régression linéaire

On veut estimer le modèle linéaire

$$y = 2 + 0.7x + \varepsilon, \ x \rightsquigarrow \mathcal{N}(0, 2)$$

où ε un terme d'erreur de loi $\mathcal{N}(0, 1)$.

Dans les résultats graphique de cette section, on a :

1) La droite noire exprime la fonction de régression $m(x)$[3.1]

2) La droite en blue exprime la fonction de régression empirique $m_n(x)$[3.2]

3) L'axe des absices représente les valeurs des x et l'axe des coordonnées les valeurs des m_n (et m).

3.2.1 Paramètre de lissage h fixé et n variée

En choisissant le paramètre de lissage $h = n^{-1/5}$ (fixée) et n varié ($n = 30, 120, 400$)

– **K à support non compact**

Dans ce premier cas, on choisit un noyau gaussien $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)$ et on va utiliser le code ci-dessous pour estimer ce modèle, et le resultat graphique obtenu représenté dans la figure[3.1]

Code R :

```
rm(list=ls(all=TRUE)) # Nouveau programme
n=30
X=rnorm(n,0,2)
E=rnorm(n)
Y=2+.7*X+E
h=n^-.2
# Initiation
s=100
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=numeric(n)
fn=numeric(s)
```

```
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h)}
fn[j]=sum(V)/(n*h)}
# Fonction Hn(.)
W=numeric(n)
Hn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] }
Hn[j]=sum(W)/(n*h)}
# Graphes
op=par(mfrow=c(1,3))
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=30",type='l',col=4, lwd= 2)
abline(2,.7,lwd= 2)
#####Pour n =120
n=120
X=rnorm(n,0,2)
E=rnorm(n)
Y=2+.7*X+E
h=n^-.2
V=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
W=numeric(n)
for(j in 1 :s){
```

```
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i]}
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=120",type='l',col=4, lwd= 2)
abline(2,.7,lwd= 2)
#####Pour n =400
n=400
X=rnorm(n,0,2)
E=rnorm(n)
Y=2+.7*X+E
h=n^-.2
V=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h)}
fn[j]=sum(V)/(n*h)}
W=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i]}
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=400",type='l',col=4, lwd= 2)
abline(2,.7,lwd= 2)
par(op)
```

Remarque 3.2.1 *Par la comparaison graphique, on remarque que le graphe blue de m_n est approche beaucoup à la droite noire de m dans le troisième graphe, donc*

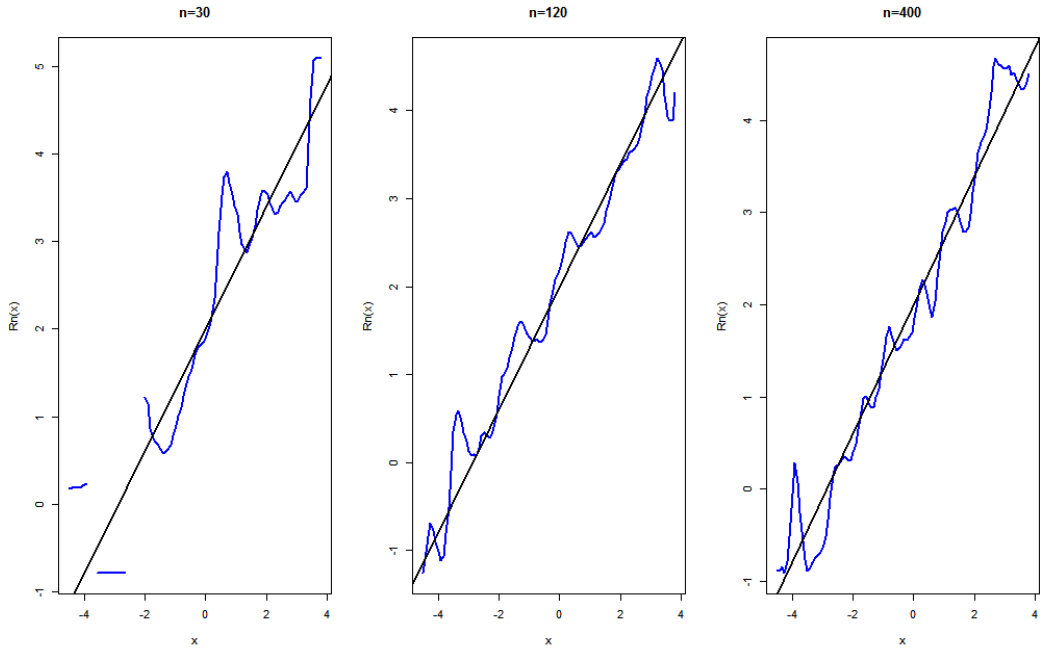

FIG. 3.1 – Régression linéaire : h fixée, n variée et K noyau gaussien

ce graphe exprime la convergence de l'estimateur m_n vers m.

– **K noyau Triangulaire (à support compact)**

Dans ce second cas, on choisit le noyau Triangulaire: $K(t) = (1 - |t|)1_{[1;1]}(x)$
Ensuite, on modifie seulement cette partie dans le programme R précédent :

Noyau $K(t)$:

```
K=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
```

On obtient la figure[3.2],

Remarque 3.2.2 *on arrive au même conclusion de la convergence de l'estimateur (voir la figure [3.1], i :e ; convergence de l'estimateur pour n assez grand).*

– **K noyau d'Epanechnikov (à support compact)**

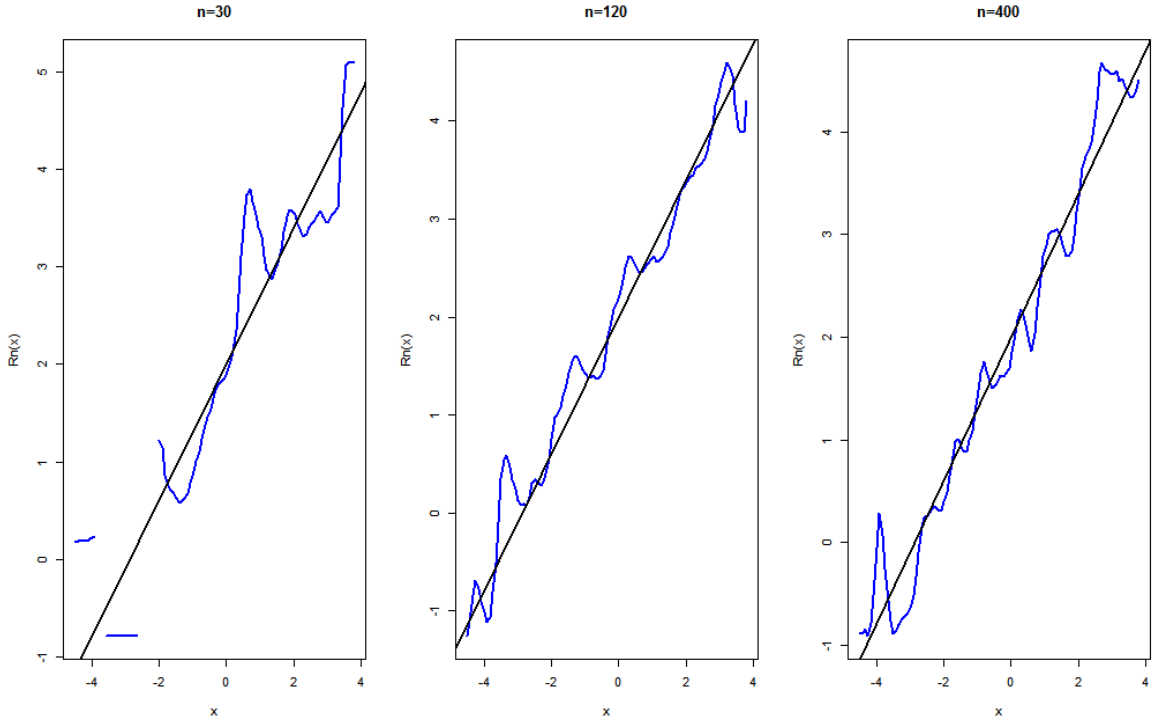

FIG. 3.2 – Régression linéaire : h fixé, n variée et K noyau Triangulaire

Dans ce troisième cas, on choisit le noyau d'Epanechnikov $K(t) = \frac{3}{4}(1-t^2)\,1_{[1;1]}(x)$. Ensuite, on modifie seulement cette partie dans le programme R précédent :

```
K=function(t){ifelse(abs(t)<1,(3/4)*(1-t^2),0)}
```

On obtient la figure[**??**] ;

Remarque 3.2.3 *on arrive au même conclusion de la convergence de l'estimateur (voir la[3.2], i :e ; convergence de l'estimateur pour n assez grand)*

3.2.2 Choix graphique du paramètre de lissage

Dans cette partie, on va prendre le paramètre de lissage dans l'intervalle $[0;1]$ de même façon pour la régression linéaire, et avec des tests graphiques en va diterminer le paramètre h optimal (au sens graphique).

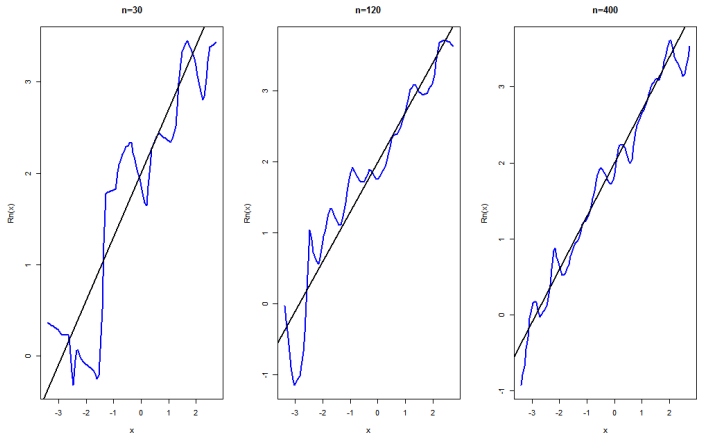

FIG. 3.3 – Régression linéaire : h fixé, n variée et K noyau d'Epanechnikov

– **K à support non compact :**

On fixe la taille de l'échantillon $n = 300$ et le noyau K est normal, l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure [3.4]

Code R

```
n=300
X=rnorm(n,0,2)
E=rnorm(n)
Y=2+.7*X+E
# Noyau Normal K(t)
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=seq(.1,.9,length=9)
# Initiation
s=100 # taille de l'intervalle [a,b]
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
```

```
V=array(dim=c(n,s,9))
fn=array(dim=c(s,9))
W=array(dim=c(n,s,9))
Hn=array(dim=c(s,9))
# density fn(x)
for(k in 1 :9){
for(j in 1 :s){
for(i in 1 :n){ V[i,j,k]=K((x[j]-X[i])/h[k]) }
fn[j,k]=sum(V[,j,k])/(n*h[k])}}
# fonction Hn(x)
for(k in 1 :9){
for(j in 1 :s){
for(i in 1 :n){ W[i,j,k]=K((x[j]-X[i])/h[k])*Y[i] }
Hn[j,k]=sum(W[,j,k])/(n*h[k])}}
Rn=array(dim=c(s,9))
for(k in 1 :9){ Rn[,k]=Hn[,k]/fn[,k]}
# Graphes
x11() # nouvelle fenettre graphique
op=par(mfrow=c(3,3))
for(k in 1 :9){
plot(x,Rn[,k],xlab="x", ylab="Rn(x)", main=" ",type='l',col=4, lwd= 2)
abline(2,.7,lwd= 2)
}
par(op)
```

Remarque 3.2.4 *Il est clair que la valeur du h optimal est de $h = 0.7$ (ligne 3, colonne 1).*

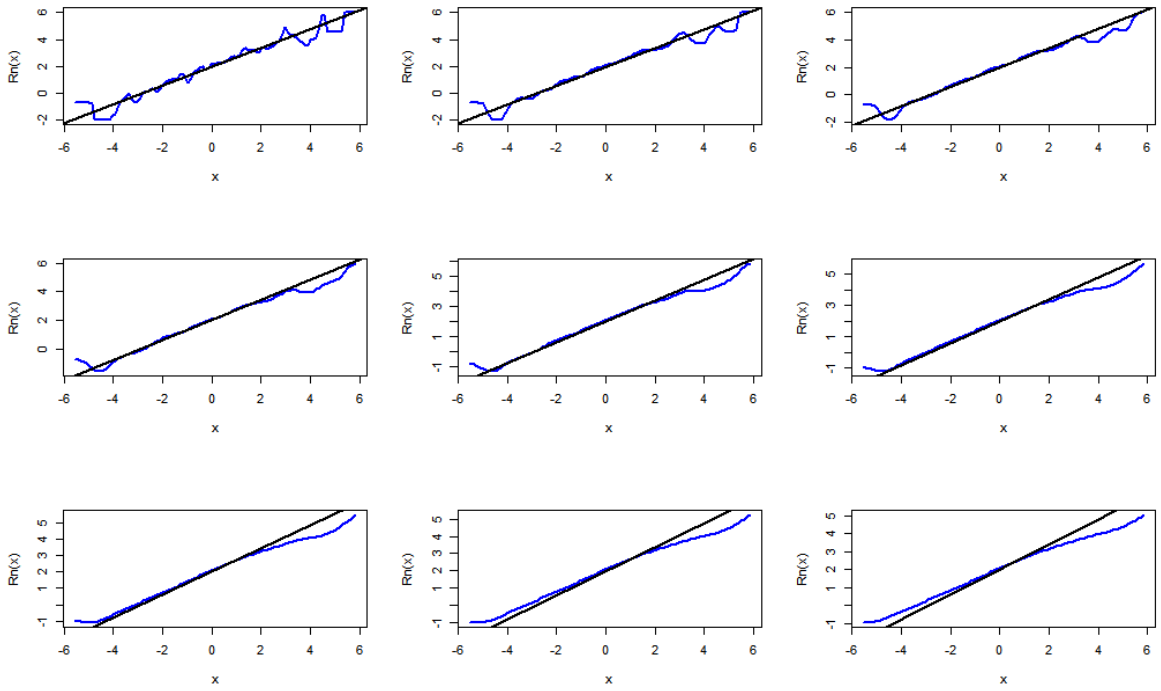

FIG. 3.4 – Régression linéaire avec h varié, n fixéé et K noyau gaussien.

– **K à support compact (noyau Triangulaire) :**

Identique aux choix précédents, mais on change le noyau $K(t) = (1 - |t|)1_{[1;1]}(x)$: (noyau Triangulaire). On obtenu la figure [3.5]qui explique l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9.

Remarque 3.2.5 *Il est clair que la valeur du h optimale est de $h = 0.9$(ligne 3, colonne 3).*

– **K noyau d'Epanechnikov (à support compact) :**

On change le noyau par $K(t) = \frac{3}{4}(1 - t^2)\, 1_{[1;1]}(x)$: (noyau d'Epanechnikov). On obtenu la figure [3.6] qui explique l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9.

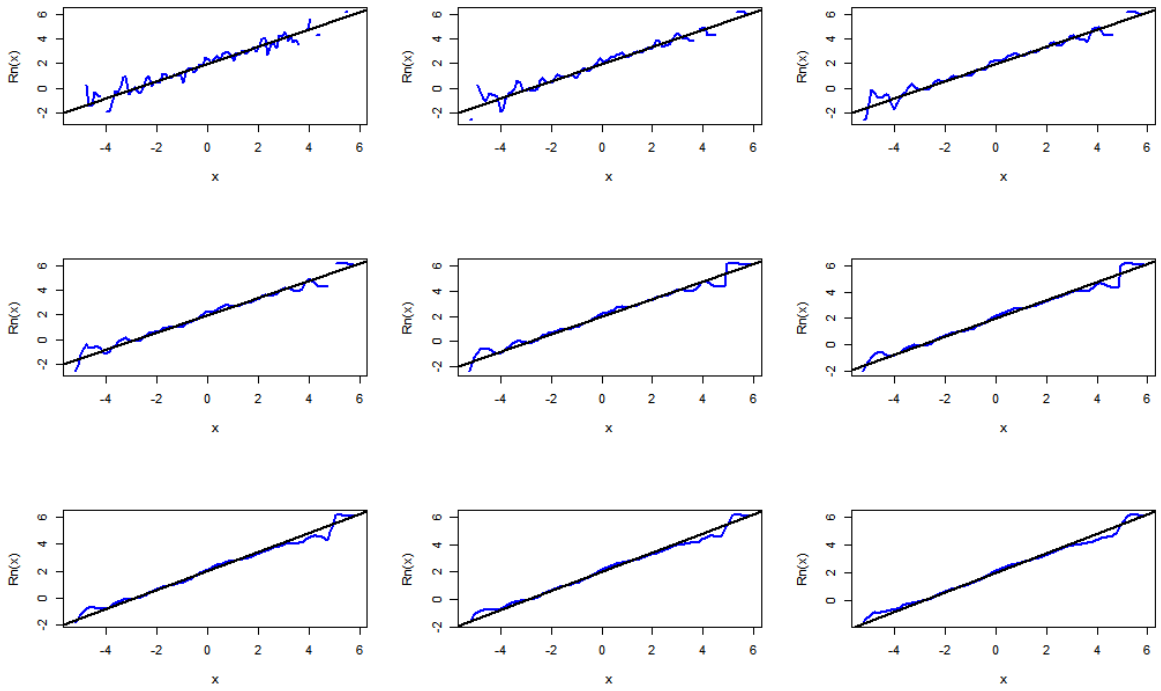

FIG. 3.5 – Régression linéaire avec h varié, n fixée et K Triangulaire

Remarque 3.2.6 *On remarque que la valeur du h optimal est de* $h = 0.9$*(ligne* 3*, colonne* 3*).*

3.3 Etude du régression non linéaire

Dans cette section, nous allons répeter les mêmes étapes que dans la régression linéaire mais avec un modèle non linéaire :

$$y = \sin x + \varepsilon, x \rightsquigarrow \mathcal{N}(0, 2)$$

Où ε un terme d'erreur de loi $\mathcal{N}(0, 1)$

Toujours, la ligne noire exprime la fonction de régression théorique $m(x)$ et la

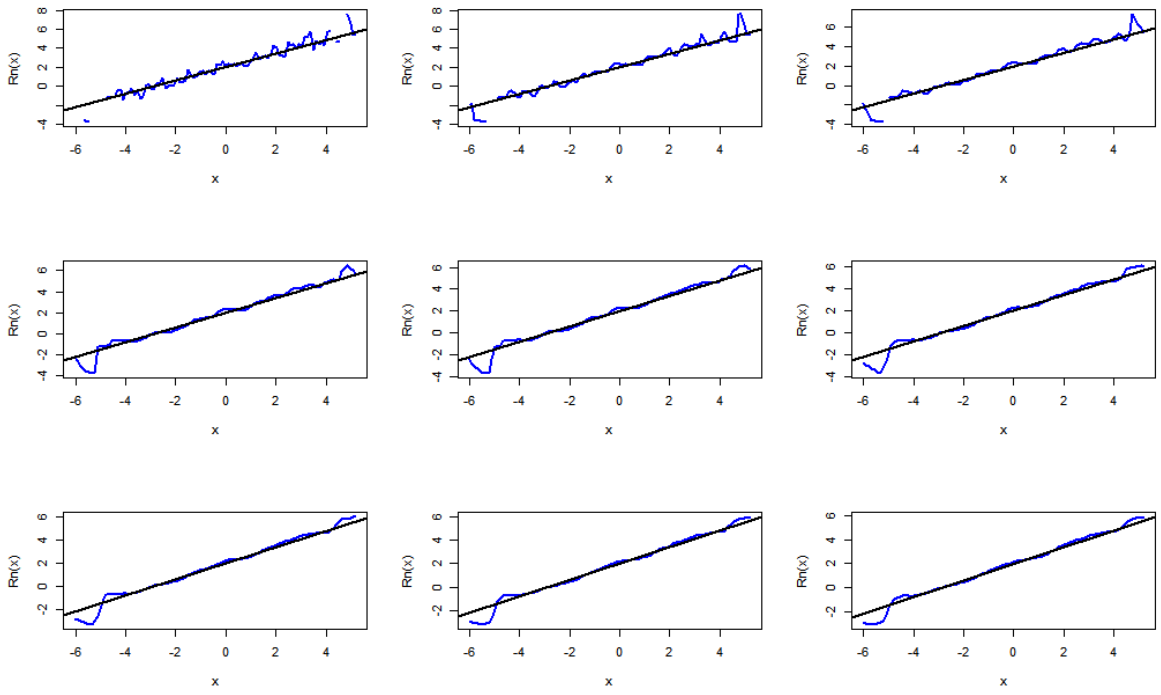

FIG. 3.6 – Régression linéaire avec h varié, n fixée et K d'Epanechnikov

ligne blue exprime la fonction de régression empirique $m_n(x)$.

3.3.1 Paramètre de lissage h fixé, n variée

– **K à support non compact**

En choisissant le paramètre de lissage $h = n^{-1/5}$ (fixée) et n variée ($n = 30, 120, 400$) et K est un noyau gaussien $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)$. On obtenu la figure |3.7],

Code R :

```
rm(list=ls(all=TRUE)) # Nouveau programme
n=30
X=rnorm(n,0,2)
E=rnorm(n)
```

```
Y=sin(X)+E
# Noyau Normal
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=n^-.2
# Initiation
s=100 # taille de l'intervalle [a,b]
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=numeric(n)
fn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
# Fonction Hn(.)
W=numeric(n)
Hn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] }
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
# Graphes
x11() # nouvelle fenettre graphique
op=par(mfrow=c(1,3))
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=30",type='l',col=4, lwd= 2)
lines(x,sin(x),lwd= 2)
```

```
#####Pour n =120
n=120
X=rnorm(n,0,2)
E=rnorm(n)
Y=sin(X)+E
h=n^-.2
V=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
W=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] }
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=120",type='l',col=4, lwd= 2)
lines(x,sin(x),lwd= 2)
#####Pour n =400
n=400
X=rnorm(n,0,2)
E=rnorm(n)
Y=sin(X)+E
h=n^-.2
V=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
```

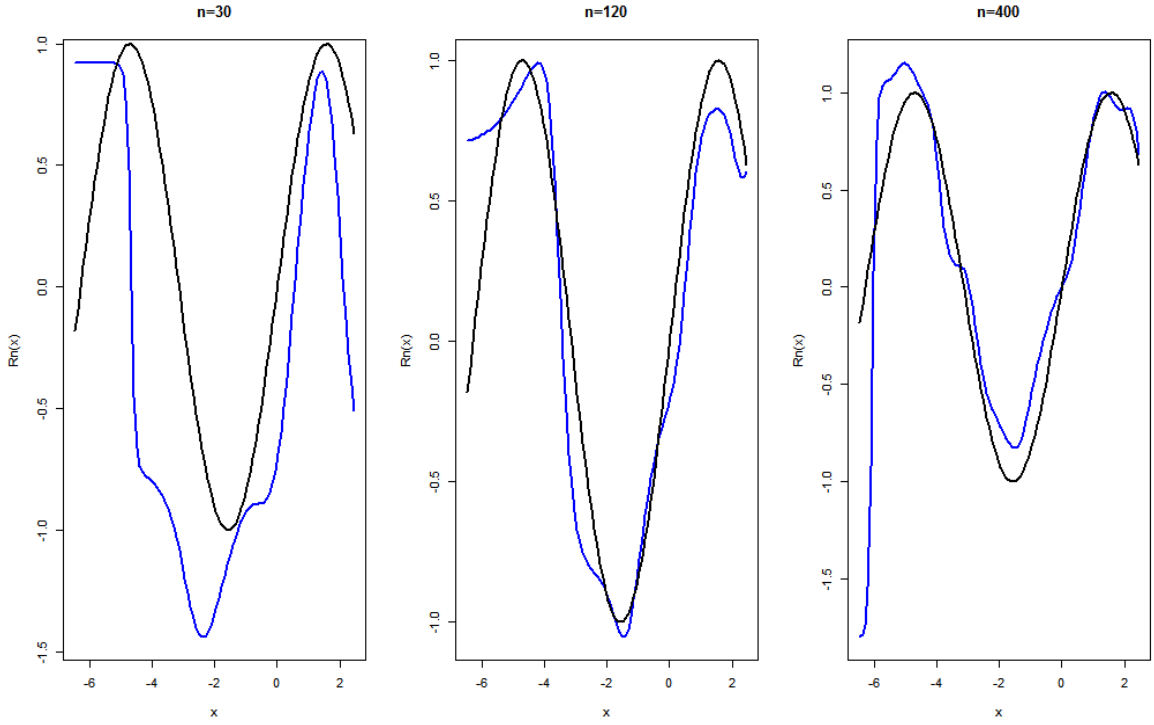

FIG. 3.7 – Régression non linéaire : h fixé, n variée et K noyau gaussien

```
fn[j]=sum(V)/(n*h)}
W=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] }
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=400",type='l',col=4, lwd= 2)
lines(x,sin(x),lwd= 2)
par(op)
```

Remarque 3.3.1 *On remarque la même conclusion pour le cas non linéaire que le cas linéaire (i.e ; convergence de l'estimateur pour n assez grand).*

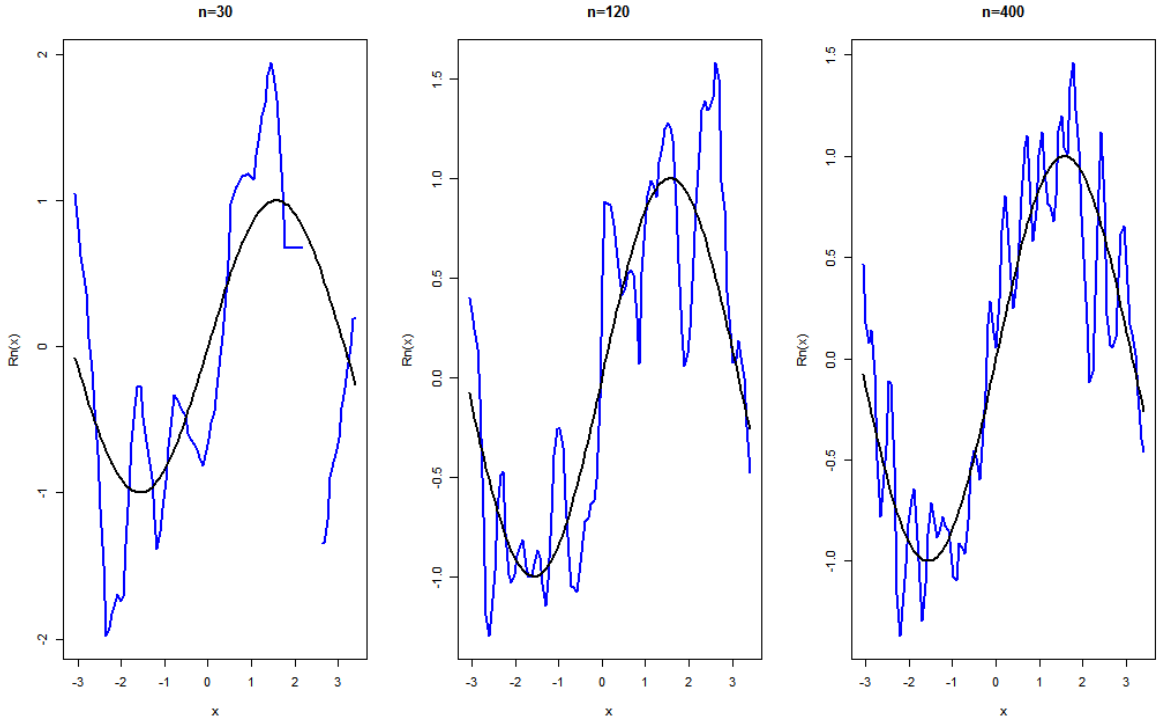

FIG. 3.8 – Régression non linéaire : h fixé, n variée et K noyau Triangulaire

– **K noyau Triangulaire (à support compact) :**

Dans ce second cas, on choisit le noyau Triangulaire: $K(t) = (1 - |t|)1_{[1;1[}(x)$ En suite, on modifie seulement cette partie dans le programme R précédent :

Noyau Triangulaire K(t) **:**

```
K=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
```

On obtient la figure [3.8] ,

Remarque 3.3.2 *on arrive au même conclusion de la convergence de l'estimateur (voir la figure [3.7], i,e ; convergence de l'estimateur pour n assez grand).*

– **K noyau d'Epanechnikov (à support compact) :**

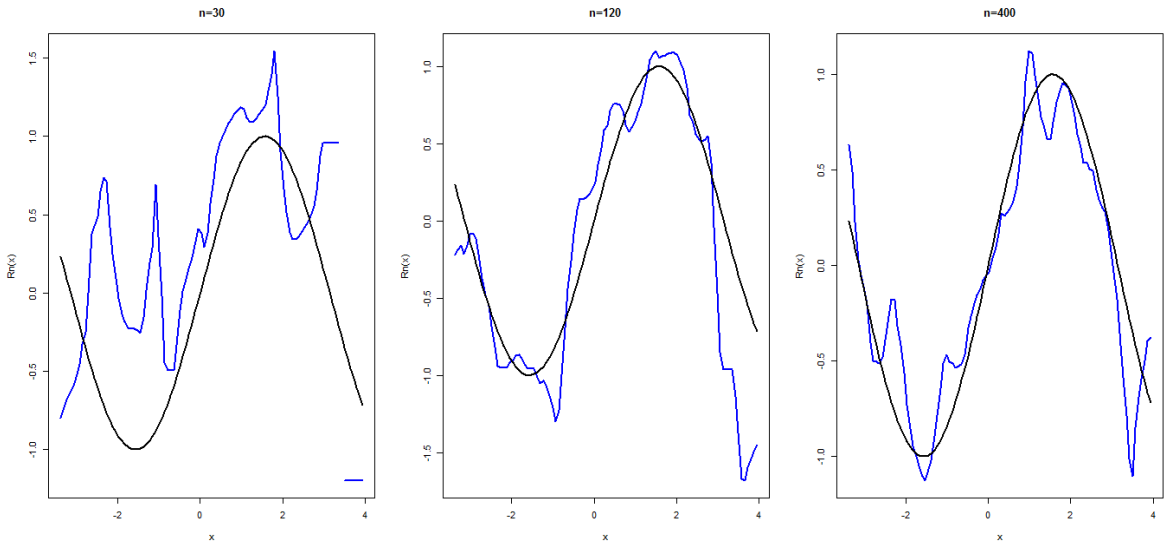

FIG. 3.9 – Régression non linéaire : h fixé, n variée et K noyau d'Epanechnikov.

Dans ce second cas, on choisit le noyau d'Epanechnikov : $K(t) = \frac{3}{4}(1-t^2)\,1_{[1;1]}(x)$

En suite, on modifie seulement cette partie dans le programme R précédent :

Noyau Epanechnikov K(t)

```
K=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}\textbf{ }
```

On obtient la figure [3.9] , on arrive au même conclusion de la convergence de l'estimateur (voir la figure [3.8], i,e ; convergence de l'estimateur pour n assez grand).

3.3.2 Choix graphique du paramètre de lissage

Dans cette partie, on va prendre le paramètre de lissage dans l'intervalle $[0;1]$ de même façon pour la régression linéaire, et avec des tests graphiques en va déterminer le paramètre h optimal (au sens graphique).

On fixée la taille de l'échantillon $n = 300$ et le noyau K est normal, l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure [3.10].

Code R :

```
n=300
X=rnorm(n,0,2)
Y=sin(X)+E
# Noyau Normal K(t)
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=seq(.1,.9,length=9)
# Initiation
s=100 # taille de l'intervalle [a,b]
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=array(dim=c(n,s,9))
fn=array(dim=c(s,9))
W=array(dim=c(n,s,9))
Hn=array(dim=c(s,9))
# density fn(x)
for(k in 1 :9){
for(j in 1 :s){
for(i in 1 :n){ V[i,j,k]=K((x[j]-X[i])/h[k]) }
fn[j,k]=sum(V[,j,k])/(n*h[k])}}
# fonction Hn(x)
for(k in 1 :9){
for(j in 1 :s){
for(i in 1 :n){ W[i,j,k]=K((x[j]-X[i])/h[k])*Y[i] }
Hn[j,k]=sum(W[,j,k])/(n*h[k])}}
```

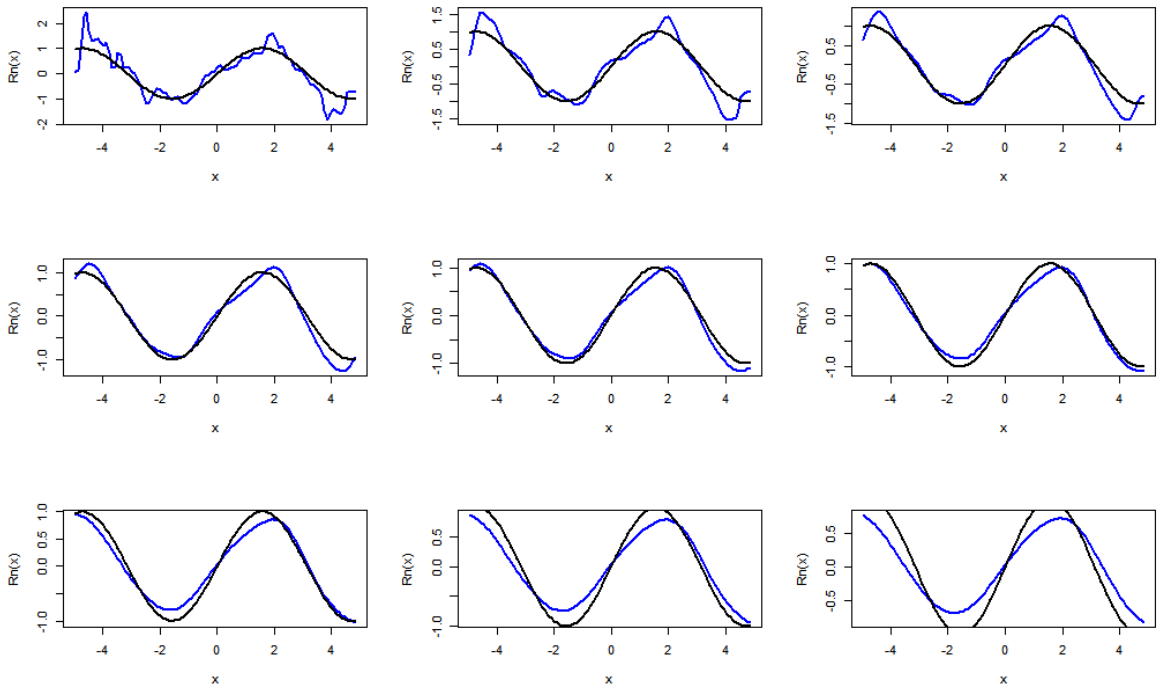

FIG. 3.10 – Régression non linéaire avec h varié, n fixée et K gaussien

```
Rn=array(dim=c(s,9))
for(k in 1 :9){ Rn[,k]=Hn[,k]/fn[,k]}
# Graphes
x11() # nouvelle fenettre graphique
op=par(mfrow=c(3,3))
for(k in 1 :9){
plot(x,Rn[,k],xlab="x", ylab="Rn(x)", main=" ",type='l',col=4, lwd= 2)
lines(x,sin(x),lwd= 2)
}
par(op)
```

Il est clair que la valeur du h optimale est de$h = 0.5$ (ligne2 , colonne 2).

– **K à support compact (Triangulaire)**

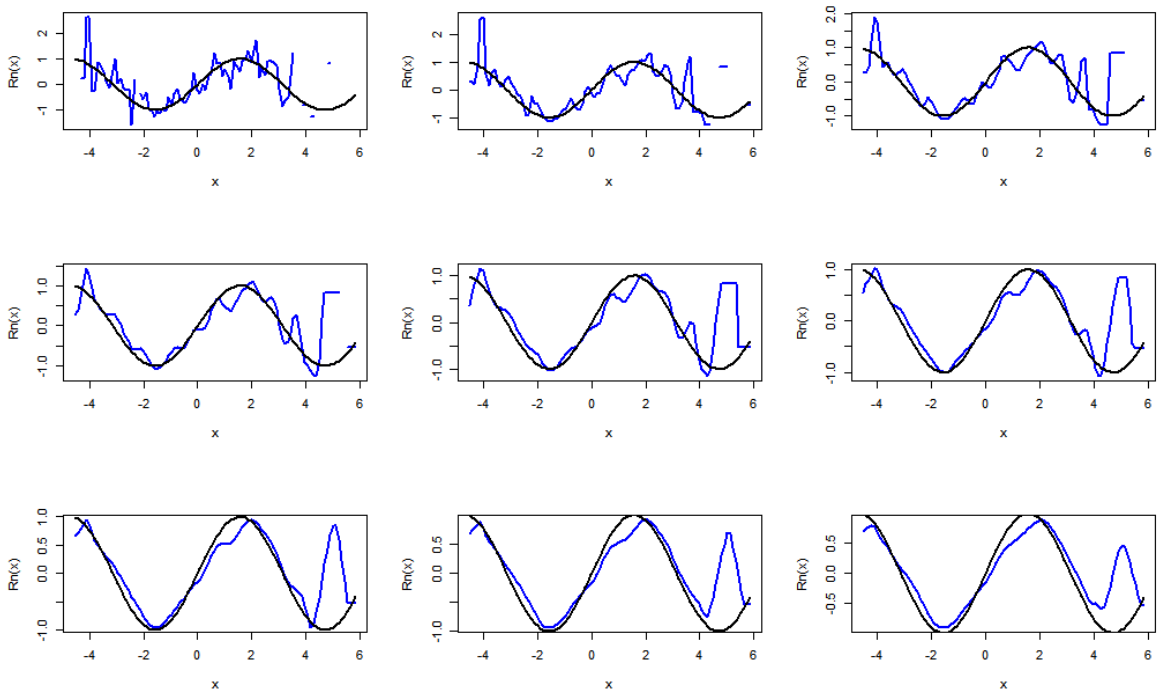

FIG. 3.11 – Régression non linéaire avec h varié, n fixée et K Triangulaire.

Si nous gardons le même modèle non linéaire $y = sinx + \varepsilon$; mais avec le noyau Triangulaire. On note, que la valeur du h optimale est de $h = 0.9$ (ligne 3 ; colonne 3) ; voir la [3.11]

– **K à support compact (noyau d'Epanechnikov)**

Si nous gardons le même modèle non linéaire $y = sinx + \varepsilon$; mais avec le noyau d'Epanechnikov. On note, que la valeur du h optimale est $h = 0.9$ (ligne 3 ; colonne 3) ; voir la figure [3.12] *En conclusion, ce chapitre montre l'importance de*

paramètre de lissage h et du noyau K dans l'estimation non paramétrique de la régression linéaire et non linéaire. Mais à noter que le choix de h est plus crucial que le choix de noyau.

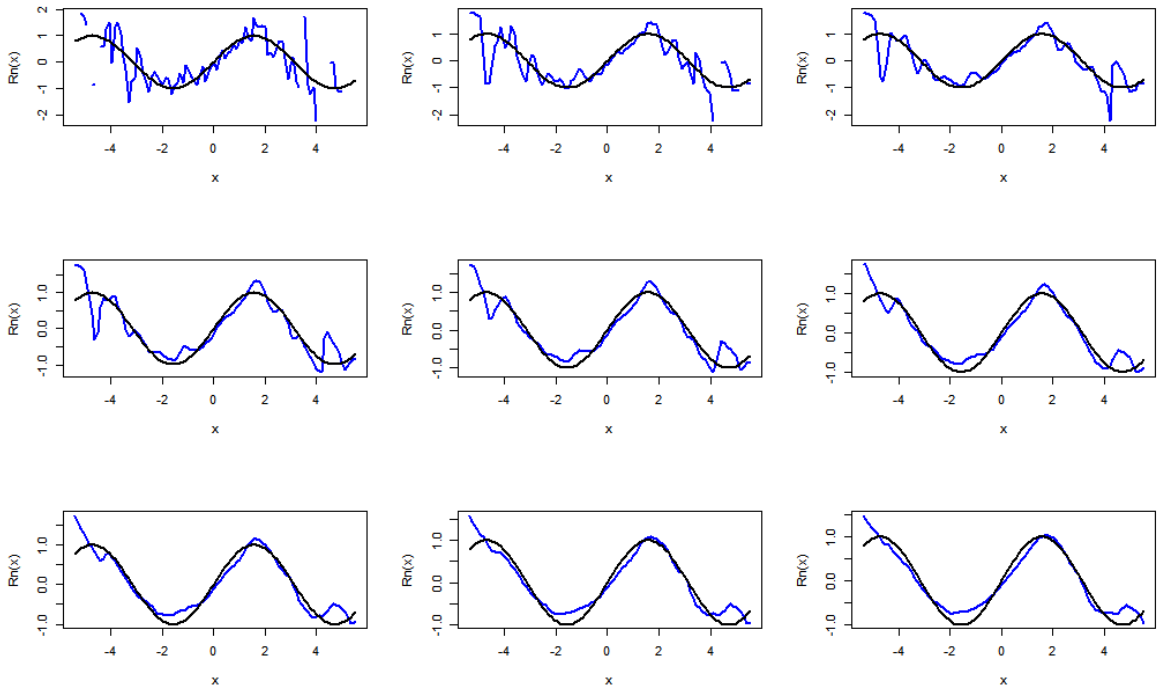

FIG. 3.12 – Régression non linéaire avec h varié, n fixée et K d'Epanechnikov

Conclusion

Dans ce mémoire, on a présenté une méthode d'estimation non paramétrique pour estimer la fonction de régression , c'est la méthode du noyau. Elle est basée sur le noyau K et le paramètre de lissage h. On a montré que la méthode d'estimation de régression non paramétrique est assez simples et peuve être très utile dans plusieurs situations. Par exemple, dans un contexte d'analyse des données, lorsque l'on désire comprendre et observer les relations qui existent entre les variables..

Dans la régression non paramétrique, la méthode du noyau joue un grand rôle. Pour que son soit plus utilisée par les praticiens, il est nécessaire que les programmes informatiques permettant d'appliquer ces méthodes soient facilement accessibles et assez simples d'utilisation. Cela favorise aussi les échanges entre statisticiens et utilisateur.

Pour confirmer notre étude, nous avons fait des simulations des données par le logiciel R qui a effectivement validé nos résultats. A travers les résultats obtenus, nous concluons que : le noyau K est peu influent sur l'estimateur, par contre le paramètre h joue un rôle très important et dont le choix est crucial. Les cas des données incomplètes : tronquées ou censurées est intéressant pour une étude future.(voir ould-saïd (2006))

Bibliographie

[1] Beghriche, H., & Messaci, F. Sur l'estimation de la fonction de régression (Doctoral dissertation, Constantine : Université Mentouri Constantine).

[2] Ben Khalifa, I, 2007. Estimation non-paramétrique par noyaux associés et données de panel en marketing, université du 7 Novembre à Carthage. Tunisie

[3] Blondin, D. (2004). Lois limites uniformes et estimation non-paramétrique de la régression (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).

[4] Bochner, S. (1946). Vector fields and Ricci curvature. Bulletin of the American Mathematical Society, 52(9), 776-797.

[5] Bosq D. (1998). Nonparametric Statistics for Stochastic Processes. Springer, New York, Berlin, Heidelberg.

[6] Carbonez, A., Györfi, L., & van der Meulen, E. C. (1995). Partitioning-estimates of a regression function under random censoring. Statistics & Risk Modeling, 13(1), 21-38.

[7] Collomb G. (1977). Quelques Propriétés de la Méthode du noyau versez l'estimation non paramétrique de la-régression en Point Fixe des Nations Unies., CR Acad. Sc. Paris 285 : 289-92.

[8] Collomb, G. (1981). Estimation non paramétrique de la régression : Revue bibliographique, ISI 49 : 75-93

[9] Demir, S., & Toktamiş, o. (2010). On the adaptive Nadaraya-Watson kernel regression estimators. Hacettepe Journal of Mathematics and Statistics, 39(3), 429-437.

[10] Devroye, L., Györfi, L., (1985) *Nonparametric density estimation. The L1 view*. Wiley, New York.

[11] Epanechnikov, V.A. (1969) Nonparametric estimation of a multivariate probability density. *Theory Probab. Appl.* 14, 153-158.

[12] Eubank, R. L. (1988). Spline smoothing and nonparametric regression (No. 04 ; QA278. 2, E8.).

[13] Eubank, R. L. (1999). Nonparametric regression and spline smoothing. CRC press.

[14] Kadi, A., & Zelliche, W. (2019). Analyse de la régression non paramétrique (Doctoral dissertation, University of Jijel).

[15] Kohler M., Mathé K., Pintér M. (2002). Prediction from randomly right censored data. Journal of Multivariate Analysis, 80 : 73–100.

[16] Lejeune, M. (2004). Statistique : La théorie et ses applications. Springer Science & Business Media

[17] Nadaraya, E.A. (1964). On estimating regression. Theory Probab. Appl. 9 : 141–142.

[18] Nadaraya, E. A. (1989). Nonparametric estimation of probability densities and regression curves. Dordrecht : Kluwer Academic Publishers.

[19] Ould-Saïd, E., & Lemdani, M. (2006). Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. Annals of the Institute of Statistical Mathematics, 58(2), 357-378.

[20] Parzen, E. (1962) On estimation of a probability density function and mode, *Ann. Math. Stat.* 33 : 1065-1076.

[21] Rosenblatt, F. (1962). Principles of Neurodynamics Spartan. New York, 10, 318-362.

[22] Rao P. (1983). Nonparametric Functional Estimation. Academic Press, Inc., London.

[23] Schuster, E.F. (1972), Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. The Annals of Mathematical Statistics, 43(1) : 84–88.

[24] Silverman B.W. (1986). Density Estimation. London : Chapman and Hall.

[25] Tsybakov, A. B. (2003). Introduction à l'estimation non paramétrique (Vol. 41). Springer Science & Business Media.

[26] Wasserman L. (2005). All of Statistics : A Concise Course in Statistical Inference, Springer Texts in Statistics.

[27] Watson, G. S. (1964). Smooth regression analysis. Sankhyā : The Indian Journal of Statistics, Series A, 359-372.

[28] https ://www.ceremade.dauphine.fr/~turinici/images/stories/work/cours/M1_stat_nonpara

[29] http ://www.univ-bejaia.dz/xmlui/bitstream/handle/123456789/7299/Estimation%20non%2