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

Option : **Statistique**

Par

KHADRAOUI Romaissa

Titre :

Estimation à noyau de la fonction de régression

Membres du Comité d'Examen :

Dr.	SAYAH Abdallah	M.C.A.	UMKB	Président
Dr.	KHEIREDDINE Souraya	M.C.B.	UMKB	Encadreur
Dr.	SOLTANE Louiza	M.C.B.	UMKB	Examinateur

Septembre 2020

Dédicace

Je dédie ce travail à mon père adoré

à ma chère mère

à mes soeurs

à mes amies

et à toute ma famille.

Romaissa

REMERCIEMENTS

Tout d'abord je tiens à remercier Dieu de m'avoir donné le courage, le morale et la santé pour mener à bien ce travail.

Je remercie mon encadreur Dr. Kheireddine Souraya pour sa disponibilité, son soutien et ses remarques précieuses qui m'ont aidé à bien présenter ce travail.

Mes remerciements s'adressent aussi à Dr. Sayah Abdallah d'avoir accepté de présider mon jury de soutenance. Je suis également très reconnaissante à Dr. Soltane Louiza pour son soutien et sa gentillesse et aussi d'avoir accepté d'examiner ce travail.

Je tiens à remercier toute ma famille, mes amies et mes condisciples de la promotion 2020.

Enfin, je remercie chaleureusement toutes personnes qui m'ont aidé, et qui ont contribué de proche ou de loin à la réalisation de ce travail.

Romaissa

Résumé

Dans ce mémoire nous étudions l'estimateur non paramétrique de la fonction de régression. La construction de l'estimateur est basée sur l'utilisation d'une densité K(.) appelée noyau et d'un paramétre de lissage h. Nous rappelons les propriétés de l'estimateur : convergence faible et forte,.... Nous parlons aussi du choix de noyau et de paramètre de lissage. 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.**

Table des matières

Table des figures

Introductions

La statistique fonctionnelles constitue un champ de recherches d'actualité,à la fois diversifie par ses aspects fondamentaux et par les différents domaines qu'elle recoupe : statistique non-paramétrique, statistique des opérateurs, variables et /ou modèles fonctionnels .Sans prétendre à l'exhaustivité, l'objectif de cette mémoire est de présenter quelques uns de ces aspects fonctionnels de la statistique en nous contonnant autour des modèles non paramétriques de régression.

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 à 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, etet 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 travail est subdivisé en trois chapitres :

Dans le premier chapitre nous introduisons la défintion de l'estimation fonctionnelle et l'estimateur à

noyau de la densité (Rosenblatt, 1956 et Parzen, 1962*).*

Dans le deuxiéme chapitre, nous présentons l'estimation non paramétrique de la régression et les propriétés de l'estimateur. Nous étudions ici, la normalité asymptotique de l'estimateur et choix du noyau et de la largeur de fenêtre.

Dans le troisiéme chapitre où nous donnons des exemples par simulation qui expriment l'importance de paramètre de lissage h, la taille de l'échantillon utilisé et le noyau K dans l'estimation non paramétrique de la fonction de régression. Les chapitres un et trois comprennent des simulations effectuées en utilisant le Logiciel ***R*** *(****Annexe A****). Les codes* ***R*** *utilisés sont donnés avec les sorties graphiques correspondantes.*

*Finalement, Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées dans l'***annexe B**

Chapitre 1

Estimation fonctionnelle

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

1.1 Estimation paramétrique et non paramétrique

Premièrement, nous appelons modèle statistique, le triplet $(E, \mathcal{A}, \mathbb{P})$ où E est l'éspace des observations (par exemples des réels), $\mathcal{A}$ une tribu sur E et P une famille de probabilité sur $(E, \mathcal{A})$ .Soit $X : \Omega \to E$ 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 E. Donc, le paramètre à estimer est l'application

$$h : P \to \Theta' \qquad \text{ou} \quad \Theta \to \Theta'$$
$$\theta \mapsto h(P_\theta)$$

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

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 $\Theta \subset \mathbb{R}^s$ et $h(.,.)$ est une fonction connue, on parle alors d'estimation paramétrique, car éstimer h revient à éstimer le paramétre fini-dimensionnel θ.

1.1.2 Estimation non paramétrique

Par contre, si l'on sait seulement que h appartient à $\mathbb{P}$ ensemble des lois de probabilités qui est un éspace de dimension infinie, alors on dit que l'on fait de l'estimation non paramétrique ou de 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 β. De plus,on suppose que les $\{X_i,\ i = 1....., n\}$ sont indépendantes et identiquemment distribuées $(i.i.d)$ $\mu \in \rho_0$ une famille de loi sur $(\mathbb{R}^s, \beta)$;

La fonction de répartition :

Soit X une variable aléatoire à valeurs dans un intervalle I de la forme $[a;b]$ qui suit une loi de probabilité P, on appelle fonction de répartition de X la fonction F, $F(X) = P(X \leq x) = \int_a^x f(t)\,dt$, soit $(X_1, X_2, ...X_n)$ un échantillon de taille n observation issue de même loi de probabilité de densité f

$$F_n(X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} I_{(X_i \leq x)}(x) = \frac{\sharp\{X_i \leq x\}}{n}$$

La densité de probabilité :

Une loi à densité est loi continue on appelle densité de probabilité sur un intervalle I toute fonction f continue et postive telle que :

- $\int_I f(t)\,dt = 1$
- $f(x) = \frac{\partial F(x)}{\partial x}$

La fonction des quantiles :

Pour $s = 1$,la fonction quantile d'ordreρ definie par

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

F_{μ}^{-1}est un paramétre à valeur dans l'espace de fonction réelles définies sur $]0; 1[$ monotonses non décroissantes et continues à gauche.

La fonction caractéristique :

Elle est définie par

$$\hat{\mu}(t) = E_{\mu}[\exp\{i \langle t, x \rangle\} \text{ où } t, x \in \mathbb{R}^s$$

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

Le paramètre 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 $\mu_y^x, x \in \mathbb{R}^{s_1}$une famille de versions des lois conditionnelles de Y sachant $X = x$. Toute fonction de la forme $r : x \to r(\mu_y^x)$ est un paramétre de régression,.les plus usuels sont :

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

2 La densité conditionnelle

3 La fonction de répartition conditionnelle

4 Le quantile conditionnelle

1.2 Estimation non paramétrique de la densité

On peut justifier la construction de l'estimateur à noyau de deux façons, une premiére ideé (développée par Rosenblatt en 1956) va être de le construire à partir de l'estimateur de la fonction de répartition F,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(X_i \leq x)}$$

$$\forall h > 0, F(x+h) - F(x-h) = P(x-h \leq X \leq x+h) = \int_{]x-h,x+h[} f(y)dy$$

soit x une v.a de densité f, $(X_1, X_2, ..., Xn)$ un échantillon iid issu de x

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

comme $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(X_i \leq x)} \to_{n\to\infty} F(x)$, F_n est un estimateur fort de F soit $h \to 0$

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{2h}\left\{\hat{F}_n(x+h) - \hat{F}_n(x-h)\right\} \\
&= \frac{1}{2h}\left\{\frac{1}{n}\sum_{i=1}^{n}(I_{(X_I \leq x+h)} - I_{(X_i \leq x-h)}\right\} \\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}(I_{(\frac{X_i+x}{h} \leq 1)} - I_{(\frac{X_i-x}{h} \leq -1)}) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}I_{(-1 \leq \frac{X_i-x}{h} \leq 1)}
\end{aligned}$$

Donc

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) \tag{1.1}$$

on a : $K_0(t) = \frac{1}{2}I_{(|t| \leq 1)}$, K_0

Parzen (1962) a suggéré une généralisation de cet estimateur

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{1.2}$$

où $K : \mathbb{R}^s \to \mathbb{R}$ est une fonction intégrable t.q. $\int K(u)\,du = 1$.

C'est l'estimateur à noyau de la densité ou éstimateur de Parzen-Rosenblatt. La fonction K est dite

noyau et le paramétre h fenêtre (en englais bindwight).

Exemple 1.2.1 *La figure si-dessous (1.1), présente l'allure (en noire) d'une densité d'une variable aléatoire de loi log-normale contrée de variance* $\frac{1}{2}$:

$$f(x) = \frac{1}{\sqrt{x\pi}} \exp\left(-2\log x^2\right), \qquad x > 0.$$

la courbe en rouge est celle de l'estimateur à noyau de la densité (1.2), calculer pour $n = 300$, $h = 0.07$ *et* $K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right)$ *(i.e., un noyau gaussien).*

Code R utilisé :

```
n=300;m=100
X=rlnorm(n,0,.5);h=1.26*n^-.2
K=function(t){dnorm(t)}
f=rep(0,100)
x=seq(min(X),max(X),length=m)
for(i in 1:m){f[i]=sum(K((x[i]-X[1:n])/h))/(n*h)}
plot(x,dlnorm(x,0,.5),type='s',ylab="f(x)")
lines(x,f,col=2)
legend(1.5,.6,c("densite theorique","densite empirique"),col=c(1,2), lty=c(1,2),
 lwd=c(2,2), bty = "n", cex=1)
```

Propriété 1.2.1

1. K est borné
2. $\int K(t)dt = 1$
3. $\int tK(t)dt = 0$
4. $\int t^2 |K(t)| < \infty$

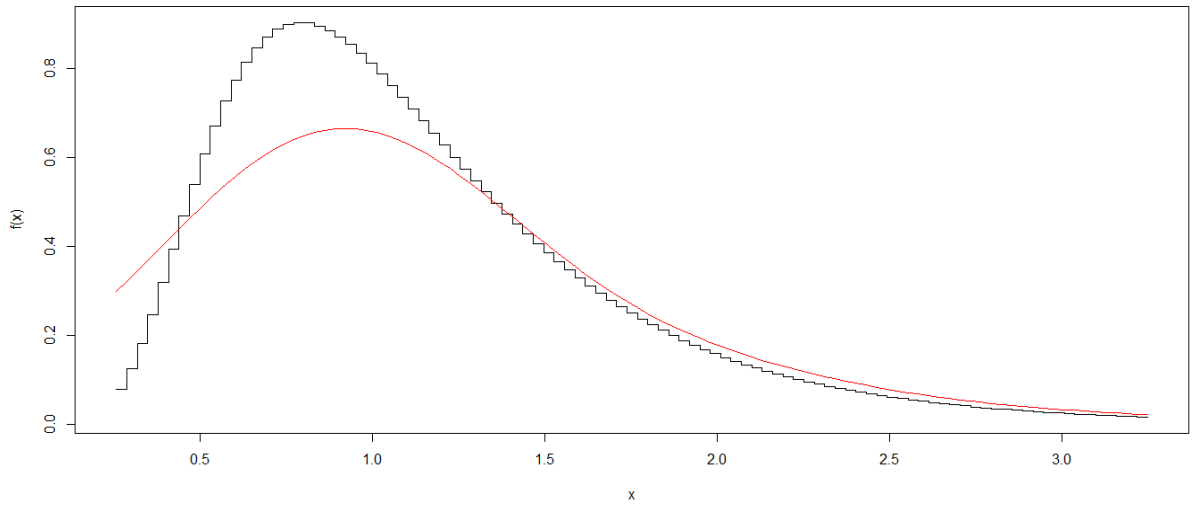

FIG. 1.1 – Estimateur à noyau d'une densité d'une v.a. lognormal.

5. $\int K^2(t)dt < \infty$

Quelques exemples de noyaux classiques :

Le tableau suivant donne l'efficacitée relative des noyaux les plus utilisées dans l'estimation de la densitée de probabilitée. Comme ces efficacitées sont trèes rapprochéees, en pratique nous choisissons K en tenant compte de la facilitée des calculs plutôt que de l'efficacité relative.

Noyaux	support	Densité	Efficacité relative
Noyau d'Epanechnikov	$[-1,1]$	$K(t) = \frac{3}{4}(1-t^2)I_{(\lvert t\rvert \prec 1)} \forall t \in \mathbb{R}$	1
Noyau Cosinus	$[-1,1]$	$K(t) = \frac{\pi}{4} cos(\frac{\pi}{2}x)$	0.999
Noyau biweight	$[-1,1]$	$K(t) = \frac{15}{16}(1-t^2)^2 I_{(\lvert t\rvert \prec 1)} \forall t \in \mathbb{R}$	0.994
Noyau triongulaire	$[-1,1]$	$K(t) = (1-\lvert t\rvert) I_{(\lvert t\rvert \prec 1)} \forall t \in \mathbb{R}$	0.986
Noyau Gaussien	$\mathbb{R}$	$K(t) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-t^2}{2}) \forall t \in \mathbb{R}$	0.946
Noyau rectongulaire(uniforme)	$[-1,1]$	$K(t) = \frac{1}{2} I_{(\lvert t\rvert \prec 1)} \forall t \in \mathbb{R}$	0.930

Code R utilisé :

```
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)
```

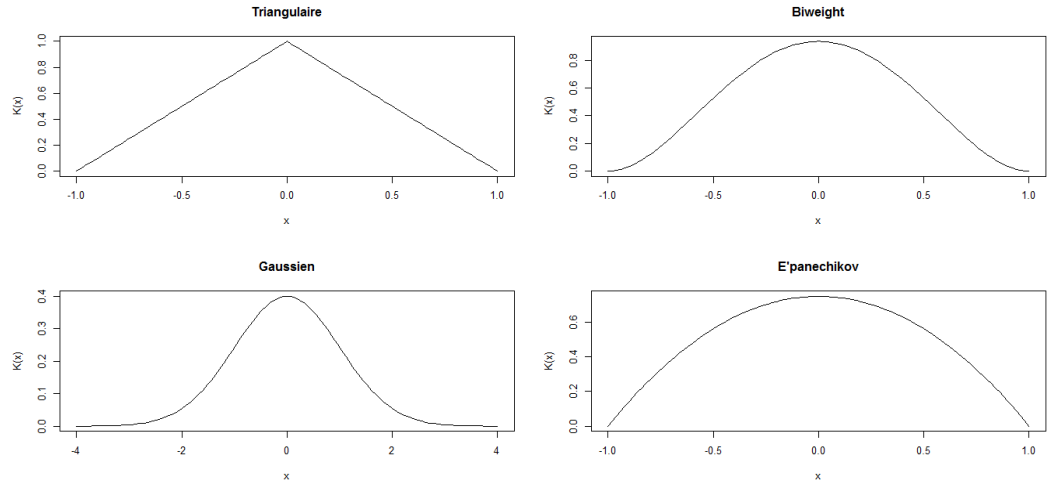

FIG. 1.2 – Allures des noyaux : Triangulaire, Biweight, Gaussien, Epanechnikov.

Théorème 1.2.1 *si K est positive et $\int_{\mathbb{R}} K(u) = 1$ alors $\hat{f}_n$ est une densité de probabilié .De plus $\hat{f}_n$ est continue si K est continue*

Preuve. L'estimateur à noyau est positive et continue car la somme des fonctions positives et continues

est elle même une fonction positive et continue .Il faut donc verifier que l'intègrale de $\hat{f}_n(x)$ vaut

$$\begin{aligned}\int_{\mathbb{R}} \hat{f}_n(x)\,dx &= \int_{\mathbb{R}} \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{\mathbb{R}} K(\frac{X_i - x}{h}) \quad \text{on.pose} \quad (u = \frac{X_i - x}{h}) \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{\mathbb{R}} K(u)h\ du \\ &= 1\end{aligned}$$

■

Notons par $*$ le produit de convolution i.e :

$$f * g(x) = \int f(y)\, g(x-y)\, dy = \int g(y)\, f(x-y)\, dy.$$

Définition 1.2.1 *Un noyau est dit de Parzen-Rosenblatt si*

$$\lim_{\|x\|\to\infty} \|x\|^s K(x) = 0, \qquad \|.\| \text{ est la norme euclidienne.}$$

Nous terminons cette section, en donnant un resultat fondamental qui exprime le fait que, lorsque h est petit, la convolution avec $K_h(.)$ perturbe peu une fonction de L^1. Alors, nous avons :

Lemme 1.2.1 ***(Lemme de Bochner)***

1 *Soit K un noyau de Parzen-Rosenblatt et $f \in L^1$ alors en tout point x de continuité 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*

$$\lim_{h\to 0}\sup_{x\in\mathbb{R}} \mid f * K_h(x) - f(x) \mid = 0$$

Maintenant, nous donnons l'inégalité suivante :

Théorème 1.2.2 ***(Inégalité de Bernstien-Frechet)*** *Soit ξ_i une suite de v.a. i.i.d, $\exists c > 0$ tq*

$$\forall k > 2 : E\left|\xi_i - E\left(\xi_i\right)\right|^k \leq c^{k-2}\ k! var\xi_i,\ \forall \lambda > 0,$$

Alors

$$P\left(|\sum_{i=1}^{n}\xi_i - E\left(\xi_i\right)| > \lambda\right) \leq 2\exp\left\{-\lambda^2(4\sum_{i=1}^{n}var\xi_i + 2c\lambda)^{-1}\right\}.$$

1.3 Théorèmes de convergences de variables aléatoires

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

1. **Convergence presque sûre :** Une suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$, définie sur $(\Omega, \mathcal{A}, P)$, converge presque sûrement (p.s.) vers la variable aléatoire X, définie sur $(\Omega, \mathcal{A}, P)$, si

   $$P\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = 1$$

   Dans ce cas, on note $\lim_{n\to\infty} X_n = X$ p.s. ou $X_n \to X$ p.s. lorsque $n \to \infty$.

2. **Convergence en probabilité :** Soient $X_n, n \in \mathbb{N}$, X, des variables aléatoires réelles sur $(\Omega, \mathcal{A}, \mathbb{P})$. On dit que X_n converge en probabilité vers X, et on note $X_n \xrightarrow{P} X$, ou $\lim_{n\to\infty} X_n = X$ en probabilité, ou $P-\lim_{n\to\infty} X_n = X$, si pour tout $\varepsilon > 0$,

   $$\lim_{n\to\infty} \mathbb{P}\left\{|X_n - X| \geq \varepsilon\right\} = 0$$

3. **Convergence dans L^p** : Soient, $X_n, n \in \mathbb{N}$, X, des variables aléatoires réelles dans $L^p(\Omega, \mathcal{A}, \mathbb{P}), 0 < p < \infty$. On dit que X_n converge vers X dans L^p si $\lim_{n\to\infty} \|X_n - X\|_p = 0$, ou de façon équivalente, $\lim_{n\to\infty} \mathbb{E}\left(|X_n - X|_p\right) = 0$.

4. **Convergence en loi :** Soient $X_n, n \in \mathbb{N}$ et X, des variables aléatoires réelles, définies sur $(\Omega, \mathcal{A}, P)$.

   On dit que X_n converge en loi vers X, ou que les lois P^{X_n} convergent étroitement vers la loi P^X, si l'une des quatre conditions équivalentes suivantes est vérifiée :

   i) $\lim_{n\to\infty} F^{X_n}(t) = F^X(t)$ en tout point de continuité t de F^X ;

   ii) $\lim_{n\to\infty} \int \phi(X_n)dP = \int \phi(X)dP$ pour toute fonction continue bornée $\phi : R \to R$;

   iii) $\lim_{n\to\infty} \varphi^{X_n}(t) = \varphi^X(t)$ pour tout $t \in R$;

   iv) Il existe un espace probabilisé $(\Omega', \mathcal{A}', P')$ sur lequel sont définies des variables aléatoires $X'_n, n \in \mathbb{N}$ et X' , telles que X_n et X'_n ont même loi pour tout n, X et X' ont même loi, et $\lim_{n\to\infty} X'_n = X'$ p.s.

   On note alors $X_n \xrightarrow{L} X$ ou $X_n \xrightarrow{d} X$ (X_n converge «en distribution» vers X).

5. **Convergence en moyenne quadratique :** Une suite de v.a.r. $(X_n)_{n\in\mathbb{N}}$ converge en moyenne quadratique vers une v.a.r. X si

$$\lim_{n\to\infty} \left(\mathbb{E}\left(X_n(\omega) - X(\omega)\right)^2\right) = 0,$$

   et on note dans ce cas : $X_n \xrightarrow{m.q} X$.

Chapitre 2

L'estimateur à noyau 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 choix du noyau et le paramétre de lissage.

2.1 Définition

On suppose que l'on a observé un échantillon $\{(X_i, Y_i), i = 1, ..., n\}$et on veut expliquer la variable aléatoire Y_i par $X_i$et nous considérons le modéle de régression non paramétrique donné pour $i = 1, ..., n$

$$Y_i = r(X_i) + \epsilon_i$$

où ϵ_i est l'aléatoire centré et indépendant de $X_i$et r est une application mesurable réelle.

La fonction de régression $r(.) = E[Y / X = .]$, apportant de l'information sur la relation de dépendance inconnue de Y et X ; un probléme important est l'estimation de r à partir de l'observation de n copies (X_i, Y_i), $i = 1, ..., n$ qui suivent la même loi que(X_i, Y_i)

Suppose que (X, Y)a une densité $f : (x, y) \longrightarrow f(x, y)$sur $\mathbb{R}^2$et que

$f_X : x \longrightarrow f_X(x) = \int f(x,y)dy > 0$ (densité de X)

$$\forall x \in \mathbb{R}, r(x) = E\left[Y \diagup X = x\right] = \frac{\int y f(x,y)\partial y}{f_X(x)}$$

Les densités f et f_X sont inconnues mais on peut les estimer via $\forall (x,y) \in \mathbb{R}^2$

$$f_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{X_i - x}{h_n})K(\frac{Y_i - y}{h_n})$$

$$f_{n,X}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{X_i - x}{h_n})$$

puis on considère l'estimateur de la régression.

$$\forall x \in \mathbb{R}, r_n(x) = \frac{\int y f_n(x,y)\partial y}{f_{nX}(x)} I_{(f_{n,X}(x)\neq 0)} \tag{2.1}$$

Définition 2.1.1 *Si K est un noyau d'ordre 1 l'estimateur défini par (2.1)verifié*

$$\begin{aligned}\forall x \in \mathbb{R}, r_n(x) &= \frac{\sum Y_i K(\frac{X_i - x}{h_n})}{\sum K(\frac{X_i - x}{h_n})} \\ &= \frac{\sum Y_i K_{h_n}(X_i - x)}{\sum K_{h_n}(X_i - x)}\end{aligned}$$

où $K_{h_n}(.) = K(./h_n)$,donc l'estimateur à noyau de régression est donné

$$r_n(x) = \frac{\sum Y_i K_{h_n}(X_i - x)}{\sum K_{h_n}(X_i - x)} = \frac{g_{n,X}(x)}{f_{n,X}(x)} \tag{2.2}$$

C'est l'estimateur à noyau introduit par Nadaraya -Watson (Nadaraya,1964 et Watson,1964)

2.2 Propriétés asymptotiques de l'estimateur

Nous étudions dans cette partie deux modes de convergence, la convergence en moyenne quadratique et la convergence presque compléte. nous supposons que K est un noyau vérifant les condition suivantes :

(H.1) K est borneé ,c'est à dire $\sup_{x\in\mathbb{R}} \mid K(x) \mid \partial x < \infty$

(H.2) $\lim \mid x \mid K(x) \longrightarrow 0$,quand $\mid x \mid \longrightarrow +\infty$

(H.3) $k \in L_1(\mathbb{R})$,c'est à dire $\int_{\mathbb{R}} \mid K(x) \mid \partial x < +\infty$

(H.4) $\int_{\mathbb{R}} \mid K(x) \mid dx = 1$

(H.5) $\int_{\mathbb{R}} u\ K(u)\partial u = 0$

(H.6) $\int_{\mathbb{R}} u^2\ K(u)\partial u < +\infty$

(H.7) K est bornée ,intégrable et à support compact.

2.2.1 Etude asymptotique de biais

L'etude asymptotique du biais repose sur la proposition suivante :

Proposition 2.2.1 *i) Si* $\mid Y \mid \leq C_1 < \infty$ *P.s et* $nh_n \longrightarrow \infty$,*quand* $n \longrightarrow \infty$, *alors :*

$$E\left[r_n(x)\right] = \frac{E\left[g_{n,X}(x)\right]}{E\left[f_{n,X}(x)\right]} + O(\frac{1}{nh_n})$$

ii) Si $EY^2 < \infty, nh_n^2 \longrightarrow \infty$, *quand* $n \longrightarrow \infty$ *alors :*

$$E\left[r_n(x)\right] = \frac{E\left[g_{n,X}(x)\right]}{E\left[f_{n,X}(x)\right]} + O(\frac{1}{\sqrt{n}h_n})$$

Mantenant nous sommes e mesure d'énoncer le resultat suivant :

Proposition 2.2.2 *Si condition* $(H.4), (H.5)$ *et* $(H.6)$ *sont vérifiées si* $f_X(.)$*et* $r(.)$*sont le classe* $C^2(\mathbb{R})$ *et si* $\mid Y \mid$*est borné. Alors :*

$$E\left[r_n(x) - r(x)\right] = \frac{h_n^2}{2}\left\{\left\{r^{''}(x) + 2r^{'}(x)\frac{f_X^{'}(x)}{f_X(x)}\right\}\int_{\mathbb{R}} u^2K(u)\partial u\right\}(1 + O(1)) \tag{2.3}$$

1. Les condition $(H.4), (H.5)$ *et* $(H.6)$peuvent être remplacées par le noyau K est d'ordre 2 au sens Gasser et Müller
2. dans la relation (**??**) est égle à$O(h) + O((nh)^{-1})$

Preuve.

$$E\left[r_n(x)-r(x)\right]=\left[EK(\frac{x-X}{h_n})\right]^{-1}\left\{\int_{\mathbb{R}}\frac{1}{h_n}K(\frac{x-t}{h_n})g(t)\partial t-r(x)\int_{\mathbb{R}}\frac{1}{h_n}K(\frac{x-t}{h_n})f(t)\partial t\right\}$$
$$=\left\{(f(x))^{-1}\left\{\frac{h_n^2}{2}g^{''}(x)-\frac{h_n^2}{2}r(x)f^{''}(x)\right\}\int_{\mathbb{R}}u^2K(u)\partial u+g(x)-r(x)f(x)\right\}(1+O(1))$$

comme $g(x)=r(x)f(x)$ l'équation précédente peut s'ecrite :

$$E\left[r_n(x)-r(x)\right]=\left\{\frac{h_n^2}{2}\left\{r^{''}(x)+2r^{'}(x)\frac{f_X^{'}(x)}{f_X(x)}\right\}\int_{\mathbb{R}}u^2K(u)\partial u\right\}(1+O(1))$$

D'où

$$\lim_{n\longrightarrow\infty}E\left[r_n(x)\right]=r(x)$$

■

2.2.2 Etude asymptotique de la variance

Proposition 2.2.3 *Sous $E\left(Y^2\right)<\infty$, alors en chaque point de continuité des fonctions $r(x), f_X(x)$ et $\sigma^2(x)=var(Y\diagdown X=x)$*

on a :

$$var\left[r_n(x)\right]=\frac{1}{nh_n}\left\{\frac{\sigma^2(x)}{f_X(x)}\int K^2(u)du\right\}(O(1)+1) \tag{2.4}$$

où $f_X(x)>0$

Preuve. soit la fonction $\psi(x)=\int y^2f(x,y)\partial y$, en se basant sur le lemme de Bochner

$$var\left[g_{n,X}(x)\right]=\frac{1}{nh_n}\left\{E\left[Y^2k^2(\frac{x-X)}{h_n})\right]-\left[EY\ K(\frac{x-X}{h_n})\right]^2\right\}$$
$$=\frac{1}{nh_n}\left\{\int_{\mathbb{R}}K^2(u)\Psi(x-h_nu)\partial u-h_n(\int_{\mathbb{R}}K(u)f(x-uh_n)^2\right\}$$
$$=\frac{1}{nh_n}\Psi(x)\int_{\mathbb{R}}K^2(u)\partial u(1+O(1))$$

$$E\left[\{f_{n,X}(x) - E(f_{n,X}(x))\}\{g_{n,X}(x) - E(g_{n,X}(x)\}\right] = \frac{1}{nh_n} g(x) \int_{\mathbb{R}} K^2(u)\partial u(1 + O(1))$$

et

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

on pose $B_n(x) = \binom{f_{n,X}(x)}{g_{n,X}(x)}$ et $A(x) = (\frac{-r(x)}{[f_X(x)]^2}, \frac{1}{f_X(x)})$

La matrice de variance covariance de $B_n(x)$ est alors donnée par :

$$\sum := \frac{1}{nh_n} \begin{pmatrix} f_X(x) & g(x) \\ g(x) & \Psi(x) \end{pmatrix} \int_{\mathbb{R}} K^2(u)\partial u(1 + O(1))$$

En remarquant que :

$$\begin{aligned} var\left[r_n(x)\right] &= A \sum A^t \\ &= \frac{1}{nh_n} \left(\frac{\Psi(x)}{\mid f_X(x) \mid^2} - \frac{(g(x))^2}{\mid f_X(x) \mid^3} \right) \int_{\mathbb{R}} K^2(u)\partial u(1 + O(1)) \end{aligned}$$

ou A^t designe la transposeé de$A$,on obtient alors :

$$var\left[r_n(x)\right] = \frac{1}{nh_n} \left\{ \frac{\sigma^2(x)}{f_X(x)} \int K^2(u)du \right\} (O(1) + 1)$$

■

2.3 Choix du noyau et paramètre de lissage

Comme nous vous le disons l'estimateur r_n dépend de deux paramétres : le noyau K et la largeur de la fenêtre h.

Etude de critére d'erreur quadratique moyenne de $r_n(x)$

L'erreur quadratique moyenne EQM (en anglais : mean squared error MSE) est une mesure permettant d'évaluer la similarité de r_n par rapport à la fonction de régression inconnue r, au point x.

Notre but est de minimiser

$$MSE(r_n(x)) = E\left[r_n(x) - r(x)\right]^2$$

Le développement de cette expression faite précédemment , nous donne

$$MSE(r_n(x)) = var\left[r_n(x)\right] + \left[biais(r_n(x))\right]^2$$

Nous constatons d'une part que les expressions du biais de $r_n(x)$et de la variance de $r_n(x)$

$$MSE(r_n(x)) = \frac{h_n^4}{4}\left[(r^{\backprime\backprime}(x) + 2r^{\backprime}(x)\frac{f_X^{\backprime}(x)}{f_X(x)})(u^2K(u)) + 0(1)\right]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right] + (1 + O(1)) \quad (2.5)$$

où $[u^pK^q(u)] = \int t^pK^q(t)dt$

Pour trouver donc un compromis entre le biais et la variance nous minimisions par rapport à h_nl'expression de l'erreur quadratique moyenne asymptotique$AMSE$ (asymptotic mean square error) donnée par :

$$AMSE\left[r_n(x)\right] = \frac{h_n^4}{4}\left\{r^{\backprime\backprime}(x) + 2r^{\backprime}(x)\frac{f_X^{\backprime}(x)}{f_X(x)}\right\}^2\left[u^2K(u)\right]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]$$

Comme $AMSE$ est fonction convexe.La fenêtre $h_{opt(r_n(x))}^{MSE} = \arg\min_h(AMSEr_n(x))]$

est solution de l'équation suivante :

$$\frac{\partial}{\partial h_n}\left[\frac{h_n^4}{4}\left\{r^{\backprime\backprime}(x) + 2r^{\backprime}(x)\frac{f_X^{\backprime}(x)}{f_X(x)}\right\}^2\left[u^2K(u)\right]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]\right] = 0$$

Lorsque $[r^{\backprime\backprime}(x) + 2r^{\backprime}(x)\frac{f_X^{\backprime}(x)}{f_X(x)}]^2\left[u^2K(u)\right] \neq 0$

d'où

$$h_{opt(r_n(x))}^{MSE} = n^{-1/5}\left\{\frac{\frac{\sigma^2(x)}{f_X(x)}[K^2(x)]}{\left\{\left\{r^{\backprime\backprime}(x) + 2r^{\backprime}(x)\frac{f_X^{\backprime}(x)}{f_X(x)}\right\}^2[tK]^2\right\}}\right\}^{1/5}$$

MISE(mean integrated squared error)

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

En appliquant le théoréme de Fubini, on a

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

Sous les mêmes hypothéses que les propositions (*2.2.1*) et (*2.2.3*),on a

$$AMISE\left[r_n(x)\right] = \frac{h_n^4}{4}\int\left\{r''(x) + 2r'(x)\frac{f_X'(x)}{f_X(x)}\right\}^2 dx\left[u^2K(u)\right] + \frac{1}{nh_n}\int\frac{\sigma^2(x)}{f_X(x)}dx\left[K^2(u)\right]$$

Théorème 2.3.1 ***(AMISE sous condition de continuité)***

Supposons que :

$$\exists\beta > 0 \text{ telle que } \inf_{x\in C} f(x) > \beta$$

et $h \to 0, nh \to \infty$ et K est borné ,intégrable ,positif ,symétrique et a support compact

On a :

$$AMISE\left[r_n(x)\right] \longrightarrow 0$$

la fenêtre $h^{MISE}_{opt(r_n(x))}$ minimisant l' $AMISE$ du critère global est :

$$h^{MISE}_{opt(r_n(x))} = n^{-1/5}\left\{\frac{\int\frac{\sigma^2(x)}{f_X(x)}[K^2(x)]dx}{\int\left\{r''(x) + 2r'(x)\frac{f_X'(x)}{f_X(x)}\right\}^2 dx\left[tK\right]^2}\right\}^{1/5}$$

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

$$h^{MSE}_{opt(f_n(x))} = n^{-1/5}\left\{\frac{f_X(x)[K^2]}{\left\{\left\{f''(x)\right\}^2\left[t^2K\right]^2\right\}}\right\}^{1/5} \tag{2.6}$$

$$h_{opt(f_n(x))}^{MISE} = n^{-1/5} \left\{ \frac{[K^2]}{\int_{\mathbb{R}} (f_X^{\backprime\backprime}(x))^2 dx \, [t^2 K]^2} \right\}^{1/5} \tag{2.7}$$

Nous notons que l'expression de h_n optimal,minimisant asymtotiquement les quatre critères d'erreurs la forme

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

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

Chapitre 3

Simulation

Dans ce dernier chapitre ,nous utilisons le logiciel **R,**pour calculer et représentons graphiquement la fonction de régression et son estimateur en vue de les *comarer* dans des situations simulées. Il s'agit de l'estimateur proposé par **Nadaraya-Watson** et présenté au **chapitre 2.** Nous donnons des exemples sur cet estimateur qui expriment l'importance de paramètre de lissage h, du noyau K.

Ensuit, nous présentons les résultats obtenus pour les différent jeux de donnée ainsi que pour les différentes noyaux K (noyau Gaussien :à support non compact et noyau d'Epanichnekov à support compact), différents valeurs de h strictement positive (h fixé ou h varié),régression linéaire et non linéaire .

3.1 Présentation des données

Rappelons qu'on suppose que l'on a observé un échantillon $\{(X_i, Y_i), i = 1, ..., n\}$ et on veut expliquer la variable aléatoire Y_i parX_i .De plus ,on suppose que le modèle est donné par l'expression :

$$Y_i = r(X_i) + \epsilon_i$$

où ϵ_i est l'aléatoire centré et indépendente de X_i. Aussi la fonction de régression

$$\forall x \in \mathbb{R}, r(x) = E\left[Y \diagup X = x\right] = \frac{\int y f(x, y) \partial y}{f_X(x)} \tag{3.1}$$

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

Nous avons vu que $r(x)$ est estimé par la quantité :

$$r_n(x) = \frac{\sum_{i=1}^{n} Y_i K_{h_n}(X_i - x)}{\sum_{i=1}^{n} K_{h_n}(X_i - x)} = \frac{g_{n,X}(x)}{f_{n,X}(x)} \tag{3.2}$$

Il dépend de la taille de l'échantillon n et aussi du noyau K et de la fenêtre h_n qu'il faut choisir pour calculer $r_n(x)$.avec $g_{n,X}(x)$est l'estimateur naturel de $g_X(x)$

$$g_{n,X}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K_{h_n}(X_i - x)$$

et $f_{n,X}(x)$ l'estimateur à noyau de densité

$$f_{n,X}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K_{h_n}(X_i - x)$$

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

$$y = r(x) + \epsilon \text{ où } \epsilon \to N(0, \sigma^2) \tag{3.3}$$

et on va estimer les deux fonctions de régression suivantes à l'aide d'un estimateur de **Nadaraya(1964)-Watson (1964) :**

- Régression linéaire : $r(x) = 3 + 0.5X + \epsilon$,
- Régression non linéaire :$r(x) = 3\cos(X) + \epsilon$

on supposons que X est de loi normal centré de variance $\sigma^2 = 2$ et ϵ un terme d'erreur de loi $N(0, 1)$.

Nous allons donc étudier les cas suivant dans chaque modéle :

1. Paramètre de lissage ou fenêtre h fixe, noyau normal (noyau à support non compact) et n varié.
2. Paramètre de lissage ou fenêtre h fixe,noyau d'Epanechnikov (noyau à support compact) et n varié.
3. n fixe et fenêtre h varié (noyau normal).
4. n fixe et fenêtre h varié (noyau d'Epanechnikov).

3.2 Régression linéaire

On veut estimer le modèle linéaire

$$r(x) = 3 + 0.5X + \epsilon$$

Dans les résultats graphique de cette section ,on a :

- la droite noire exprime la fonction régression $r(x)$
- la droite bleu exprime la fonction de régression emprique $r_n(x)$

3.2.1 Paramétre de lissage h fixé, n varié

En choisissant le paramètre de lissage $h_n = n^{-\frac{1}{5}}$ et n varié ($n = 60, 200, 600$)

Le noyau K à support non compact

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

code R utilisé :

```
rm(list=ls(all=TRUE))
n=60
X=rnorm(n,0,2)
E=rnorm(n)
Y=3+.5*X+E
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
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)
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
op=par(mfrow=c(1,3))
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=60",type='l',col=4, lwd= 2)
abline(3,.5,lwd= 2)

#####Pour n =200
n=200
X=rnorm(n,0,2)
E=rnorm(n)
Y=3+.5*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=200",type='l',col=4, lwd= 2)
abline(3,.5,lwd= 2)

#####Pour n =600
n=600
X=rnorm(n,0,2)
E=rnorm(n)
Y=3+.5*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=600",type='l',col=4, lwd= 2)
abline(3,.5,lwd= 2)
par(op)
```

L'axe des absices représent les valeurs des x et l'axe des coordonnées les valeurs des r_n(et r).Par la comparaison graphique ,on remarque que le graphe bleu de r_n est approche beaucoup à la droite noire de r dans le troisième graphe, donc ce graphe exprime la convergence de l'estimateur r_n vers r

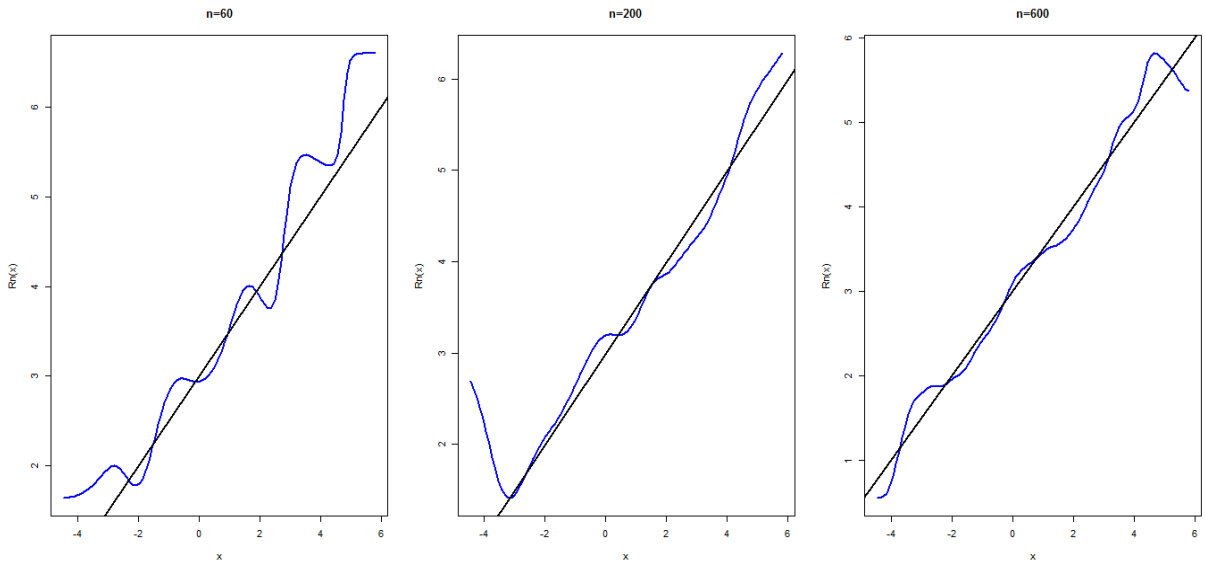

FIG. 3.1 – Régression linéaire : h fixé , n variée et K noyau normal

Le noyau K à support compact

Dans ce second cas, on choisit le noyau d'Epanechnikov :$K(t) = \frac{3}{4}(1-t^2)I_{(|t|<1)}$, $\forall t \in \mathbb{R}$, ensuit, 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 [3.2] suivante :

Même conclusion de la convergence de l'estimateur **(**voir **[3.1])**,i.e : convergence de l'estimateur pour n assez grand)

3.2.2 Choix graphique du paramètre de lissage

Dans cette section ,nous prenons le paramètre de lissage dans l'intervalle$]0;1[$ et avec des tests graphiques en va déterminer le paramètre h optimal (au sens graphique).

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[3.3]. Il est claire que la valeur du h optimale est $h = 0.8$ (ligne 3, colonne 2).

Code R utilisé :

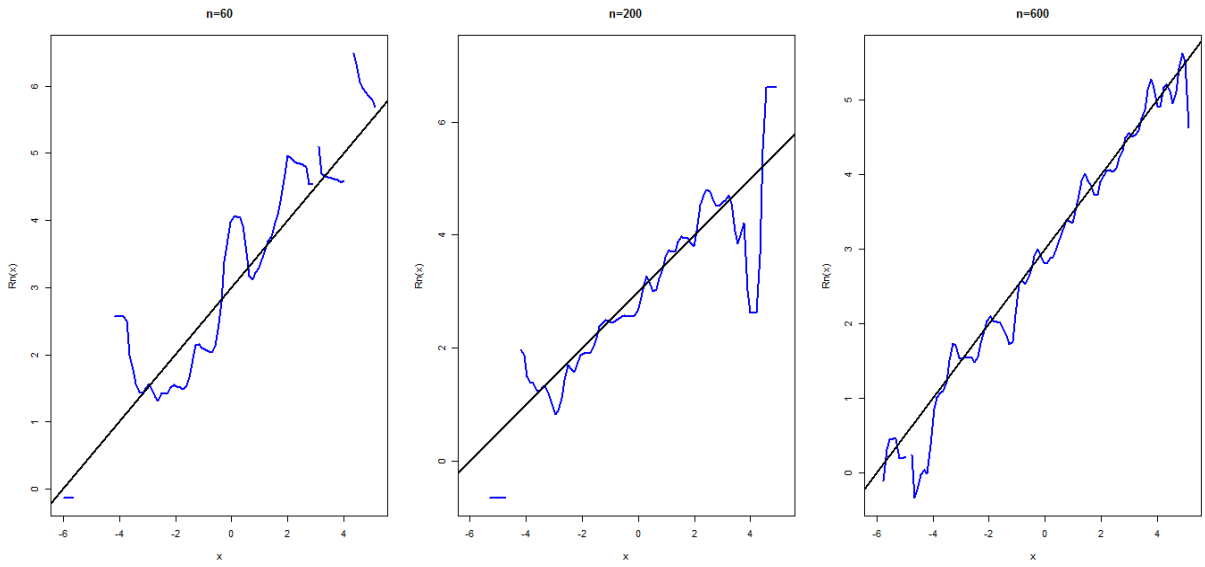

FIG. 3.2 – Régression linéaire :h fixe, n varié et K noyau d'Epanechnikov

```
#### h non fix\'{e} ####

n=300 # taille de l'\'{e}chantillon

X=rnorm(n,0,2)

E=rnorm(n)

Y=3+.5*X+E

# Noyau Normal K(t)

K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}

# param\‘{e}tre de lissage h

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(3,.5,lwd= 2)
}
par(op)
```

Identique aux choix précedents, mais on change le noyau : $K(t) = \frac{3}{4}(1-t^2)I_{(|t|\prec 1)}$ (noyau d'Epanechnikov). On obtennu figure[3.4] qui explique l'estimation obtenue avec les valeurs de h varié de $0,1$ à $0,9$

Il est claire que la valeur du h optimal est $h = 0,9$ (ligne 3 ,colonne3)

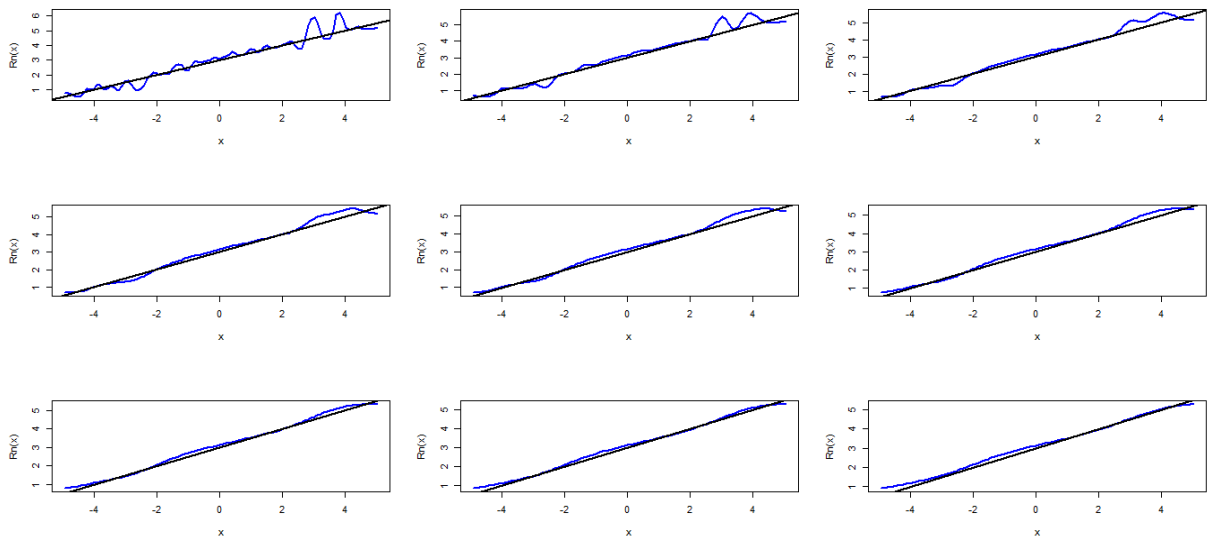

FIG. 3.3 – Régression linéaire avec h varié, n fixé et K noyau gaussien

3.3 Régression non linéaire

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

$$y = 3\cos x + \epsilon$$

où ϵ un terme d'erreur de loi $N(0;1)$.

Toujours, la ligne noire exprime la fonction de régression théorique $r(x)$ [Eq 3.1]et la ligne bleu exprime la fonction de régression empirique $r_n(x)$ donnée par l'équation [Eq 3.2]

3.3.1 Paramètre de lissage h fixé, n varié

Dans ce cas, on choisit le paramètre de lissage $h = n^{-\frac{1}{5}}$ (fixé), n varié ($n = 60, 200, 600$) et K est un noyau gaussien $K(t) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-t^2}{2})$

Code R utilisé :

```
rm(list=ls(all=TRUE)) # Nouveau programme
n=60
X=rnorm(n,0,2)
```

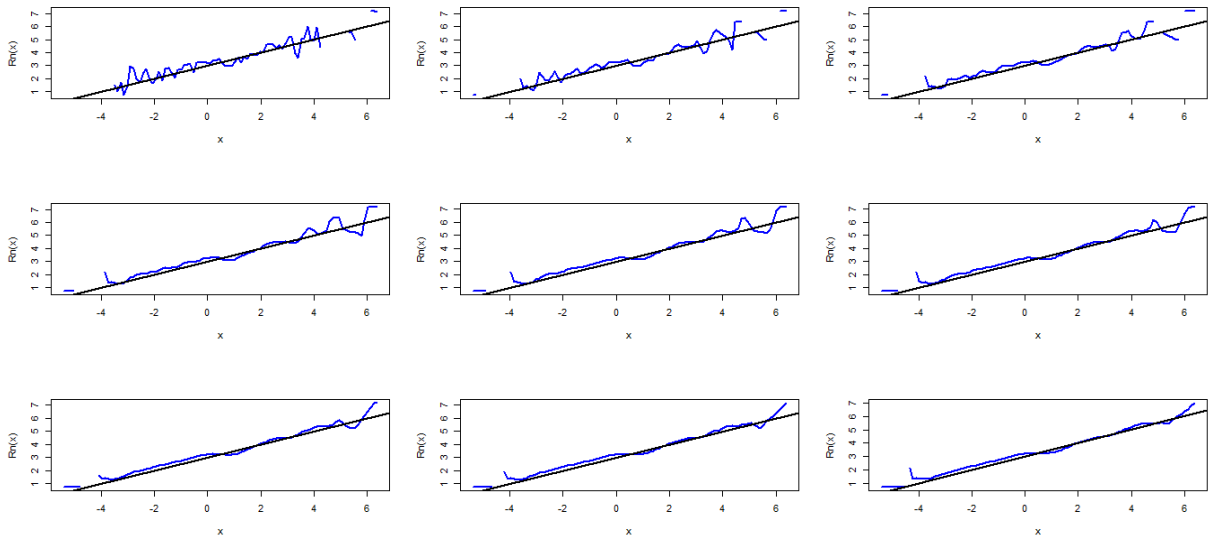

FIG. 3.4 – Régression linéaire avec h varié, n fixe et K noyau d'Epanechnikov

```
E=rnorm(n)

Y=3*cos(X)+E

# Noyau Normal K(t)

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=60",type='l',col=4, lwd= 2)
lines(x,3*cos(x),lwd= 2)
#####Pour n =200
n=200
X=rnorm(n,0,2)
E=rnorm(n)
Y=3*cos(X)+E # Mod\`{e}le Cosinus Non Lin\'{e}aire
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=200",type='l',col=4, lwd= 2)
lines(x,3*cos(x),lwd= 2)
```

```
#####Pour n =600

n=600

X=rnorm(n,0,2)

E=rnorm(n)

Y=3*cos(X)+E # Mod\‘{e}le cosinus Non Lin\’{e}aire

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=600",type='l',col=4, lwd= 2)

lines(x,3*cos(x),lwd= 2)

par(op)
```

On obtient la figure [3.5], la même conclusion pour le cas non linéaire que le cas linéaire (i.e., convergence de l'estimateur pour n assez grand).

Dans ce second cas, on choisit le noyau d'Epanechnikov : $K(t) = \frac{3}{4}(1-t^2)I_{(|t|<1)}$. En suit , on modifié seulement cette partie dans le programme **R** précédent :

#Noyau d'Epanechnikov K(t)

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

On obtient la figure[3.6] ; et on arrive au même conclusion de la convergence de l'estimateur (voir la [3.5] i.e : convergence de l'estimateur pour n assez grand)

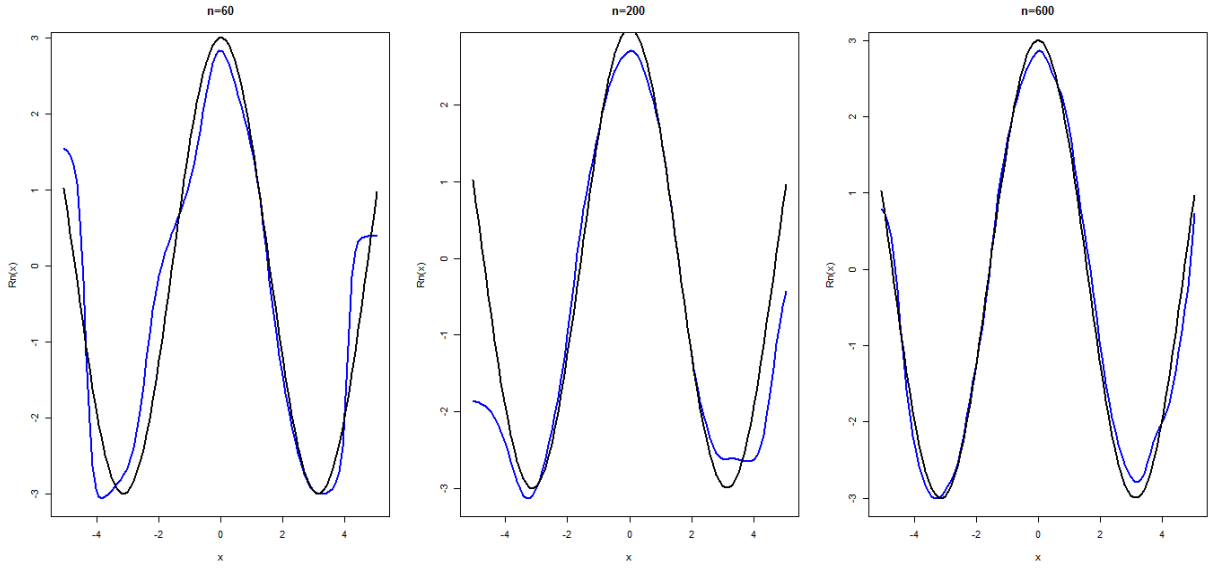

FIG. 3.5 – Régression non linéaire : h fixé, n varié et K noyau normal

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 que pour la régression linéaire, et avec des tests graphique en va diterminer le paramétre h optimal (au sens graphique)

On fixé la taille de l'échantillon $n = 300$ et le noyau K est normal , l'estimation obtennue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure [3.7].

Code R utilisé :

```
n=300 #
X=rnorm(n,0,2)
Y=3*cos(X)+E
# Noyau Normal
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
# param\‘{e}tre de lissage h
h=seq(.1,.9,length=9)
# Initiation
s=100 # taille de l’intervalle [a,b]
```

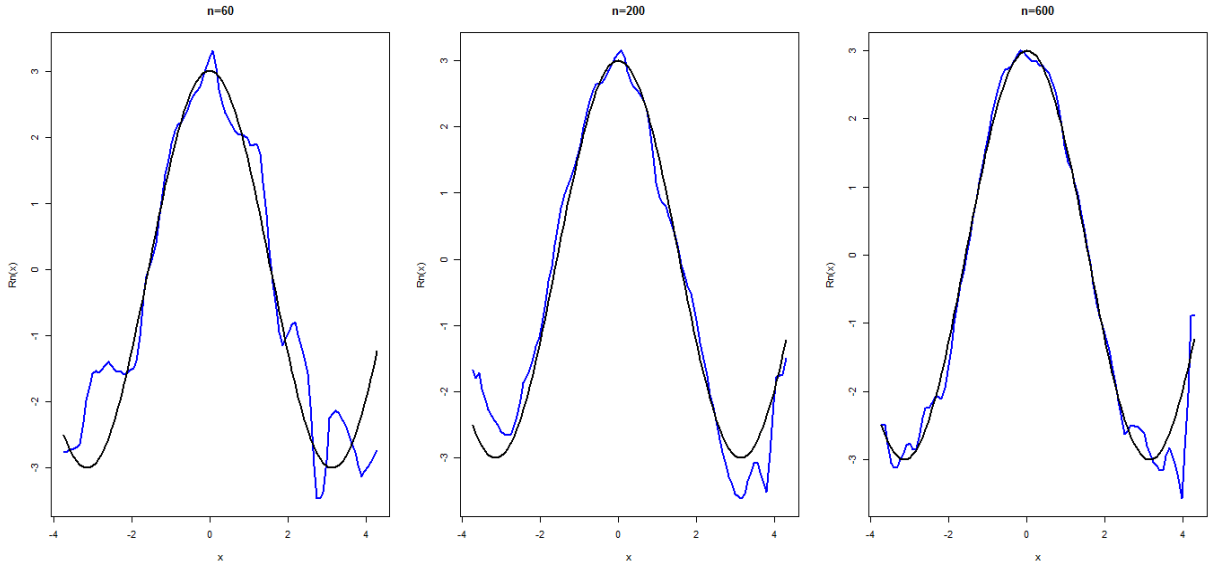

FIG. 3.6 – Régression non linéaire : h fixé, n varié et K noyau d'Epanechnikov

```
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)
lines(x,3*cos(x),lwd= 2)
}
par(op)
```

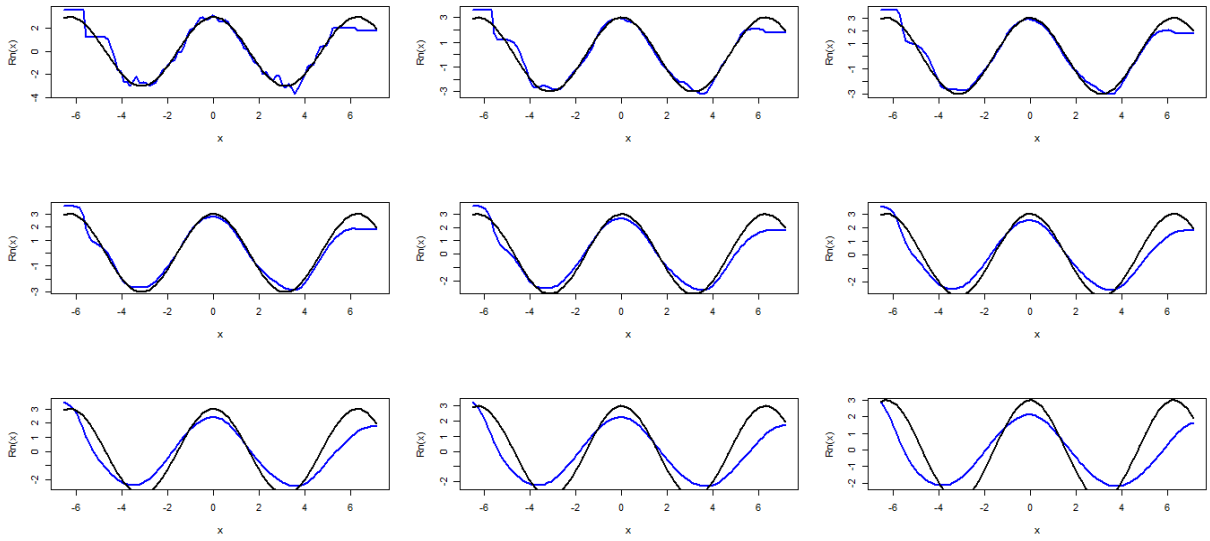

FIG. 3.7 – Régression non linéaire avec h varié, n fixé et K gaussien

A noté que la valeur du h optimale est de $h = 0.3$ (ligne 1, colonne 3).

Si nous gardons le même modèle non linéaire $y = 3 * \cos(x) + \epsilon$, mais avec le noyau d'Epanechnikov, on note que la valeur de $h = 0.5$ (ligne 2,colone 2, voir la [3.8])

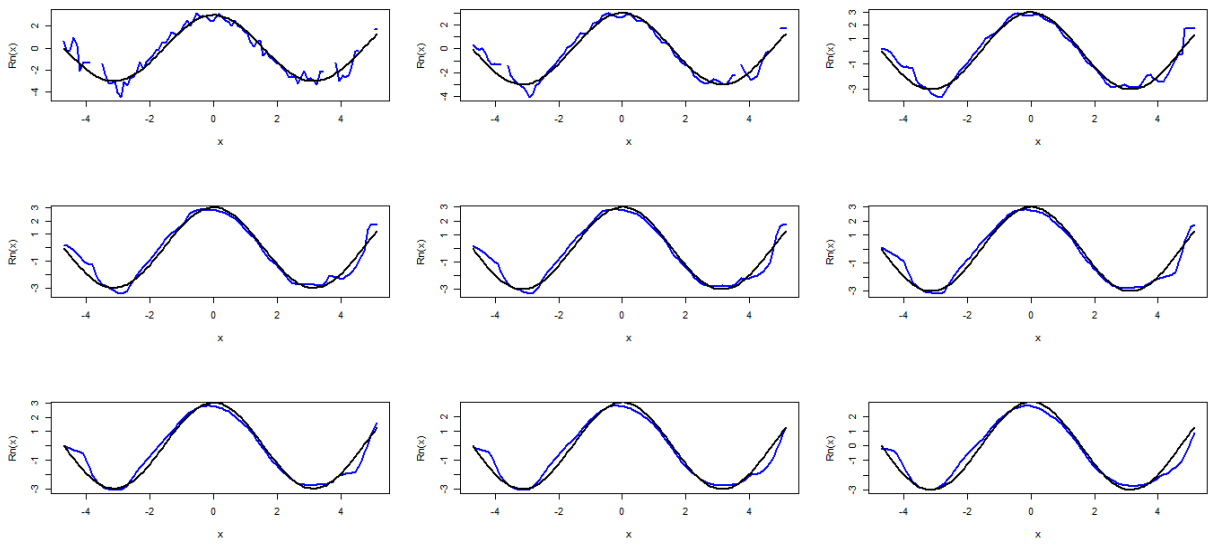

FIG. 3.8 – Régression non linéaire avec h varié, n fixé et K d'Epanechnikov

Finallement, ce chapitre montre l'importance de paramètre de lissage h et du noyau K dans l'estimation non paramètrique de la régression. Mais noté que le choix de h est plus crucial que le choix de noyau.

Conclusion

Dans ce mémoire,on a présenté la méthode d'estimation à noyau ,qui permettant d'effectuer de la régression non paramétrique.Ce travail a montré que la méthode d'estimation de régression non paramétrique est simple et peut être très utile dans plusieurs situations. Par exemple ,dans l'analyse des données, lorsque l'on désire comprendre et observer les relations qui existent entre les variables.

La méthode du noyau joue un grand rôle.en régression non paramétrique. 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 utilisateurs.

Dans la pratique, on utilisé le logiciel ***R*** *pour présenté des exemples sur cet estimateur, et à travers les résultats obtennus, nous concluons que : le noyau K est peu influence sur l'estimateur, par contre le paramètre h est un grand influence, et dont le choix est crucial. Les cas des données incomplètes : tronquées ou censurées est intéressant pour une étude future.*

Bibliographie

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

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

[3] Carbonez A , Györfi, L., van derMeulin, E.C. (1995). Partition-estimate of a regression function under random censoring. Statistics and Decisions, 13 : 21–37.

[4] 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.

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

[6] Devroye, L. (1983) The equivalence of weak, strong and complete convergence in L1 for kernel density estimates. *The Annals of Statistics*, 896-904.

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

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

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

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

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

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

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

[14] Rosenblatt, F. (1962). Principles of Neurodynamics. Spartan, New York.

[15] 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.

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

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

[18] Watson, G.S., (1964). Smooth regression analysis. Sankhya Ser. A 26 : 359–372.

Annexe A : Logiciel *R*

Les chapitres de ce mémoire comprennent des simulations effectuées en utilisant le Logiciel **R**. Les codes **R** utilisés sont donnés avec les sorties graphiques correspondantes .L'étude de simulation est basé sur l'observation des résultats d'une estimation de la régression avec la méthode du noyau .L'influence de plusieurs paramétres tels que le nombre de données générées (taille de l'échantillon noté n),la valeur choisie pour la fenêtre h et le noyau K est bient détaller

Présention du logiciel R

R est un système ,communément appelé langage et logiciel, qui permer de réaliser des analyses statistiques.Plus particuliérement ,il comporte des moyens qui rendent possible la

manipulation des données , les calculs et les représentations graphiques.**R** a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand

nombre de procédures statistiques appelées paquets. Il été créé ,en 1996,par Robert Gentleman et Ross Ihaka du département de statique de l'Université d'Auckland en Nouvelle Zélande.

le logiciel **R** est disponible sur le site

$$http://cran.r-project.org/$$

Il existe des versions

- Windows
- MacOs

Linux

Outils disponible :

– un langage de programmation orienté objet

– des fonctions de "bse"

– des librairies complémentaires (1800 sur le site CRAN)

Objets :

On site quelques objets de basse sur **R** :

1 Fonctions

2 Vecteurs,Matrices,etc

3 Listes : C'est une structure qui regroupe des objets (pas nécessirement de méme type)

4 Boucles et calculs vectoriels

5 Graphiques

Annexe B : Abréviations et Notations

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
$\widehat{\theta}$	Estimateur de θ
$i.i.d$	indépendantes et identiquement distribuées
$h := h_n$	Paramètre de lissage ou fenêtre
$K(.)$	Noyau
E	Espérance de probabilité
$biais$	Biais d'un estimateur
$\mathbb{R}$	Ensemble des nombres réels
var	Variance d'un estimateur
f_X	Densité de X
$\xrightarrow{p}$	convergence en probabilité
$\xrightarrow{p.s.}$	convergence presque sûre.
$\xrightarrow{m.p}$	convergence en moyenne d'ordre p.
F	Fonction de répartition
$f_{n,X}$	Estimateur de f

$v.a$	Variable aléatoire
L^1	espace des fonctions intégrables
r	Fonction de regression
r_n	Estimateur de r
$I_{(.)}$	Fonction indicatrice
$N(,)$	loi normal
EQM	Erreur Quadratique Moyenne
MSE	Mean Squared Error