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

MANCER Aicha

Titre :

Estimation non paramétrique de la fonction de régression

Membres du Comité d'Examen :

Mme.	TOUBA Sounia	UMKB	Président
Mlle.	KHEIREDDINE Souraya	UMKB	Encadreur
Mme.	OUANOUGHI Yasmina	UMKB	Examinateur

Juin 2018

REMERCIEMENTS

Je tiens à dire que j'ai réussie à réaliser ce modeste travail grâce à ***Dieu*** *qui m'à donner le pouvoir, la santé, la volonté et le courage d'être arrivé jusqu'au là.*

A mon très cher père et ma très chère mère.

Je remercie en particulier mon encadreur Dr. ***Kheireddine Souraya*** *qui m'a honorés par son encadrement, sa direction, son orientation, sa modestie, sa patience, ses conseils et toutes, ses remarques constructives pour le bon déroulement de mon travail.*

J'exprime aussi mes remerciements à tous le personnel de département de Mathématiques, à toute personne qui a posée son empreinte de loin ou de prés pour faire aboutir ce modeste travail.

Merci du fond du cœur

Table des matières

Table des figures

Introduction

La théorie de l'estimation est une des préoccupations majeures des statisticiens. Ainsi l'estimation non paramétrique réelle a reçu un intérêt croissant tant sur le plan théorique que pratique. Elle consiste généralement à estimer à partir des observations une fonction inconnue, élément d'une certaine classe fonctionnelle, telle que la fonction de densité ou la fonction de régression à titre d'exemples.

Les travaux de Rosenblatt (1956)[14] et Parzen (1962)[13] puis de Nadaraya- Watson (1964)[[11], [17]] portant respectivement sur les estimateurs non paramétriques 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 du noyau.

Ce travail est subdivisé en trois chapitres :

Dans le premier chapitre nous introduisons la définition de l'estimation fonctionnelle et l'estimateur à noyau de la densité (estimateur de Parzen-Rosenblatt), et ses propriétés asymptotiques.

Nous présentons, au chapitre 2, l'estimateur non paramètrique de la fonction de régression (estimateur de Nadaraya-Watson). Nous traitons ici, les propriétés asymptotiques de cet

estimateur (la convergence en moyenne quadratique et la convergence presque complète) et sa normalité asymptotique.

Finalement, Nous terminons notre mémoire par un troisième chapitre où nous utilisons le logiciel R pour donnons des exemples sur l'estimateur non paramétrique de régression (linéaire et non linéaire) par la méthode du noyau qui expriment l'importance de paramètre de lissage h, du noyau K.

Chapitre 1

Estimation non paramétrique de la densité

1.1 Définition d'estimation paramétrique et non paramétrique

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

Soit $X : \Omega \longrightarrow \mathbb{E}$ 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 paramètre à estimer est l'application

$$h : P \longrightarrow \Theta' \quad \text{ou} \quad \begin{array}{l} \Theta \longrightarrow \Theta' \\ \theta \longmapsto h(P_\theta) \end{array}$$

Un éstimateur de h est une fonction $h_n : x \longmapsto h_n(X_1, ..., X_n)$ 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 $\Theta \subset \mathbb{R}^s$ et $h(.,.)$ est une fonction connue, on parle alors d'éstimation paramétrique, car éstimer h revient à éstimer le paramétr finidimensionnel θ.*

Définition 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'éstimation non paramétrique ou de l'éstimation 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 indépendantes 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 loi λ, donc elle admet (théorème de Radon-Nykadim) une densité $f = \frac{d\mu}{d\lambda}$ c'est un paramètre dans L^1. Si $\frac{d\mu}{d\lambda}$ admet une version bornée (respectivement continue et bornée) alors on peut la considérer comme un paramétre dans L^2 (respectivement dans $C_b(\mathbb{R}^s)$).
Enfin, si f_μ est différentiable, on définit de nouveaux paramètres fonctionnels : les dérivées partielles de f_μ.

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

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

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

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

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 définie par

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

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

iv) **Le paramètre de régression :** Supposons que l'on observe un échantillon $\{(X_i, Y_i), \quad i = 1, ..., n\}$ d'un couple (X, Y) à valeurs dans $\mathbb{R}^{s_1} \times \mathbb{R}^{s_2}$ est soit μ_Y^x, $x \in \mathbb{R}^{s_1}$ une famille de versions des lois conditionnelles de Y sachant $X = x$: Toute fonction de la forme $r : x \longmapsto r(\mu_Y^x)$ est un paramètre de régression. Les plus usuels sont :

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

2) La densité conditionnelle,

3) Le mode conditionnel,

4) La fonction de répartition conditionnelle,

5) Le quantile conditinnel.

1.2 Estimation non paramétrique de la densité par la méthode du noyau

Soient $X_1, ..., X_n$ des variables aléatoires réelles (v.a) $i.i.d$ de densité de probabilité f_X par rapport à la mesure de Lebesgue sur $\mathbb{R}$ et de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)\, dt$.

Considérons la fonction de répartition empirique

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

où $1(.)$ désigne la fonction indicatrice. D'après la loi forte des grands nombres, presque sûrement,

$$F_n(x) \longrightarrow F(x), \qquad \forall x \in \mathbb{R},$$

quand $n \longrightarrow \infty$. Donc F_n est un estimateur convergent (consistant) de F pout tout $x \in \mathbb{R}$. Comment peut-on estimer f ? Une des premières solutions intuitives a été proposée par Rosenblatt (1956)[14]. Pour $h > 0$ assez petit on a

$$f_X(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

En remplaçant ici F par son estimateur F_n, on obtient.

$$f_{n,X}^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

La fonction $f_{n,X}^R$ est un estimateur de f appelé estimateur de Rosenblatt. On peut aussi l'écrire sous la forme

$$f_{n,X}^R(x) = \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{x-h<X_i\leq x+h\}} = \frac{1}{nh} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),$$

où $K_0(u) = \frac{1}{2} 1_{\{-1<u\leq 1\}}$. Parzen(1962)[13] a suggéré une généralisation de cet estimateur :

$$f_{n,X}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{1.1}$$

où $K : \mathbb{R} \to \mathbb{R}$ est une fonction intégrable, positive et telle que $\int K(u)\, du = 1$. C'est l'estimateur à noyau de la densité ou estimateur de Parzen-Rosenblatt. La fonction K est dite noyau. Et $h > 0$ un paramètre de lissage "fenêtre" (en anglais "bandwidth") de l'estimateur.

Dans le cadre asymptotique où $n \to \infty$ on supposera que la fenêtre h dépend de n et on la notera h_n, la suite $(h_n)_{n\geq 1}$ tendant vers 0 lorsque $n \to \infty$.

Voici quelques exemples de noyaux classiques :

• Noyau rectangulaire :

$$K_1(x) = \begin{cases} \frac{1}{2}, & \text{si } |x| \leq 1; \\ 0, & \text{si } |x| > 1. \end{cases}$$

• Noyau triangulaire :

$$K_2(x) = \begin{cases} 1 - |x|\,, & \text{si } |x| \leq 1; \\ 0, & \text{si } |x| > 1. \end{cases}$$

• Noyau d'Epanechnikov ou parabolique :

$$K_3(x) = \begin{cases} \frac{3}{4}(1 - x^2), & \text{si } x \in [-1, 1]; \\ 0, & \text{sinon.} \end{cases}$$

• Noyau quadratique :

$$K_4(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, & \text{si } x \in [-1, 1]; \\ 0, & \text{sinon.} \end{cases}$$

• Noyau cubique :

$$K_5(x) = \begin{cases} \frac{35}{32}(1 - x^2)^3, & \text{si } x \in [-1, 1]; \\ 0, & \text{sinon.} \end{cases}$$

• Noyau gaussien :

$$K_6(x) = \frac{1}{\sqrt{2\Pi}} \exp\left(\frac{-1}{2}x^2\right), \quad x \in \mathbb{R}.$$

On remarquera que si le noyau K est positif et si $X_1, ..., X_n$ sont fixés, la fonction $x \to f_{n,X}(x)$ est une densité de probabilité.

1.2.1 Propriétés de l'estimateur

Le pilier des premiers résultats de la convergence de cet estimateur est le théorème de Bochner (1955)[3], rappelé ci dessous :

Théorème 1.2.1 *(Bochner)*

Soit $K : (\mathbb{R}^m, \beta^m) \to (\mathbb{R}, \beta)$ une fonction mesurable, où β^p est la tribu borélienne de $\mathbb{R}^p$, vérifiant :

$$\exists M \quad \text{(constante) telleque,} \qquad \forall z \in \mathbb{R}^m, |K(z)| \leq M,$$
$$\int_{\mathbb{R}^m} |K(z)|\, dz < \infty,$$

et

$$\|z\|^m |K(z)| \to 0 \quad quand \, \|z\| \to \infty.$$

par ailleurs, soit

$g : (\mathbb{R}^m, \beta^m) \to (\mathbb{R}, \beta)$ une fonction tq

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

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

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

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

1.2.2 La consistance de l'estimateur

L'estimateur à noyau de la densité dépend de deux paramètres la fenêtre h et le noyau K. Le noyau K établit l'aspect du voisinage de x et h contrôle la taille de ce voisinage, donc h est le paramètre prédominant pour avoir de bonnes propriétés asymptotiques, néanmoins le noyau K ne doit pas être négligé, comme le montre le travail de Parzen (1962)[13] cité ci dessous sur la consistance de cet estimateur. Cette dernière est obtenue, en se basant

sur l'étude asymptotique du biais, de la variance et de la décomposition suivante :

$$E\left[f_{n,X}(x)-f_X(x)\right]^2 = Var\left[f_{n,X}(x)\right] + \left[Biais\left\{f_{n,X}(x)\right\}\right]^2.$$

Dans la suite, nous supposons que K est un noyau vérifiant les conditions suivantes.

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

(H.2) $\lim |x|\, K(x) = 0$, quand $|x| \to \infty$,

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

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

1) Etude du biais :

Proposition 1.2.1 *Sous les hypothèses [**(H.1)**, **(H.2)**, **(H.3)** et **(H.4)**] et si f_X est continue alors*

$$\forall x \in \mathbb{R} \lim_{n\to\infty} E\left[f_{n,X}(x)\right] = f_X(x). \tag{1.2}$$

Preuve. En effet :

$$\begin{aligned} E\left[f_{n,X}(x)\right] &= E\left[\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\right] \\ &= \frac{1}{h_n} E\left[K\left(\frac{x-X_i}{h_n}\right)\right] \\ &= \frac{1}{h_n}\int_{\mathbb{R}} K\left(\frac{x-t}{h_n}\right) f_X(t)dt. \end{aligned}$$

En posanant $x - t = z$, on arrive à :

$$E\left[f_{n,X}(x)\right] = \frac{1}{h_n}\int_{\mathbb{R}} K\left(\frac{z}{h_n}\right) f_X(x-z)dz.$$

Comme K et f vérifient les conditions du théorème de Bochner, et $\lim_{n\to\infty} h_n = 0$, $n \to \infty$, on a alors

$$\lim_{n\to\infty} \frac{1}{h_n}\int_{\mathbb{R}} K\left(\frac{x-t}{h_n}\right) f_X(t)dt = f_X(x)\int_{\mathbb{R}} K(z)dz.$$

d'où

$$\lim_{n\to\infty} E\left[f_{n,X}(x)\right] = f_X(x).$$

■

Nous constatons que le biais de l'estimateur converge vers zéro quand la fenêtre tend vers zéro, de plus vu son expression, on constate qu'il ne dépend pas du nombre des variables, il dépend surtout du noyau K.

2) Etude de la variance de $f_{n,X}(x)$

Proposition 1.2.2 *Sous les conditions [**(H.1)**, **(H.2)**, **(H.3)** et **(H.4)**] et si* f_X *est continue en tout point* x *de* $\mathbb{R}$, *alors*

$$\lim_{n\to\infty} Var\left[f_{n,X}(x)\right] = 0$$

Preuve. En effet :

$$\begin{aligned} Var\left[f_{n,X}(x)\right] &= E\left[f_{n,X}(x)\right]^2 - \left[Ef_{n,X}(x)\right]^2 \\ &\leq E\left[f_{n,X}(x)\right]^2 \\ &\leq \frac{1}{n} E\left[\frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)\right]^2 \\ &\leq \frac{1}{nh_n^2}\int K^2\left(\frac{x-t}{h_n}\right) f_X(t)dt \\ &\leq \frac{1}{nh_n^2}\int K^2\left(\frac{z}{h_n}\right) f_X(x-z)dz. \end{aligned}$$

Remarquons que **(H.1)** et **(H.3)** impliquent que le noyau est de carré intégrable et les hypothèses sur h_n, K et f_X assurent que :

$$\frac{1}{nh_n^2}\int K^2\left(\frac{z}{h_n}\right) f_X(x-z)dz \sim \frac{1}{nh_n} f_X(x)\int K^2(z)\,dz,$$

d'où

$$\lim_{n\to\infty} Var\left[f_{n,X}(x)\right] = 0, \qquad \text{quand } nh_n \to \infty.$$

■

Ces deux propositions impliquent la convergence en moyenne quadratique et donc, à foriori, la consistance de l'estimateur.

Chapitre 2

Estimation non paramétrique de la fonction de régression

2.1 L'estimateur non paramétrique de régression

Nous disposons d'un échantillon composé de n couples indépendants de variables aléatoires $(X_1, Y_1), ..., (X_n, Y_n)$ et nous considérons le modèle de régression non paramétrique donné, pour $i = 1, ..., n$, par

$$Y_i = r(X_i) + \varepsilon_i. \tag{2.1}$$

où ε_i est l'aléatoire centré et indépendante de X_i et r est une application mesurable réelle. La fonction de régression $r(.) = E[Y \backslash X = .]$, apportant de l'information sur la relation de dépendance inconnue de Y et X, un problème 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, Y).

Supposons 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). Alors,

$$\forall x \in \mathbb{R},\ r(x) = E[Y \backslash X = x] = \frac{\int y f(x, y)\, dy}{f_X(x)}.$$

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

$$f_n(x,y)=\frac{1}{nh_n^2}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h_n}\right)K\left(\frac{Y_i-y}{h_n}\right),$$
$$f_{n,X}(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h_n}\right),$$

puis on considère l'estimateur de la régression

$$\forall x\in\mathbb{R},\ r_n(x)=\frac{\int y f_n(x,y)\,dy}{f_{n,X}(x)}1_{f_{n,X}(x)\neq 0}. \tag{2.2}$$

Proposition 2.1.1

Si K est un noyau d'ordre 1, l'estimateur défini par [2.2] vérifie

$$\begin{aligned}\forall x\in\mathbb{R},\ r_n(x) &= \frac{\sum_{i=1}^{n}Y_iK\left(\frac{X_i-x}{h_n}\right)}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h_n}\right)}1_{\left\{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h_n}\right)\neq 0\right\}}\\ &= \frac{\sum_{i=1}^{n}Y_iK_{h_n}(X_i-x)}{\sum_{i=1}^{n}K_{h_n}(X_i-x)}1_{\left\{\sum_{i=1}^{n}K_{h_n}(X_i-x)\neq 0\right\}},\end{aligned}$$

où $K_{h_n}(.)=K_{h_n}(./h)$.

donc l'estimareur à noyau de la régression est donné par :

$$r_n(x)=\begin{cases}\dfrac{\sum_{i=1}^{n}Y_iK_{h_n}(X_i-x)}{\sum_{i=1}^{n}K_{h_n}(X_i-x)}=\dfrac{\Phi_{n,X}(x)}{f_{n,X}(x)}, & \text{si } \sum_{i=1}^{n}K_{h_n}(X_i-x)\neq 0;\\ \dfrac{1}{n}\sum_{i=1}^{n}Y_i, & \text{si } \sum_{i=1}^{n}K_{h_n}(X_i-x)=0.\end{cases}$$

C'est l'estimateur à noyau introduit par Nadaraya-Watson (Nadaraya, 1964[11] et Watson, 1964[17]).

La construction de cet estimateur dépend de deux paramètres, le paramètre de lissage h dont le choix est crucial pour obtenir de bonnes propriétés asymptotiques et la noyau K dont on ne peut pas négliger le rôle pour la réduction du biais.

2.2 Les propriétés de l'estimateur

D'une manière analogue aux propriétés asymptotiques de l'estimateur de Parzen Rosenblatt, nous étudions dans cette partie deux modes de convergence, la convergence en moyenne quadratique et la convergence presque complète.

En plus des condition **(H.1-H.4)** sur le noyau K, nous avons besoin des hypothèses suivantes.

(H.5) $\int_{\mathbb{R}} uK(u)\,du = 0$,

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

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

2.2.1 La consistance

En vu de la décomposition suivante :

$$E\left[r_n(x) - r(x)\right]^2 = Var\left[r_n(x)\right] + \left[Er_n(x) - r(x)\right]^2.$$

L'étude asymptotique du biais et de la variance de l'estimateur de Nadaraya-Watson détermine les conditions suffisantes à la consistance de cet estimateur.

1) Etude asymptotique de la variance

Proposition 2.2.1 *Sous les hypothèse de la proposition* (1.3.2) *et si* $EY^2 < \infty$, *alors en chaque point de continuité des fonctions* $r(x)$, $f_X(x)$ *et* $\sigma^2(x) = Var(Y \backslash X = x)$

on a

$$Var\left[r_n(x)\right] = \frac{1}{nh_n}\left\{\frac{\sigma^2(x)}{f_X(x)}\int_{\mathbb{R}} K^2(u)\,du\right\}(o(1)+1).$$

où $f_X(x) > 0$.

Preuve.

Soit la fonction $\psi(x) = \int_{\mathbb{R}} y^2 f(x,y)dy$, en se basant sur le lemme de Bochner on a

$$\begin{aligned} Var\left[\Phi_{n,X}(x)\right] &= \frac{1}{nh_n^2}\left\{E\left[Y^2K^2\left(\frac{x-X}{h_n}\right)\right]-\left[EYK\left(\frac{x-X}{h_n}\right)\right]^2\right\} \\ &= \frac{1}{nh_n}\left\{\int_{\mathbb{R}}K^2(u)\,\psi(x-h_nu)du-h_n\left(\int_{\mathbb{R}}K(u)\,f(x-uh_n)\,r(x-h_nu)\right)^2\right\} \\ &= \frac{1}{nh_n}\psi(x)\int_{\mathbb{R}}K^2(u)\,du(1+o(1)), \end{aligned}$$

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

et

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

posons

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

et

$$A(x)=\left(\frac{-r(x)}{[f_X(x)]^2},\ \frac{1}{f_X(x)}\right).$$

La matrice de variance covariance de $B_n(x)$ est alors donnée par l'expression suivante

$$\sum:=\frac{1}{nh_n}\begin{pmatrix} f_X(x) & \Phi(x) \\ \Phi(x) & \psi(x)\end{pmatrix}\int_{\mathbb{R}}K^2(u)\,du(1+o(1)).$$

En remarquant, que

$$\begin{aligned} Var\left[r_n(x)\right] &= A\sum A^t \\ &= \frac{1}{nh_n}\left(\frac{\psi(x)}{[f_X(x)]^2}-\frac{(\Phi(x))^2}{[f_X(x)]^3}\right)\int_{\mathbb{R}}K^2(u)\,du(1+o(1)), \end{aligned}$$

où A^t désigne la transposée de A, on obtient alors

$$Var\left[r_n(x)\right]=\frac{1}{nh_n}\left\{\frac{\sigma^2(x)}{f_X(x)}\int_{\mathbb{R}}K^2(u)\,du\right\}(o(1)+1).$$

■

2) Etude asymptotique du biais

L'étude asymptotique du biais repose sur la proposition suivante.

Proposition 2.2.2 *Sous les hypothèse de la proposition* (2.2.1) *et*

a) Si $|Y| \leq C_1 < \infty$ $P.S$ et si $nh_n \longrightarrow \infty$, quand $n \longrightarrow \infty$, alors :

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

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

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

Maintenant nous sommes en mesure d'énoncer le resultat suivant.

Proposition 2.2.3 *Si les condition* ***(H.4), (H.5)*** *et* ***(H.6)*** *sont vérifiées et si* $f_X(.)$ *et* $r(.)$ *sont le classe* $C^2(\mathbb{R})$ *et si* $|Y|$ *est borné.*

Alors :

$$E[r_n(x)] - r(x) = \frac{h_n^2}{2}\left\{\left\{r''(x) + 2r'(x)\frac{f_X'(x)}{f_X(x)}\right\}\int_{\mathbb{R}} u^2 K(u)\,du\right\}(1 + o(1)). \qquad (2.3)$$

Remarque 2.2.1

1)- Les conditions **(H.4), (H.5)** et **(H.6)** peuvent être remplacées par le noyau K est d'ordre 2 au sens de Gasser et Müller.

2)- $o(1)$ dans la relation [2.3] est égale à $O(h) + O\left((nh)^{-1}\right)$.

Preuve.

$$\begin{aligned} E[r_n(x)] - r(x) &= \left[EK\left(\tfrac{x-X}{h_n}\right)\right]^{-1}\left\{\int_{\mathbb{R}} \tfrac{1}{h_n}K\left(\tfrac{x-t}{h_n}\right)\Phi(t)\,dt - r(x)\int_{\mathbb{R}} \tfrac{1}{h_n}K\left(\tfrac{x-t}{h_n}\right)f(t)\,dt\right\} \\ &= \left\{(f(x))^{-1}\left\{\tfrac{h_n^2}{2}\Phi''(x) - \tfrac{h_n^2}{2}r(x)f''(x)\right\}\int_{\mathbb{R}} u^2K(u)\,du + \Phi(x) - r(x)f(x)\right\}(1 + o(1)) \end{aligned}$$

comme $\Phi(x) = r(x) f(x)$. L'équation précédente peut s'écrire :

$$E[r_n(x)] - r(x) = \left\{ \frac{h_n^2}{2} \left\{ r''(x) + 2r'(x) \frac{f_X'(x)}{f_X(x)} \right\} \int_{\mathbb{R}} u^2 K(u)\, du \right\} (1 + o(1)). \qquad (2.4)$$

D'où

$$\lim_{n\to\infty} E[r_n(x)] = r(x).$$

■

2.2.2 Convergence presque complète

En se basant sur la preuve données dans Ferraty et Vieu (2003)[7], nous traitons dans ce paragraphe la convergence presque complète de l'estimateur à noyau de la fonction de régression. Nous gardons quelques conditions précédentes, aux quelles nous rajoutons les hypothèses suivantes.

- $f_X,\ r$ sont des fonctions continues au voisinage de x, un point fixé de $\mathbb{R}$. (2.5)

La densité f_X et la variable Y sont telles que

$$f_X > 0 \qquad (2.6)$$

• Le paramètre de lissage h_n est tel que

$$\lim_{n\to\infty} h_n = 0 \text{ et } \lim_{n\to\infty} \frac{\log n}{n h_n} = 0, \qquad (2.7)$$

et

$$|Y| < M < \infty, \qquad (2.8)$$

où M est une constante réelle positive.

Théorème 2.2.1 *Sous les hypothèses* (2.5) , (2.6) , (2.7) , (2.8) , ***(H.4) et (H.7),*** *on a :*

$$\lim_{n\to\infty} r_n(x) = r(x). \quad p.co$$

2.3 La normalité asymptotique de l'estimateur

La première démonstration de la normalité asymptotique de l'estimateur est due à Schuster (1972) [15]. On se réfère également aux théorème 1.3 et 1.4 p. 117-120 de Nadaraya (1989) [12] et au théorème 4.2.1 p. 99 de Härdle (1990) [8], qui proposent d'autres méthodes de démonstration. Le noyau K est supposé borné, à support compact et d'ordre 2. La fenêtre h_n est choisie égale à $cn^{-1/5}$.

Théorème 2.3.1 ***Härdle (1990)***

Supposons Y bornée ou admettant un moment d'ordre $l > 2$. Les fonctions $f_X(.)$ et $r(x)$ sont supposées deux fois continûment dérivables sur $\mathbb{R}$. A chaque point de continuité de $\sigma^2(x)$, tel que $f_X(x) > 0$,

$$(nh)^{1/2}\{r_n(x) - r(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}\left(B(x), v^2(x)\right), \tag{2.9}$$

avec

$$v^2(x) := \frac{\sigma^2(x)}{f_X(x)} \int_{\mathbb{R}} K^2(u)\, du, \quad \textit{(la variance asymptotique),}$$

et

$$B(x) := \left\{ r''(x) + 2r'(x)\frac{f'_X(x)}{f_X(x)} \right\} \times \int_{\mathbb{R}} u^2 K(u)\, du, \quad \textit{(le biais asymptotique).}$$

2.4 Choix du paramètre de lissage

Les vitesses de convergence dépendent de deux paramètres : la fonction de noyau K dont l'efficacité est peu influente et le paramètre de lissage h_n, dont le choix est crucial

aussi bien pour l'approche ponctuelle que pour la globale que nous exposons ci après.

2.4.1 Critère d'erreur quadratique moyenne de $r_n(x)$

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

Notre but est de minimiser

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

Le développement de cette expression faite précédemment, nous donne

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

Nous constatons d'une part que les expressions du $biais$ de $r_n(x)$ et de la variance de $r_n(x)$ (voir les propositions (2.2.3), (2.2.1)) permettent de conclure qu'une grande valeur de h_n donne une augmentation du $biais$ et une diminution de la variance (estimation fortement biaisée) et qu'un faible paramètre h_n, donne une diminution du $biais$ et une augmentation de la variance (phénomène de sous lissage).

D'autre part, sous les hypothèses de ces mêmes propositions, nous obtenons

$$\begin{aligned} MSE\left(r_n(x)\right) = \frac{h_n^4}{4}\left[\left(r''(x) + 2r'(x)\frac{f_X'(x)}{f_X(x)}\right)\left(u^2K(u)\right) + o(1)\right]^2 \\ + \frac{1}{nh_n}\left(\frac{\sigma^2(x)}{f_X(x)}\left[K^2(u)\right]\right)(1 + o(1)), \end{aligned} \tag{2.10}$$

où $u^pK^q(u) = \int t^pK^q(t)\,dt$. Pour trouver donc un compromis entre le $biais$ et la variance nous minimisons par rapport à h_n l'expression de l'erreur quadratique moyenne asympto-

tique $AMSE$ (asymptotique mean square error) donnée par

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

Comme $AMSE$ est une fonction convexe. La fenêtre $h^{MSE}_{opt(r_n(x))}=\arg\min_h\left[AMSE\left(r_n\left(x\right)\right)\right]$ est solution de l'équation suivante

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

lorsque $\left[r''\left(x\right)+2r'\left(x\right)\frac{f'_X(x)}{f_X(x)}\right]^2\left[u^2K\left(u\right)\right]\neq 0,$

d'où

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

On s'intéresse maintenant à l'approche globale pour la sélection du paramètre h_n, pour cela on introduit le critère d'erreur quadratique intégrée moyenne ou $MISE$ (mean integrated squared error) de $r_n\left(x\right)$.

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

En appliquant le théorème de Fubini, on a

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

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

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

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

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

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

$$h^{MISE}_{opt\left(f_{n,X}(x)\right)} = n^{-1/5}\left\{\frac{f_X(x)\left[K^2\right]}{\left(f''_X(x)\right)^2\left[t^2K\right]^2}\right\}^{1/5}, \tag{2.11}$$

$$h^{MISE}_{opt\left(f_{n,X}(x)\right)} = n^{-1/5}\left\{\frac{\left[K^2\right]}{\left[t^2K\right]^2\int_{\mathbb{R}}\left(f''_X(x)\right)^2 dx}\right\}^{1/5}. \tag{2.12}$$

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

En insérant $h^{MSE}_{opt\left(f_{n,X}(x)\right)}$ dans $MSE\left[f_{n,X}(x)\right]$ on peut montrer que le taux de convergence est d'ordre $n^{-4/5}$, il est plus faible que celui de l'histogramme dont l'ordre est égale à $n^{-2/3}$.

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

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

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 représenter graphiquement la fonction de regression 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 Epanichnekov : à support compact), différents valeurs de h strictement positif (h fixé ou h varié), régression linéaire et non linéaire.

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

$$Y_i = r(X_i) + \varepsilon_i.$$

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

$$r(x) = E[Y \backslash X = x] = \frac{\int y f(x, y)\, dy}{f_X(x)} \tag{3.1}$$

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

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

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

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

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

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

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

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

$$y = r(x) + \varepsilon, \quad \text{où } \varepsilon \longrightarrow \mathcal{N}(0, \sigma^2), \tag{3.3}$$

et on va estimer les 2 fonctions de régression suivantes à l'aide d'un estimateur de Nadaray-Watson :

- Régression linéaire : $r(x) = 3 + 0.9x + \varepsilon$.
- Régression non linéaire : $r(x) = \cos(x) + \varepsilon$.

on supposons que : X suit la loi exponentielle de paramètre 0.5 : $X \sim \mathcal{E}(0.5)$

Nous allons donc étudier les cas suivants dans chaque modèle :

◦ Paramètre de lissage ou fenêtre h fixe, noyau normale (noyau à support non compact) et n varié.

◦ Paramètre de lissage ou fenêtre h fixe, noyau d'Epanechnikov (noyau à support non compact) et n varié.

∘ n fixe et fenêtre, h varié (noyau normale).

∘ n fixe et fenêtre, h varié (noyau d'Epanechnikov).

3.1 Régression linéaire

On veut estimer le modèle linéaire

$$y = 3 + 0.9x + \varepsilon.$$

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

Dans les résultats graphique de cette section, on a :

∗ La droite noire exprime la fonction de régression $r(x)$ [Eq.3.1] .

∗ La droite en rouge exprime la fonction de régression empirique $r_n(x)$ [Eq.3.2] .

3.1.1 Paramètre de lissage h fixé, n varié

En choisissant le paramètre de lissage $h_n = n^{-1/5}$ (fixé) et n varié ($n = 40,\ 200,\ 400$)

3.1.1.1 K à support non compact

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

Code R utilisé :

```
rm(list=ls(all=TRUE))
n=40
X=rexp(n,0.5)
E=rnorm(n)
Y=3+.9*X+E
```

```
# Noyau Normale 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
op=par(mfrow=c(1,3))
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=40",type='l',col=2, lwd= 2)
abline(3,.9,lwd= 2)
#####Pour n =200
n=200
X=rexp(n,0.5)
```

```
E=rnorm(n)
Y=3+.9*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=2, lwd= 2)
abline(3,.9,lwd= 2)
#####Pour n =400
n=400
X=rexp(n,0.5)
E=rnorm(n)
Y=3+.9*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=2, lwd= 2)
abline(3,.9,lwd= 2)
par(op)
```

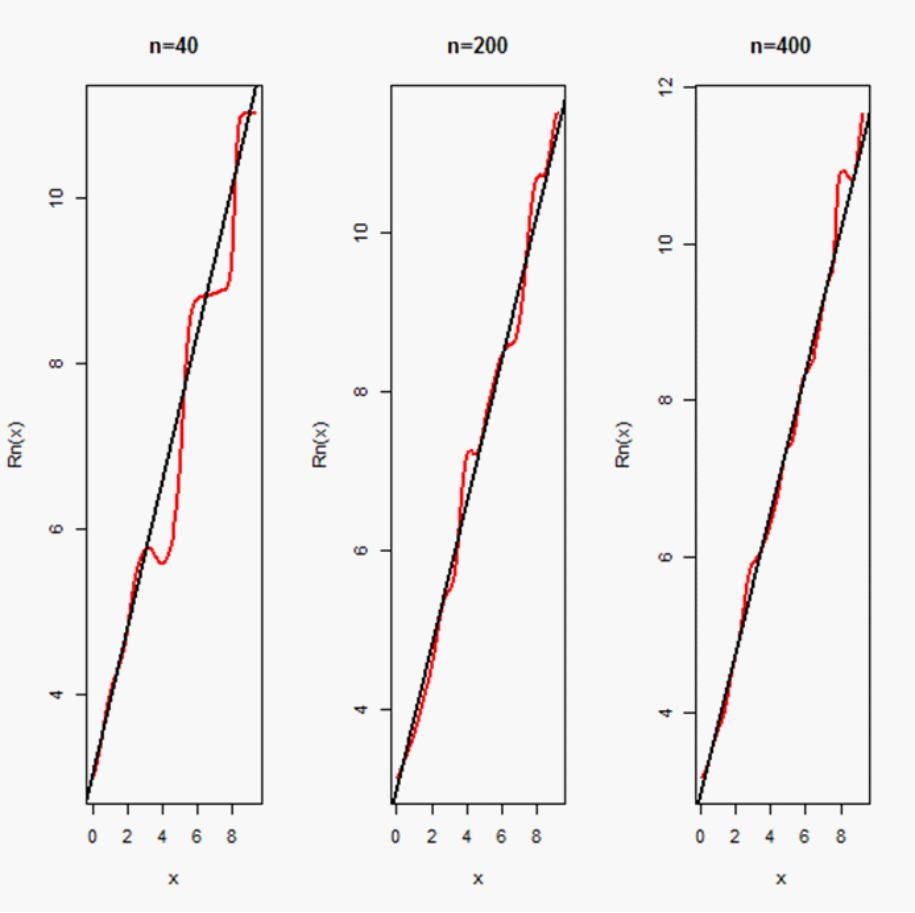

FIG. 3.1 – Régression linéaire : h fixé, n varié et K noyau normale.

L'axe des absices représente 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 rouge 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.

3.1.1.2 K à support compact

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

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

On obtient la figure [FIG-3.2] , et on arrive au même conclusion de la convergence de l'estimateur (voire la [FIG-3.1], *i.e*, convergence de l'estimateur pour n assez grand).

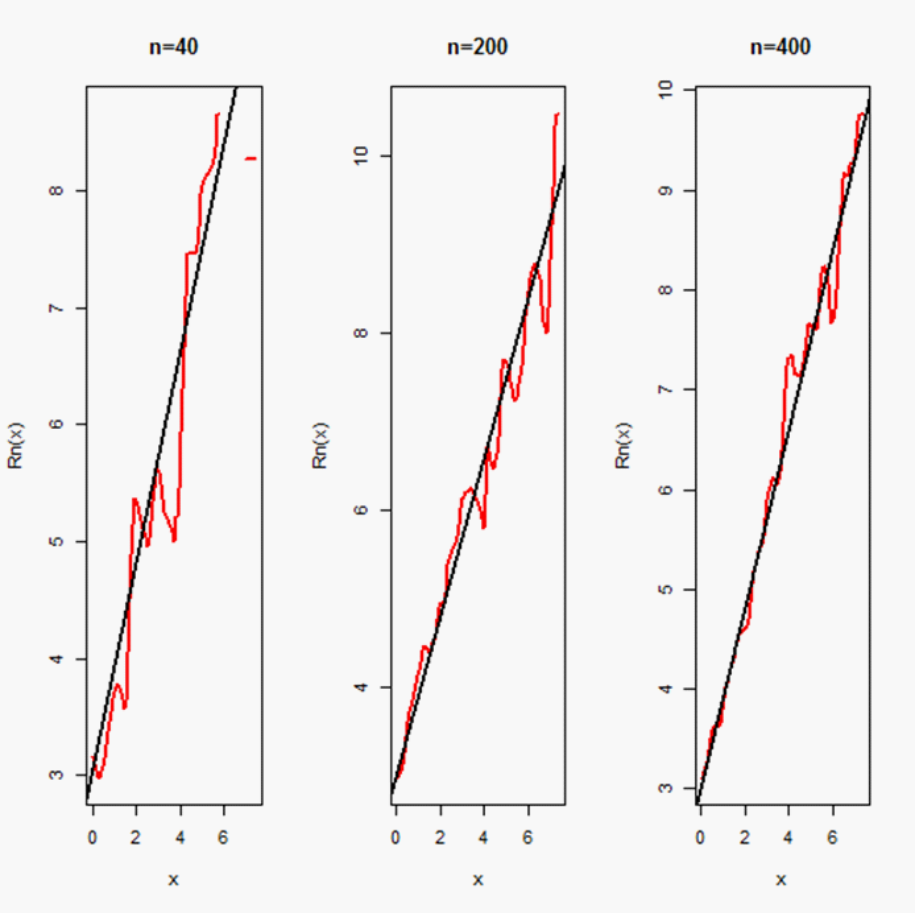

FIG. 3.2 – Régression linéaire : h fixé, n varié et K noyau d'Epanechnikov

3.1.2 Choix graphique du paramètre de lissage

Dans cette section, nous prenons le paramètre de lissage dans l'intervalle $(0; 1)$ et avec des tests graphique en va diterminer le paramètre h optimal (au sens graphique).
On fixe la taille de l'échantillon $n = 350$ et le noyau K est normale, l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure [FIG-3.3]. Il est clair que la valeur du h optimale est de $h = 0.7$ (ligne 3, colonne 1).

Code R

```
n=350
X=rexp(n,0.5)
E=rnorm(n)
Y=3+.9*X+E
# Noyau Normale 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))
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=2, lwd= 2)
abline(3,.9,lwd= 2)
}
par(op)
```

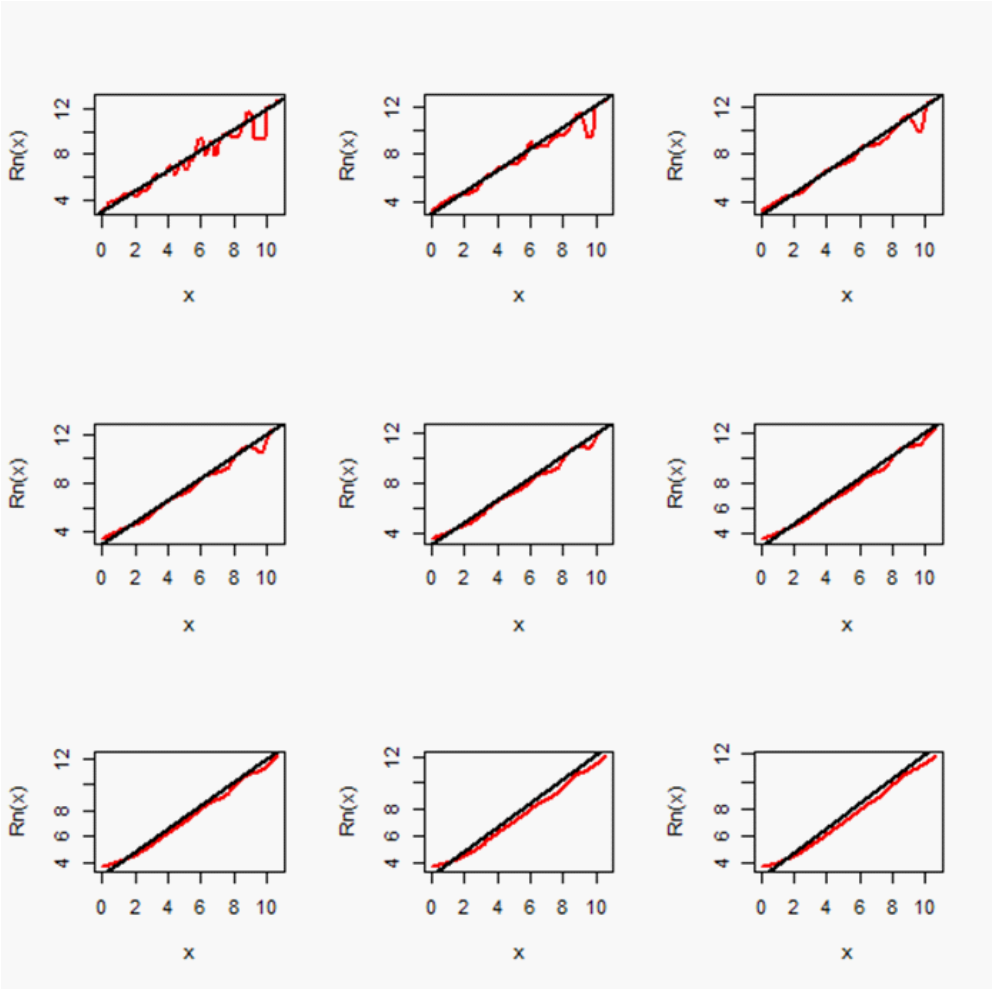

FIG. 3.3 – Régression linéaire avec h varié, n fixé et K noyau gaussion.

Identique aux choix précédents, mais on change le noyau : $K(t) = \frac{3}{4}(1-t^2)$ (noyau d'Epanechnikov). On obtenu la figure [FIG-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 optimale est de $h = 0.9$ (ligne 3, colonne 3).

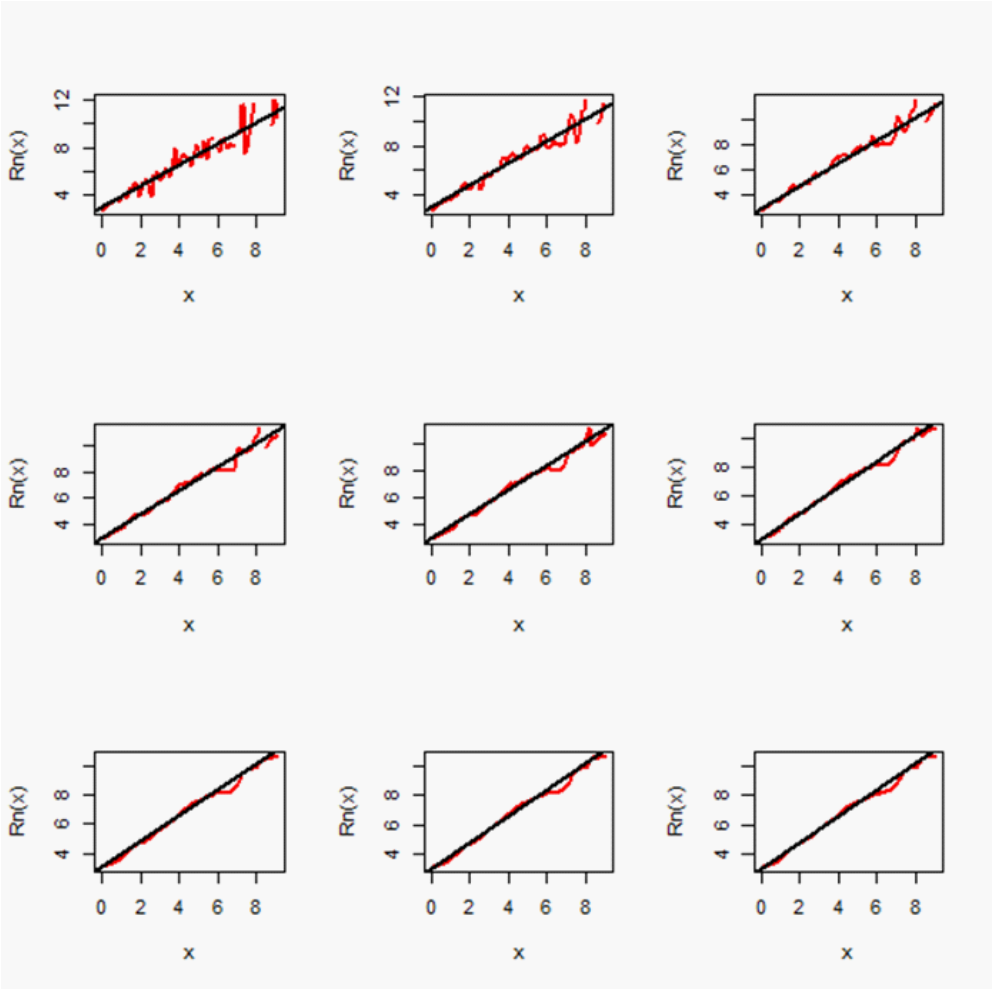

FIG. 3.4 – Régression linéaire avec h varié, n fixé et K d'Epanechnikov

3.2 Régression non linéaire

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

$$y = \cos x + \varepsilon. \tag{3.4}$$

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

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

3.2.1 Paramètre de lissage h fixé, n varié

Dans ce cas, on choisit le paramètre de lissage $h_n = n^{-1/5}$ (fixé), n varié ($n = 40,\ 200,\ 400$) et K est un noyau gaussien $K(t) = \frac{1}{\sqrt{2\Pi}} \exp\left(\frac{-1}{2}t^2\right)$.

Code R :

```
rm(list=ls(all=TRUE))
n=40
X=rexp(n,0.5)
E=rnorm(n)
Y=cos(X)+E
# Noyau Normale 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=40",type='l',col=2, lwd= 2)
lines(x,cos(x),lwd= 2)
#####Pour n =200
n=200
X=rexp(n,0.5)
E=rnorm(n)
Y=cos(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=2, lwd= 2)
lines(x,cos(x),lwd= 2)
#####Pour n =400
n=400
X=rexp(n,0.5)
```

```
E=rnorm(n)
Y=cos(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=2, lwd= 2)
lines(x,cos(x),lwd= 2)
par(op)
```

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

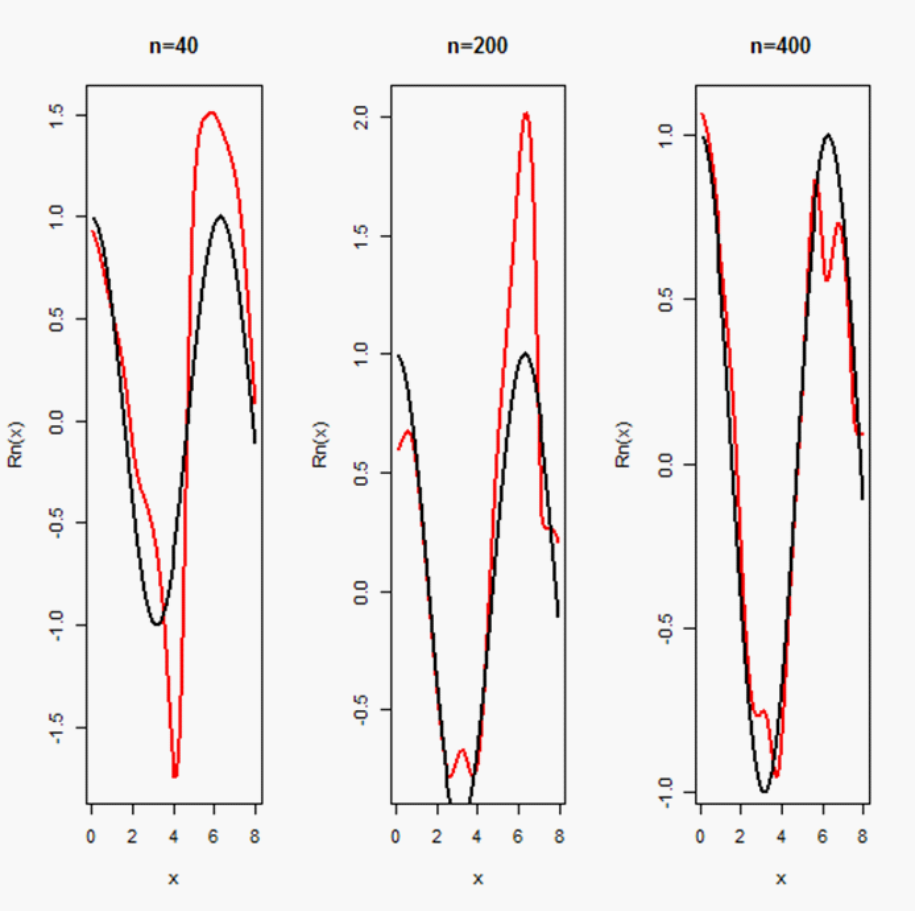

FIG. 3.5 – Régression non linéaire : h fixé, n varié et K noyau normale

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

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

On obtient la figure [FIG-3.6] , et on arive au même conclusion de la convergence de l'estimateur (voir la [FIG-3.5], *i.e*, convergence de l'estimateur pour n assez grand).

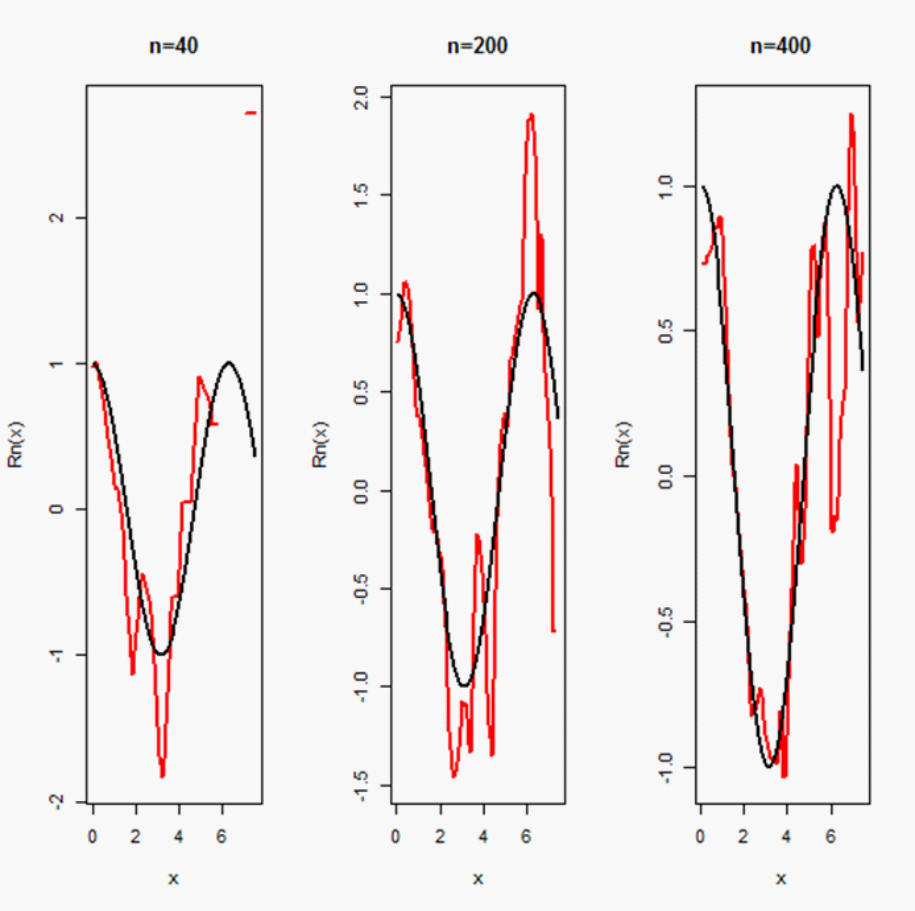

FIG. 3.6 – Régression non linéaire : h fixé, n varié et K noyau d'Epanechnikov

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

On fixe la taille de l'échantillon $n = 350$ et le noyau K est normale, l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure [FIG-3.7]. Il est clair que la valeur du h optimale est de $h = 0.5$ (ligne 2, colonne 2).

Code R :

```
n=350
X=rexp(n,0.5)
```

```
Y=cos(X)+E
# Noyau Normale 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))
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=2, lwd= 2)
lines(x,cos(x),lwd= 2)
}
par(op)
```

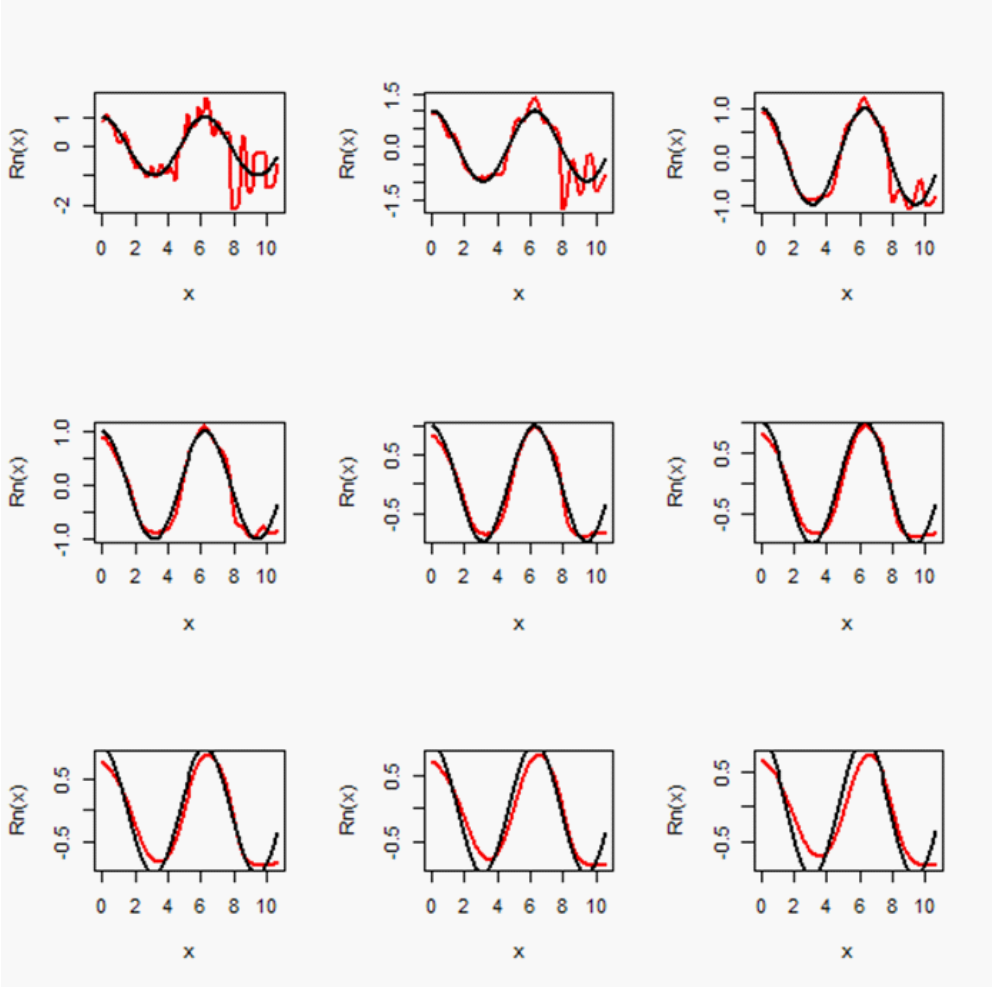

FIG. 3.7 – Régression non linéaire avec h varié, n fixé et K gaussien.

Si nous gardons le même modèle non linéaire $y = \cos x + \varepsilon$, mais avec le noyau d'Epanechikov. On note, que la valeur du h optimale est de $h = 0.9$ (ligne 3, colonne 3, voir la FIG-3.8).

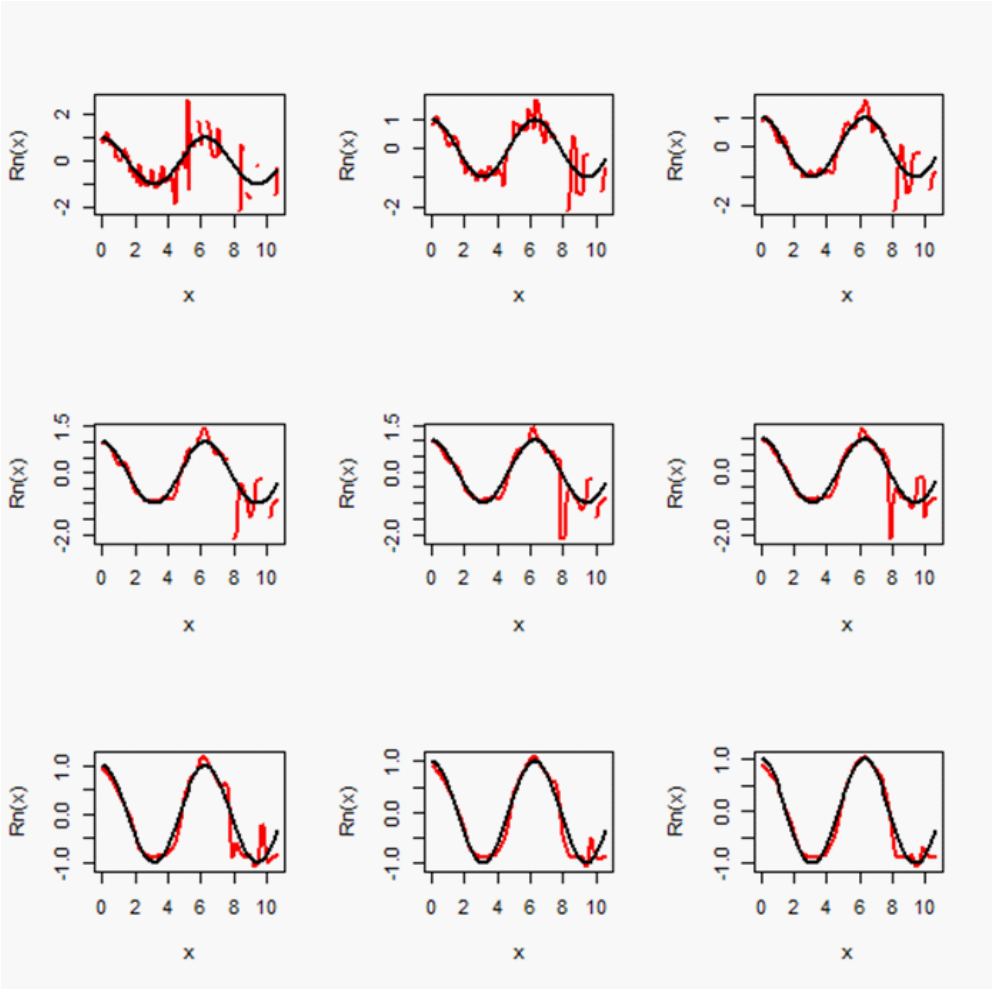

FIG. 3.8 – Régression non linéaire avec h varié, n fixé et K d'Epanechnikov.

Finallement, ce chapitre montre l'importance de paramètre de lissage h et du noyau K dans l'estimation non paramétrique de la régression linéaire et non linéaire. Mais à 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.

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.

L'estimateur à noyau de la régression non paramétrique dépend de deux paramètres le noyau K et le paramètre de lissage h. Dans la pratique, on utilisé le logiciel **R** pour présenté des exemples sur cet estimateur, et à travers les résultats obtenus, 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 .

Bibliographie

[1] Benchoulak, H. (2012). Bandes de confiance pour les fonctions de densité et de régression.

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

[3] Bochner, S. (1955). Harmonic Analysis and the Theory of probability. University of Chicago Press, Chicogo, Illinois.

[4] Bosq, D. and Lecoutre, J. P., (1987). Théorie de l'Estimation Fonctionnelle. Economica, Paris.

[5] Cruz, C. M. R. T. D. (1995). Estimation fonctionnelle : applications aux tests d'adéquation et de paramètre constant (Doctoral dissertation).

[6] Durrieu, G., Nguyen, T. M. N., & Sow, M. (2009). Comparaison d'estimateurs de régression non paramétriques : application en valvometrie. In 41èmes Journées de Statistique, SFdS, Bordeaux.

[7] Ferraty, F. et Vieu, P. (2002/2003). Statistique fonctionnelle : Modèles Non paramétrique de Régréssion. Cours de DEA.

[8] Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, Cambridge.

[9] Kebabi, K. (2014). Estimation non-paramétrique de la fonction de régression.

[10] Matias, C., & Atelier, S. F. D. S. (2013). Introductiona la statistique non paramétrique. Laboratoire Statistique et Génome, Évry, 2014.

[11] Nadaraya, E. A. (1964). On estimating Regression.Theory. Probab. Applic., 9, 141 142.

[12] Nadaraya, E. A. (1989). Nonparametric Estimation of Probability Densities and Regression Curves. Kluwer, Dordrecht.

[13] Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3), 1065-1076.

[14] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 832-837.

[15] Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of points. Annals of Mathematical Statistics, 43.1, 84-88.

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

[17] Watson, G. S. (1964). Smooth Regression analysis. Sankhyà Ser. A, 26, 359-372.

Annexe A : Rappels

A.1- Convergence presque complète : On dit que la suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \longrightarrow \infty$ (et on note $\lim_{n\longrightarrow\infty} X_n = X \quad p.co$), si et seulement si :

$$\forall \varepsilon > 0, \qquad \sum_{n\in\mathbb{N}} P\left[|X_n - X| > \varepsilon\right] < \infty.$$

A.2- Convergence presque sûre : On dit que a suite de v.a $(X_n)_{n\in\mathbb{N}}$ définie sur $(\Omega, \mathcal{A}, P)$, converge presque sûrement $(p.s.)$ vers la variable aléatoire X définie sur $(\Omega, \mathcal{A}, P)$, si

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

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

A.3- Convergence en probabilité : On dit que a suite de v.a (X_n) converge en probabilité vers une v.a. X si, pour tout $\varepsilon > 0$:

$$P\left[|X_n - X| < \varepsilon\right] \longrightarrow 1 \ \text{ quand } n \longrightarrow \infty, \ \text{ on écrit alors, } \ X_n \xrightarrow{\mathcal{P}} X.$$

A.4- Convergence en loi : On dit que la suite de v.a (X_n), de fonction de répartition 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 $n \longrightarrow \infty$.

A.5- Convergence en moyenne quadratique : On dit que la suite de v.a (X_n) converge en moyenne quadratique vers une v.a X si :

$$E\left|X_n - X\right|^2 \longrightarrow 0, \qquad \text{quand } n \longrightarrow \infty, \quad \text{on écrit alors, } \; X_n \xrightarrow{m.q} X.$$

A.6- Théorème (Lois des grands nombres) : Si $(X_1, ..., X_n)$ est un échantillon provenant d'une v.a. X telle que $E\left|X\right| < \infty$, alors :

$$\bar{X}_n \xrightarrow{\mathcal{P}} E\left(X\right) \quad \text{quand } n \longrightarrow \infty, \quad \text{(loi faible)}$$
$$\bar{X}_n \xrightarrow{p.s.} E\left(X\right) \quad \text{quand } n \longrightarrow \infty, \quad \text{(loi forte)}.$$

A.7- 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 : $E\left(\hat{\theta}_n\right) = \theta$.
De même, $\hat{\theta}_n$ est dite asymptotiquement sans biais si pour tout $\theta \in \Theta$:

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

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

A.8- Définition $(o\left(h_n\right),\ O\left(h_n\right))$: Soit x_n et y_n deux suites de nombres réels. Alors, lorsque $n \to \infty$,

$$i) \quad x_n = O(y_n) \Leftrightarrow \limsup_{n\to\infty} |x_n/y_n| < \infty,$$
$$ii) \quad x_n = o(y_n) \Leftrightarrow \lim_{n\to\infty} |x_n/y_n| = 0.$$

Annexe B : Notations et Abréviations

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

$h := h_n$	Paramètre de lissage ou Fenêtre
h_{opt}	Fenêtre optimale
$K(.)$	Noyau
$\mathcal{A}$	Tribu
$\mathbb{E}$	Espace des observations
$\mathcal{B}$	Tribu borélienne
$\mathbb{P}$	une famille de probabilité
Θ	Ensemble des paramètres
L^1	Espace des fonctions intégrables
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
r	Fonction de regression
r_n	Estimateur de r
$1\,(.)$	Fonction indécatrice
$\|.\|$	Norme euclidienne
$\xrightarrow{\mathcal{P}}$	Convergence en probabilité.
$\xrightarrow{\mathcal{L}}$	Convergence en loi.
$\xrightarrow{p.s.}$	Convergence presque sûre.
$\xrightarrow{m.p}$	Convergence en moyenne d'ordre p.
$p.co$	Convergence presque complète.
MSE	L'erreur quadratique moyenne (mean square error).
$AMSE$	L'erreur quadratique moyenne asymptotique (asymptotique mean square error).
$MISE$	L'erreur quadratique intégrée moyenne (mean integrated squared error).