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

BOUREGHISSA Biya

Titre :

Estimation de la fonction de régression par la méthode du noyau

Membres du Comité d'Examen :

M.C.B	**HASSOUNA Houda**	UMKB	Président
M.A.A	**DHIABI Samra**	UMKB	Encadreur
M.A.A	**ROUBI Afef**	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie cette Mémoire

À mes parents.

À mes frères.

REMERCIEMENTS

-Avant tout, nous remercions **ALLAH** le tout puissant qui nous a guidé tout au long de notre vie, nous a permis de nous instruire et d'arriver aussi loin dans nos études, et qui nous a donné courage et patience pour traverser tous les moments difficiles.

-J'exprime ma reconnaissance et toute ma gratitude à Mme Dhiabi Samra pour son encadrement, sa confiance et sa gentillesse. Je remercie également tous les autres membres du jury qui ont acceptés de juger mon travail.

-Un grand merci à tous ceux et celles qui ont contribués de prés ou de loin à la realisation de ce travail.

-A tous mes collègues et amies pour leur encouragement et leur aide.

L'étudiante : Boureghissa Biya.

Table des matières

Table des figures

Introduction

La régression est l'une des méthodes les plus connues et les plus appliquées en statistique. Elle est utilisée pour établir la liaison entre une ou plusieurs variables (qualitatives ou quantitatives) dans un but prédictif. Si on s'intéresse à la relation entre deux variables ; on parle de la régression simple en exprimant une variable en fonction de l'autre. Si la relation porte entre une variable et plusieurs autres variables ; on parle alors de la régression multiple.

Nous cherchons le lien entre Y appeler variable expliquée (dépendante) à l'aide d'autre variable X dites variables explicative (indépendante), modélisé par la relation : $Y = m(X) + \varepsilon$, où ε est l'erreur supposée centrée et indépendante de X. Le problème consiste donc à déterminer pour chaque réalisation x de la variable X, la valeur de la fonction $m(x)$. Pour caractériser cette fonction, une première approche consistait à utiliser un modèle de régression paramétrique. On suppose que cette fonction peut s'écrire comme une fonction explicite des valeurs de X. Cette dernière peut être linéaire, logarithmique,... ; dépendant d'un nombre fini de paramètres que l'on cherche ensuite à estimer par la méthode la plus appropriée (moindre carrées, maximum de vraisemblance,...). L'utilisation d'un modèle paramétrique n'était pas justifiée ; il suffisait alors d'utilisé les données de l'échantillon pour réaliser une estimation. Cela se faisait à l'aide d'un modèle non paramétrique. Dans ce cas on ne dispose d'aucune forme paramétrique pour m. Parmi les méthodes usitées nous présentons dans ce mémoire l'une d'elle qui est : l'estimation par la méthode du noyau (due Nadaraya, 1964 ; Watson, 1964).

Ce mémoire est composé de trois chapitres qui sont :

Premier chapitre : (Quelques notions de bases en estimation non paramétrique)

Dans ce chapitre, nous présenterons les concepts les plus importants de l'estimation non paramétrique dont nous avons besoin dans notre sujet principal comme l'estimation de la fonction de densité par la méthode de noyau, le théorème de Lyapounov ...etc.

Deuxième chapitre : (Estimation non paramétrique de la fonction de régression)

Au départ, on présente d'une manière générale la fonction de régression, son estimation, particulièrement la méthode qui nous interesse le plus ; qui est la méthode de noyau due Nadaraya et Watson (1964), Apres avoir présenté la construction de cet estimateur, ainsi que ses caracteristiques, Nous détaillerons à la fin de ce chapitre les techniques utilisées pour bien choisir les paramètres influents sur cet estimateur, afin de realisé la meilleure estimation possible. la normalité asymptotique de cet estimateur est établi.

Troisième chapitre : (Application sous **R)**

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

Chapitre 1

Quelques notions de bases en estimation non paramétrique

L'estimation consiste à donner des valeurs approchées aux paramètres inconnus d'une population à l'aide d'un échantillon de n observations issues de cette population. On distingue deux approches : approche paramétrique et non paramétrique, dont le but de choisir parmi toutes les statistiques possibles le meilleur estimateur.
On va donner quelques notions élémentaires, définitions et exemples dans l'estimation non paramétrique.

1.1 Estimateur

Définition 1.1.1 (Estimateur) *Nous appelons statistique toute application :*

$$\begin{array}{rccc} S: & \mathbb{R}^n & \longrightarrow & \mathbb{R}^d \\ & X=(X_1,X_2,...,X_n) & \longrightarrow & S(X_1,X_2,...,X_n), \end{array}$$

telle que $S(X_1,X_2,...,X_n)$ soit bien définie et une variable aléatoire (v.a).
Si cette application est utilisée pour évaluer un paramètres inconnus θ (où $\theta \in \mathbb{R}^d$), alors

elle est appelée estimateur, on la note par $\hat{\theta}$ et ses réalisations sont des estimations.

1.1.1 Recherche du meilleur estimateur

On mesure généralement la précision d'un estimateur $\hat{\theta}$ par l'erreur quadratique moyenne, il faut de recherche un estimateur d'erreur quadratique minimale, autrement dit soit sans biais de θ et de variance minimale.

1.1.2 Qualités d'un estimateur

Biais d'un estimateur

On appelle biais de $\hat{\theta}$ pour θ la valeur : $b_{\hat{\theta}} = E[\hat{\theta}] - \theta$.

Si $b_{\hat{\theta}} = E[\hat{\theta}] - \theta = 0$, on dit que θ est sans biais.

On dit que $\hat{\theta}$ est asymptotique sans biais pour θ si : $b_{\hat{\theta}} \underset{n\to\infty}{\longrightarrow} 0$.

On dit que $\hat{\theta}$ est un estimateur convergent si : $E[\hat{\theta}] \underset{n\to\infty}{\longrightarrow} \theta$.

Variance et erreur quadratique moyenne d'un estimateur

On appelle erreur quadratique moyenne de $\hat{\theta}$ par rapport à θ la valeur notée $MSE_{\hat{\theta}}$, définie par : $MSE_{\hat{\theta}} = E[(\hat{\theta} - \theta)^2] = b^2_{\hat{\theta}} + Var(\hat{\theta})$, telle que $Var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$.

On dit $\hat{\theta}$ est un estimateur consistant de θ si : $\hat{\theta} \xrightarrow{\mathcal{P}} \theta$;

si $\hat{\theta}$ est convergent et si $Var(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} 0$ alors $\hat{\theta}$ est consistant.

Convergence d'un estimateur

Lorsque la taille de l'échantillon tend vers l'infini, il faut que l'estimateur se rapproche du paramètre estimé.

-L'estimateur $\hat{\theta}$ est convergent en probabilité si : $\forall \varepsilon > 0 : P(\mid \hat{\theta} - \theta \mid > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0$.

-L'estimateur $\hat{\theta}$ est presque sûrement convergent si : $P(\lim \hat{\theta} \neq \theta) \underset{n\to\infty}{\longrightarrow} 0$.

-L'estimateur $\hat{\theta}$ est convergent en moyenne quadratique si : $MSE_{\hat{\theta}} \underset{n\to\infty}{\longrightarrow} 0$.

1.2 Normalité asymptotique d'un estimateur

Il est très utile de connaître la loi d'un estimateur, elle permet en effet de déduire les caractéristiques de cet estimateur et de construire un intervalle de confiance pour le paramètre. Cette loi peut être détèrminée par l'utilisation de la convergence en loi en général, ou le théorème centrale limite, en particulier.

Théorème 1.2.1 (Centrale limite "TCL") *Si $(X_n)_{n\geq 1}$ une suite de v.a's (i.i.d) d'espérance $\mu < \infty$ et de variance $\sigma^2 < \infty$, alors :*

$$\sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \ \textit{quand } n \to \infty.$$

où $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Nous présentons maintenant le théorème de Lyaponov qui va nous servir pour la démonstration de la normalité asymptotique de l'estimateur présenté dans le deuxième chapitre.

Théorème 1.2.2 (Lyapounov) *Soit $(X_n)_{n\geq 1}$ une suite de v.a's indépendantes de $L^{2+\delta}$ pour un certain $\delta > 0$, définies sur le même espace de probabilité. Supposons que, pour $n \geq 1$, X_n sont centées et ait un écart-type fini σ_n^2, et posons*

$$S_n^2 = \sum_{i=1}^{n} \sigma_i^2 \ \textit{et} \ Z_n = \frac{1}{S_n}\sum_{i=1}^{n} X_i.$$

Il existe $\delta \in]0, \infty[$

$$\lim_{n\to\infty} S_n^{-(2+\delta)} \sum_{i=1}^{n} E|X_i|^{2+\delta} = 0,$$

alors :

$$Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \ \textit{quand } n \to \infty.$$

1.3 Estimation non paramétrique

On s'intéresse dans ce sujet à l'estimation non paramétrique que nous difinissons comme suit :

Définition 1.3.1 (Estimation non paramétrique) *Si l'on sait que h appartient à l'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 pour approcher h.*

1.3.1 Estimation non paramétrique de la densité

On va examiner l'estimation non paramétrique de la fonction de densité. En particulier, l'utilisation de la méthode du noyau, qui servira par la suite à l'estimation de la fonction de régression par la méthode du noyau.
Afin d'estimer la fonction de densité, on sait que l'estimateur naturel de la fonction de répartition F est la fonction de répartition empirique F_n définie par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)}(x), \ \forall x \in \mathbb{R}. \tag{1.1}$$

C'est un estimateur non paramétrique.
On suppose pour la suite que l'on ait $\{x_1, x_2, \ldots, x_n\}$ n observations issues d'une même loi de probabilité de densité f, où f est à support borné $[a, b[$

Histogramme mobile

Soit h un réel strictement positif et assez petit on a bien que

$$f_X(t) = \lim_{h \to 0} \frac{F_X(t+h) - F_X(t-h)}{2h},$$

alors pour estimer f_X il faut estimer F_X par son meilleur estimateur F_n (équation 1.1), ce qui donne :

$$f_{X;n}(t) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{[t-h,t+h[}(x_i).$$

Remarquons que :

$$t - h \leq x_i \leq t + h \Leftrightarrow -1 \leq \frac{x_i - t}{h} \leq 1.$$

D'où :

$$f_{X;n}(t) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{[-1,1[}(\frac{x_i - t}{h}),$$

Cet estimateur est l'estimateur de l'histogramme mobile qui s'écrit alors :

$$f_{X;n}(t) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{K}(\frac{x_i - t}{h}),$$

où :

$$\mathbf{K}(x) = \frac{1}{2}\mathbf{1}_{[-1,1[}(x).$$

Estimation par noyau

Pour améliorer cet estimateur de l'histogramme, remarquons que la classe $[t - h, t + h[$ est centrée en t et tous les points de cette classe ont le même rôle quant au calcul de $f_{X;n}$. L'idée est de pondérer les observations en mettant d'autant plus de poids lorsqu'on se trouve proche de t, et d'autant moins quand on se trouve éloigné.
on a déjà vu un exemple de la fonction de poids $\mathbf{K}$ on choisira alors des fonctions de poids dans des classes plus larges, comprenant notamment des densités à support non borné alors l'estimateur à noyau s'écrit :

$$f_{X;n}(t) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{K}(\frac{x_i - t}{h}).$$

h appelé le paramètre de lissage, et $\mathbf{K}$ est le noyau définie de $\mathbb{R}$ dans $\mathbb{R}$ mesurables et satisfaisant certaines hypothèses basiques parmi celles énoncée ci-dessous :

1. $\mathbf{K}$ est bornée.
2. $\lim_{|u|\to\infty} |u|\mathbf{K}(u) = 0$.
3. $\mathbf{K} \in L_1(\mathbb{R})$.
4. $\int \mathbf{K}(u)du = 1$.

Voici quelques exemples de noyaux classiques (tableau 1.1) :

Uniforme	$\mathbf{K}(x) = \frac{1}{2}\mathbf{1}_{[-1,1[}(x)$.
Triangulaire	$\mathbf{K}(x) = (1- \mid x \mid)\mathbf{1}_{[-1,1[}(x)$.
QuarticouBiweight	$\mathbf{K}(x) = \frac{15}{16}(1-x^2)^2\mathbf{1}_{[-1,1[}(t)$.
Epanechnikov	$\mathbf{K}(x) = \frac{3}{4}(1-x^2)\mathbf{1}_{[-1,1[}(t)$.
Triweight	$\mathbf{K}(x) = \frac{35}{32}(1-x^2)^3\mathbf{1}_{[-1,1[}(t)$.
Normale	$\mathbf{K}(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$.

TAB. 1.1 – Noyaux classiques

Les courbes de ces noyaux sont présentées sur la figure 1.1.

ll reste le choix du noyau $\mathbf{K}$ et du paramètre h.

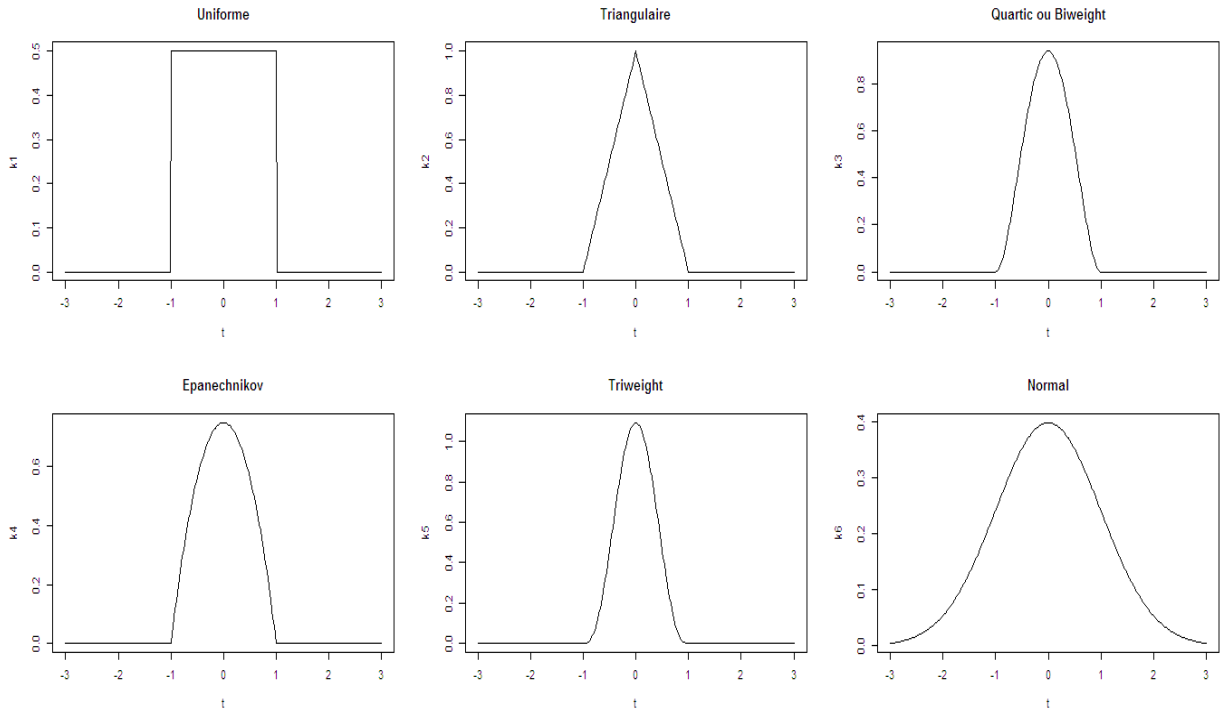

FIG. 1.1 – Courbes des noyaux.

Qualité de l'estimateur de noyau : En raison de l'utilisation de l'estimateur à noyau de la densité, nous présentons quelques propriétés de cet estimateur. Pour plus de détailles on se réfère à ([1],[8]).

Biais :

$$\begin{aligned} b_{f_{X;n}} &= E\left[f_{X;n}(x)\right] - f_X(x) \\ &\approx \frac{1}{2}h^2 f_X''(x) \int t^2 \mathbf{K}(t)dt \end{aligned}$$

Cet équivalent à :

$$E\left[f_{X;n}(x)\right] = f_X(x) + 0(h^2).$$

Variance et l'erreur quadratique :

$$Var\left(f_{X;n}(x)\right) = \frac{1}{nh} f_X(x) \int \mathbf{K}^2(t)dt + 0(\frac{1}{nh}) = O(\frac{1}{nh}).$$

$$\begin{aligned} MSE_{f_{X;n}} &\approx \frac{1}{nh} f_X(x) \int \mathbf{K}^2(t)dt + \left(\frac{1}{2}h^2 f_X''(x) \int t^2 \mathbf{K}(t)dt\right)^2 \\ &= O(\frac{1}{nh}) + 0(h^4). \end{aligned}$$

Convergence : On a $MSE_{f_{X;n}} \underset{n\to\infty}{\longrightarrow} 0$. Cet équivalent à dit $f_{X;n} \xrightarrow{m,q} f_X$, ce implique que $f_{X;n} \xrightarrow{\mathcal{P}} f_X$ et ce dernier implique $f_{X;n} \xrightarrow{\mathcal{L}} f_X$.

Chapitre 2

Estimation non paramétrique de la fonction de régression

Dans l'analyse de régression on considère un vecteur aléatoire (X, Y), où X une variable liée à la variable Y. Notre but est alors de determiner la fonction (mesurable) $m : \mathbb{R} \longrightarrow \mathbb{R}$ tel que $m(X)$ est une bonne approximation de Y, c'est-à-dire (c-à-d) $m(X)$ devrait être proche de Y dans un certain sens, qui est équivalent minimisation de $|m(X) - Y|$. puisque X et Y sont les vecteurs aléatoires, $|m(X) - Y|$ est aléatoire aussi, donc il est difficile d'établir le minimum de $|m(X) - Y|$. Ceci reviens à minimiser le risque L^2 ou l'erreur quadratique moyenne de m

$$E|m(X) - Y|^2,$$

et on exige qu'elle soit aussi petite que possible. Donc, on s'intéresse à une fonction (mesurable) $m : \mathbb{R} \longrightarrow \mathbb{R}$ telle que :

$$E|m(X) - Y|^2 = \min_{f:\mathbb{R}\longrightarrow\mathbb{R}} E|f(X) - Y|^2.$$

On différencie l'espérance $E\left[(m(X)-Y/X=x)^2\right]$ par rapport à $m(.)$ et en égalant le résultat à 0, on obtient :

$$\begin{aligned}\frac{\partial}{\partial m}E\left[(m(X)-Y/X=x)^2\right] &= 2E\left[m(X)-Y/X=x\right]\\ &= 2m(X)-2E\left[Y/X=x\right]\\ &= 0.\end{aligned}$$

Ce qui implique que :

$$m(x)=E[Y\diagup X=x].$$

La dérivée seconde est égale à 2, soit positive nous permet de conclure que c'est bien un minimum.

2.1 Estimation à noyau de la régression

Dans cette section, nous présentons une méthode d'estimation non paramétrique de la fonction de régression nommée "La méthode du noyau".

On dispose d'un couple de v.a's (X,Y) de n observation. On suppose typiquement l'existence d'une fonction $m(.)$ qui exprime la valeur moyenne de la variable expliquée Y en fonction de la variable explicative X.

La forme de régression de Y sur X soit donc défini par :

$$y_i=m(x_i)+\varepsilon_i,\ \text{et}\ i=1,...,n. \tag{2.1}$$

où : ε_i un terme d'erreur, et une variable indépendante de X.

(X,Y) est à valeurs dans $\mathbb{R}^2$; il est supposé admettre une densité jointe sur $\mathbb{R}^2$ notée $f_{(X,Y)}(.,.)$ et nous désignons par $f_X(.)$ la densité marginale (par rapport à la mesure de Lebesgue sur $\mathbb{R}$)

Le principe de la régression à noyau [N W] est d'estimer les fonctions de densités (jointe et marginale) par ses estimateurs du noyau, voir la section 1.3.1.

$$f_{X;n}(x) = \frac{1}{nh}\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h}), \tag{2.2}$$

$$f_{(X,Y);n}(x,y) = \frac{1}{nhk}\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})\mathbf{K}(\frac{y-y_j}{h}).$$

Alors, l'estimateur de Nadaraya (1964) et Watson (1964) pour la fonction inconnue m de 2.1 est défini par :

$$\hat{m}(x) = \sum_{j=1}^{n}\frac{\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}y_j, \tag{2.3}$$

La fonction noyau $\mathbf{K}$ sera supposée mesurable et satisfaisant certaines hypothèses :

h_1 : $\sup_{u\in\mathbb{R}} \mid \mathbf{K}(u) \mid < \infty$.

h_2 : $\lim_{|u|\to\infty} |u|\mathbf{K}(u) = 0$.

h_3 : $\forall u \in \mathbb{R} : \mathbf{K}(u) = \mathbf{K}(-u)$.

h_4 : $\int u^2\mathbf{K}(u)du < \infty$

h_5 : $\int \mathbf{K}(u)du = 1$.

h_6 : $\int \mid \mathbf{K}(u) \mid du < \infty$.

Preuve. On a

$$m(x) = E\left[Y \diagup X = x\right] = \frac{\int y f_{(X,Y)}(x,y)}{f_X(x)}dy = \frac{a(x)}{f_X(x)},$$
$$\Rightarrow \hat{m}(x) = \frac{\hat{a}_n(x)}{f_{X;n}(x)}.$$

alors

$$\begin{aligned}\hat{m}(x) &= \frac{\int y f_{(X,Y);n}(x,y)dy}{f_{X;n}(x)} \\ &= \int y \frac{\frac{1}{nhk}\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})\mathbf{K}(\frac{y-y_j}{k})}{\frac{1}{nh}\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}dy \\ &= \frac{\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}\int \frac{y}{k}\mathbf{K}(\frac{y-y_j}{k})dy \\ &= \frac{\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}\int \frac{y-y_j}{k}\mathbf{K}(\frac{y-y_j}{k})dy + \int \frac{y_j}{k}\mathbf{K}(\frac{y-y_j}{k})dy,\end{aligned}$$

on fait un chengement de variable : $u = \frac{y-y_j}{k}$

$$\begin{aligned}\hat{m}(x) &= \frac{\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}\left\{k\int u\mathbf{K}(u)du + y_j\int \mathbf{K}(u)du\right\} \\ &= \frac{\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}y_j \\ &= \sum_{j=1}^{n}\frac{\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{K}(\frac{x-x_i}{h})}y_j,\end{aligned}$$

$\int u\mathbf{K}(u)du = 0$ (d'après h_3), $\int \mathbf{K}(u)du = 1$ (d'après h_5). ■

2.1.1 Qualités de l'estimateur [N W]

Tout d'abord on va rappelé le lemme de Bochner :

Lemme 2.1.1 (Bochner) *Supposons que g est une fonction bornée sur $\mathbb{R}$, continue dans un voisinage de $x_0 \in \mathbb{R}$ et Q une fonction sur $\mathbb{R}$ tel que :*

$$\int \mid Q(t) \mid dt < \infty,$$

alors :

$$\int \frac{1}{h}Q(\frac{t-x_0}{h})g(t)dt \underset{h\to 0}{\to} g(x_0)\int Q(t)dt$$

Preuve. Voir [10], pages 13-14. ■

Biais

L'estimateur de [N W] de m est un estimateur asymptotique sans biais.

On commence d'abord par donner le théorème du Nadaraya, pour le biais.

Théorème 2.1.1

a. si Y est bornée et si nh tend vers ∞ lorsque n tend vers ∞, ona :

$$E(\hat{m}) = \frac{E\left[\hat{a}_n(x)\right]}{E\left[f_{X;n}(x)\right]} + O(\frac{1}{nh}).$$

si $E(Y^2) < \infty$ et si nh^2 tend vers ∞ lorsque n tend vers ∞, on a :

$$E(\hat{m}) = \frac{E\left[\hat{a}_n(x)\right]}{E[f_{X;n}(x)]} + O(\frac{1}{\sqrt{n}h}).$$

Voir [8], pages 116-117.

Preuve. Au départ, on pose :

$$\bar{\sigma}^2(x) = Var(Y/X = x) = \frac{1}{f_X(x)} \int y^2 f_{(X,Y)}(x,y)dy - m^2(x),\ \varphi(x) = \int y^2 f_{(X,Y)}(x,y)dy$$

et $\kappa = \int \mathbf{K}^2(t)dt.$

En utilisant l'identité suivante :

$$\frac{1}{f_{X;n}(x)} = \frac{1}{E[f_{X;n}(x)]} - \frac{f_{X;n}(x) - E[f_{X;n}(x)]}{E^2[f_{X;n}(x)]} + \frac{(f_{X;n}(x) - E[f_{X;n}(x)])^2}{f_{X;n}(x)E^2[f_{X;n}(x)]}$$

Et en multipliant par $\hat{a}_n(x)$ des deux côtés, et par passage à l'espérance on obtient :

$$\begin{aligned} E(\hat{m}) &= E\left[\frac{\hat{a}_n(x)}{E[f_{X;n}(x)]}\right] - E\left[\frac{\hat{a}_n(x)(f_{X;n}(x) - E[f_{X;n}(x)])}{E^2[f_{X;n}(x)]}\right] + E\left[\frac{\hat{a}_n(x)\left[(f_{X;n}(x) - E\left[f_{X;n}(x)\right])^2\right]}{f_{X;n}(x)E^2[f_{X;n}(x)]}\right] \\ &= \frac{E\left[\hat{a}_n(x)\right]}{E[f_{X;n}(x)]} + \frac{\alpha_n(x) + \beta_n(x)}{E^2[f_{X;n}(x)]} \end{aligned}$$

où : $\alpha_n(x) = E\left[\hat{a}_n(x) f_{X;n}(x)\right] - E[\hat{a}_n(x)]E[f_{X;n}(x)] = cov\left(\hat{a}_n(x), f_{X;n}(x)\right)$
et $\beta_n(x) = E\left[\dfrac{\hat{a}_n(x)}{f_{X;n}(x)}\left[f_{X;n}(x) - E[f_{X;n}(x)]^2\right]\right] = E\left[\hat{m}(x)\left[f_{X;n}(x) - E[f_{X;n}(x)]^2\right]\right]$.
a\ Lorsque la variable Y est bornée, c-à-d : $|Y| \leq M$. On remarque que l'estimateur de [N W] est lui aussi naturellement borné.

$$\hat{m}(x) = \sum_{j=1}^{n} \frac{\mathbf{k}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{k}(\frac{x-x_i}{h})} y_j \leq \frac{\sum_{j=1}^{n}\mathbf{k}(\frac{x-x_j}{h})}{\sum_{i=1}^{n}\mathbf{k}(\frac{x-x_i}{h})} M = M$$

Donc :
$\beta_n(x) = E\left[\dfrac{\hat{a}_n(x)}{f_{X;n}(x)}\left[f_{X;n}(x) - E(f_{X;n}(x))\right]^2\right] \leq M\, E\left[f_{X;n}(x) - E(f_{X;n}(x))\right]^2 = M\, Var(f_{X;n}(x))$.
On utulise le lemme Bochner**,** lorsque h tend vers 0

$$\begin{aligned} Var(f_{X;n}(x)) &= \frac{1}{nh^2}\left\{E\left[\mathbf{K}^2(\frac{x-x_1}{h})\right] - E^2\left[\mathbf{K}(\frac{x-x_1}{h})\right]\right\} \\ &= \frac{1}{nh}\int \frac{1}{h}\mathbf{K}^2(\frac{x-t}{h}) f_X(t)dt - \left(\int \frac{1}{h}\mathbf{K}(\frac{x-t}{h}) f_X(t)dt\right)^2 \\ &= \frac{1}{nh} f_X(x)\kappa + 0(1). \end{aligned} \tag{2.4}$$

Alors :

$$\beta_n(x) \leq MVar(f_{X;n}(x)) \approx \frac{M}{nh} f_X(x)\kappa = O(\frac{1}{nh}). \tag{2.5}$$

Pour $\alpha_n(x)$ on a :

$$\begin{aligned} \alpha_n(x) &= cov\left(\hat{a}_n(x), f_{X;n}(x)\right) \\ &= E\left[\hat{a}_n(x) f_{X;n}(x)\right] - E[\hat{a}_n(x)]E[f_{X;n}(x)] \\ &= \frac{1}{nh^2}E\left[Y\mathbf{K}^2(\frac{x-x_1}{h})\right] - \frac{1}{h^2}E\left[\mathbf{K}(\frac{x-x_1}{h})\right]E\left[Y\mathbf{K}(\frac{x-x_1}{h})\right] \\ &= \frac{1}{nh^2}\int\int y\mathbf{K}^2(\frac{x-t}{h}) f_{(X,Y)}(t,y)dtdy - \frac{1}{h^2}\int \mathbf{K}(\frac{x-t}{h}) f_X(t)dt \int\int y\mathbf{K}(\frac{x-t}{h}) f_{(X,Y)}(t,y)dtdy \\ &= \frac{1}{nh}\int \frac{1}{h}\mathbf{K}^2(\frac{x-t}{h}) a(t)dt - \int \frac{1}{h}\mathbf{K}(\frac{x-t}{h}) f_X(t)dt \int \frac{1}{h}\mathbf{K}(\frac{x-t}{h}) a(t)dt \\ &= \frac{1}{nh} a(x)\kappa + 0(1). \end{aligned} \tag{2.6}$$

Alors

$$\alpha_n(x) \approx \frac{1}{nh} a(x)\kappa = O(\frac{1}{nh}). \tag{2.7}$$

b\ Lorsque $E(Y^2) < \infty$, en utilisant l'inégalité de Cauchy-schwartz

$$\begin{aligned}
\beta_n(x) &= E\left[\frac{\hat{a}_n(x)}{f_{X;n}(x)}\left(f_{X;n}(x) - E[f_{X;n}(x)]\right)^2\right] \\
&\leq E\left\{\max_{1\leq i\leq n}|y_i|\left[f_{X;n}(x) - E\left[f_{X;n}(x)\right]\right]^2\right\} \\
&\leq \left(E\left[\sum_{i=1}^{n} y_i^2\right]\right)^{\frac{1}{2}}\left(E\left[f_{X;n}(x) - E[f_{X;n}(x)]\right]^4\right)^{\frac{1}{2}} \\
&\leq \sqrt{n}\left[E(y_i^2)\right]^{\frac{1}{2}}\left(E\left[f_{X;n}(x) - E[f_{X;n}(x)]\right]^4\right)^{\frac{1}{2}}.
\end{aligned}$$

D'après (1.3.1) (l'erreur quadratique moyenne de l'estimateur à noyau de la fonction de densité), $E\left[f_{X;n}(x) - E[f_{X;n}(x)]\right]^4 = O(\frac{1}{(nh)^2})$.

Alors :

$$\beta_n(x) \leq \sqrt{n}\left(E[Y_i^2]\right)^{\frac{1}{2}} O(\frac{1}{nh}) = O(\frac{1}{\sqrt{n}h}). \tag{2.8}$$

Donc 2.7 et 2.5 et 2.8 achèvent la démonstration. ■

Proposition 2.1.1 *Supposons que $m(.)$ et $a(.)$ sont de classe $C^2(\mathbb{R})$ et que le noyau $\mathbf{K}$ est d'ordre 2, tel que :*

$$\int \mathbf{K}(t)dt = 1, \quad \int t\mathbf{K}(t)dt = 0 \text{ et } \int t^2\mathbf{K}(t)dt < \infty,$$

alors lorsque h tend vers 0 et nh tend vers ∞ on a

$$E[\hat{m}(x)] - m(x) = \frac{h^2}{2}\left\{\left(m''(x) + 2m'(x)\frac{f_X'(x)}{f_X(x)}\right)\int t^2\mathbf{K}(t)dt\right\} + 0(1).$$

Preuve.

$$\begin{aligned} E[\hat{m}(x)] - m(x) &\approx \frac{E\left[\hat{a}_n(x)\right]}{E[f_{X;n}(x)]} - m(x) \\ &= \frac{1}{\frac{1}{h}E[\mathbf{K}(\frac{x-x_1}{h})]} \int\int \mathbf{K}(\frac{x-t}{h}) y f_{(x,y)}(t,y)dtdy - m(x) \\ &= \frac{1}{\frac{1}{h}E[\mathbf{K}(\frac{x-x_1}{h})]} \left\{ \int \mathbf{K}(\frac{x-t}{h}) a(t)dt - a(x) + a(x) - m(x)\frac{1}{h}E[\mathbf{K}(\frac{x-x_1}{h})] \right\}. \end{aligned}$$

En utilisant un changement de variable, on obtient

$$\frac{1}{\int \frac{1}{h}\mathbf{K}(\frac{x-t}{h}) f_X(t)dt} \left\{ \int \mathbf{K}(u)\left(a(x-hu) - a(x)\right) du - m(x) \int \mathbf{K}(u)(f_X(x-hu) - f_X(x))du \right\}.$$

Le lemme de Bochner et un développement de Taylor à d'ordre 2 pour $a(.)$ et $b(.)$ et d'ordre 1 pour $E[\mathbf{K}(\frac{x-X}{h})]$ ce qui donne :

$$\begin{aligned} E[\hat{m}(x)] - m(x) &= \frac{h^2}{2}\frac{1}{f_X(x)}\left[a''(x) - m(x)f_X''(x)\right] \int t^2\mathbf{K}(t)dt + 0(1) \qquad (2.9) \\ &= \frac{h^2}{2}\left(m''(x) + 2m'(x)\frac{f_X'(x)}{f_X(x)} \right) \int t^2\mathbf{K}(t)dt + 0(1). \end{aligned}$$

$\hat{m}$ est asymptotiquement sans biais de m car : $E[\hat{m}(x)] - m(x) \longrightarrow 0$. ■

Consistance

L'estimateur de [N W] de m est un estimateur consistant.

On commence d'abord par donner le théorème du Nadaraya, pour la consiste.

Théorème 2.1.2 *Si $E(Y^2) < \infty$, et si nh tend vers ∞. A chaque point de continuité des fonctions $m(x)$, $f_X(x)$ et $Var(Y/X = x)$ tels que $f_X(x) > 0$, $\hat{m}(x)$ est un estimateur consistant de $m(x)$.*

Voir [8], pages 116-117.

Preuve. On va cité la page 140 dans la thèse [7] :

$$Var\left(\frac{\hat{a}_n(x)}{f_{X;n}(x)}\right) \doteq \frac{E^2\left[\hat{a}_n(x)\right]}{E^2\left[f_{X;n}(x)\right]}\left\{\frac{Var\left(\hat{a}_n(x)\right)}{E^2\left[\hat{a}_n(x)\right]} + \frac{Var\left(f_{X;n}(x)\right)}{E^2\left[f_{X;n}(x)\right]} - 2\frac{cov\left(\hat{a}_n(x), f_{X;n}(x)\right)}{E\left[\hat{a}_n(x)\right]E\left[f_{X;n}(x)\right]}\right\}$$

Nous déduisons du 1.3.1 (l'espérance de l'estimateur à noyau de la densité), 2.4, 2.6,

$$\begin{aligned} Var(\hat{a}(x)) &= Var(\frac{1}{nh}\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})y_j) \\ &= \frac{1}{nh^2}\left\{E\left[Y^2\mathbf{K}^2(\frac{x-x_1}{h})\right] - E^2\left[Y\mathbf{K}(\frac{x-x_1}{h})\right]\right\} \\ &= \frac{1}{nh}\left\{\int \mathbf{K}^2(u)\varphi(x-uh)du - h^2\left(\int \mathbf{K}(u)a(x-uh)du\right)^2\right\} \\ &= \frac{1}{nh}\varphi(x)\kappa + 0(1), \end{aligned}$$

et

$$\begin{aligned} E\left[\hat{a}(x)\right] &= E\left[\frac{1}{nh}\sum_{j=1}^{n}\mathbf{K}(\frac{x-x_j}{h})y_j\right] \\ &= \frac{1}{h}E\left[\mathbf{K}(\frac{x-x_1}{h})Y\right] \\ &= \frac{1}{h}\int\int y\mathbf{K}(\frac{x-t}{h})f_{(X,Y)}(t,y)dtdy \\ &\approx a(x). \end{aligned}$$

que :

$$\begin{aligned} Var(\hat{m}(x)) &= \frac{1}{nh}\frac{a^2(x)}{f_X^2(x)}\left\{\frac{\varphi(x)}{a^2(x)} + \frac{f_X(x)}{f_X^2(x)} - 2\frac{a(x)}{a(x)f_X(x)}\right\}\kappa + 0(1). \qquad (2.10) \\ &= \frac{1}{nh}\frac{\bar{\sigma}^2(x)}{f_X(x)}\kappa + 0(1). \end{aligned}$$

$Var(\hat{m}(x)) \longrightarrow 0$ et $E[\hat{m}(x)] \longrightarrow m(x)$ quand $h \to 0$ et $nh \to \infty$, ce implique que $\hat{m}$ est un estimateur consistant de m. ■

Convergence

D'après 2.10 et 2.9 on obtient

$$MSE_{\hat{m}} = \frac{1}{nh}\left\{\frac{\bar{\sigma}^2(x)}{f_X(x)}\kappa\right\} + \left\{\frac{h^2}{2}\left(m''(x) + 2m'(x)\frac{f_X'(x)}{f_X(x)}\right)\int t^2\mathbf{K}(t)dt\right\}^2 + 0(1). \quad (2.11)$$

$$MSE_{\hat{m}} \longrightarrow 0, \text{ quand } h \to 0 \text{ et } nh \to \infty. \quad (2.12)$$

Lorsque $E\left[(\hat{m}(x) - m(x))^2\right]$ tend vers zéro, ce résultat est démontré (2.12), nous en déduisons $\hat{m}(x) \xrightarrow{L^2} m(x)$ ce qui imlique $\hat{m}(x) \xrightarrow{\mathcal{P}} m(x)$.

2.1.2 Normalitée asymptotique de l'estimateur [N W]

La première démonstration de la normalité asymptotique de l'estimateur [N W] a été fournie par Schuster (1972), qui a étendu le résultat de Nadaraya (1964) dans lequel celui-ci avait montré la normalité asymptotique de

$$\sqrt{nh}\left(\hat{m}(x) - E\left[\hat{m}(x)\right]\right)$$

sous la condition que Y soit borné et que $nh^2 \longrightarrow \infty$.

Pour notre part on a choisi de donner un théorème de normalité en reprenant le résultat de Bierens (1987) [2] qui en donne une preuve différente mais plus facile.

D'après les expressions 2.1, 2.2 et 2.3, nous remarquons que

$$\left(\hat{m}(x) - m(x)\right) f_{X;n}(x) = Q_1 + Q_2 + Q_3,$$

$$Q_1 = \frac{1}{nh}\sum_{i=1}^{n}\varepsilon_i \mathbf{K}\left(\frac{x-x_i}{h}\right),$$
$$Q_2 = \frac{1}{nh}\sum_{i=1}^{n}\left\{(m(x_i)-m(x))\,\mathbf{K}\left(\frac{x-x_i}{h}\right) - E\left[(m(x_i)-m(x))\,\mathbf{K}\left(\frac{x-x_i}{h}\right)\right]\right\},$$
$$Q_3 = \frac{1}{nh}\sum_{i=1}^{n}E\left[(m(x_i)-m(x))\,\mathbf{K}\left(\frac{x-x_i}{h}\right)\right].$$

pour établir la normalité asymptotique de l'estimateur de [N W], nous allons étudier Q_i, $i = 1, 2, 3$.

pour $f_X > 0$ posons $\sigma_\varepsilon^2(x) = E\left[\varepsilon_i / x_i = x\right]$ et pour $p > 0$ si $E|\varepsilon_i|^p < \infty$ posons $\sigma_\varepsilon^p(x) = E\left[|\varepsilon_i|^p / x_i = x\right]$

il existe quelques hypothèses nécessaires à l'établissement de la normalité asymptotique qui sont données ci-dessous :

- (h_1) Pour $p > 0$, $\sigma_\varepsilon^p(x) f_X(x)$ est continue et uniformément bornée sur $\mathbb{R}$.
- (h_2) Il existe $\delta > 0$ tel que $\sigma_\varepsilon^{2+\delta}(x) f_X(x)$ est uniformément bornée sur $\mathbb{R}$.
- (h_3) Les fonctions $m^2(x) f_X(x)$ et $\sigma_\varepsilon^2(x) f_X(x)$ sont continues et uniformément bornées.
- (h_4) Les fonctions $f_X(x)$ et $m(x) f_X(x)$ ainsi que leurs dérivées premières et secondes sont continues et uniformément bornée.
- (h_5) $\int t\mathbf{K}(t)dt = 0$ et $\int t^2\mathbf{K}(t)dt < \infty$.

Alors après étude, nous constatons que

$$\sqrt{nh}\left(\hat{m}(x) - E\left[\hat{m}(x)\right]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_\varepsilon^2(x) f_X(x)\kappa). \tag{2.13}$$

Preuve. nous allons exposer des conditions telles que

$$\sqrt{nh}Q_1 \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_\varepsilon^2(x) f_X(x)\kappa), \tag{2.14}$$

$$\lim_{n\to\infty} E\left[\sqrt{nh}Q_2\right]^2 = 0, \tag{2.15}$$

$$\lim_{n\to\infty}\frac{1}{h^2}Q_3 < \infty. \tag{2.16}$$

Si ces conditions sont remplies, l'équation 2.13 est valide.

D'abord nous prouvons 2.14. Posons que $\upsilon_i(x) = \varepsilon_i \dfrac{\mathbf{K}\left(\frac{x-x_i}{h}\right)}{\sqrt{h}}$, $i = 1, 2, ..., n$. Alors :

$$\sqrt{nh}Q_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\upsilon_i(x).$$

Il suffit de montrer que $\upsilon_i(x)$ satisfait la condition de Lyapounov.

pour $i = 1, 2, ..., n$. Il est clair que $E\left[\upsilon_i(x)\right] = 0$,

$$\begin{aligned}
E\left[\upsilon_i(x)\right]^2 &= E\left[\varepsilon_i^2\frac{\mathbf{K}^2\left(\frac{x-x_i}{h}\right)}{h}\right] \\
&= \frac{1}{h}E\left[E\left[\varepsilon_i^2\mathbf{K}^2\left(\frac{x-x_i}{h}\right)\Big/ x_i\right]\right] \\
&= \frac{1}{h}E\left[\mathbf{K}^2\left(\frac{x-x_i}{h}\right)E\left[\varepsilon_i^2/x_i\right]\right] = \frac{1}{h}E\left[\mathbf{K}^2\left(\frac{x-x_i}{h}\right)\sigma_\varepsilon^2(x_i)\right] \\
&= \frac{1}{h}\int \mathbf{K}^2\left(\frac{x-t}{h}\right)\sigma_\varepsilon^2(t)f_X(t)dt \underset{n\to\infty}{\longrightarrow} \sigma_\varepsilon^2(x)f_X(x)\kappa,
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n}E\left[\frac{\upsilon_i(x)}{\sqrt{n}}\right]^{2+\delta} &= \frac{1}{h}\left(\frac{1}{\sqrt{nh}}\right)^{\delta}E\left[\varepsilon_i^{2+\delta}\mathbf{K}^{2+\delta}\left(\frac{x-x_i}{h}\right)\right] \\
&= \left(\frac{1}{\sqrt{nh}}\right)^{\delta}\frac{1}{h}\int \mathbf{K}^{2+\delta}\left(\frac{x-t}{h}\right)\delta = \sigma_\varepsilon^{2+\delta}(t)f_X(t)dt \\
&= O(\frac{1}{\sqrt{nh}})^{\delta} \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Pour montrer 2.15. Posons $Z_i = (m(x_i) - m(x))\,\mathbf{K}\left(\frac{x-x_i}{h}\right)$, $i = 1, 2, ..., n$.

Alors

$$\begin{aligned}
\sqrt{nh}Q_2 &= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(Z_i - E[Z_i]\right), \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\upsilon_i(x),\ \upsilon_i(x) = \frac{(Z_i - E[Z_i])}{\sqrt{h}}.
\end{aligned}$$

pour $i = 1, 2, ..., n$. Il est clair que $E\left[\upsilon_i(x)\right] = 0$,

$$E\left[\upsilon_i(x)\right]^2 = E\left[\left(\frac{(Z_i - E[Z_i])}{\sqrt{h}}\right)^2\right] = \frac{1}{h}E\left[(Z_i - E[Z_i])^2\right] = \frac{1}{h}\left(E\left[Z_i^2\right] - E^2[Z_i]\right)$$
$$= \frac{1}{h}\left\{\int (m(t) - m(x))^2 \mathbf{K}^2\left(\frac{x-t}{h}\right) f_X(t)dt - \left(\int (m(t) - m(x))\mathbf{K}\left(\frac{x-t}{h}\right) f_X(t)dt\right)^2\right\}$$
$$\underset{n\to\infty}{\longrightarrow} 0.$$

car $m^2(x)f_X(x)$ et $m(x)f_X(x)$ sont uniformément bornées.

Il reste a montrer 2.16, en remarquant que

$$\sqrt{nh}Q_3 = \lambda\frac{1}{h^2}Q_3,\ \lambda = h^2\sqrt{nh}.$$

En faisant un développement de Taylors à l'ordre 2 pour des fonction $m(x-th)f_X(x-th)$ et $f_X(x-th)$ au voisinage de x, on obtient

$$\begin{aligned}
Q_3 &= \frac{1}{nh}\sum_{i=1}^{n} E\left[(m(x_i) - m(x))\mathbf{K}\left(\frac{x-x_i}{h}\right)\right] \\
&= \frac{1}{h}\int (m(t) - m(x))\mathbf{K}\left(\frac{x-t}{h}\right) f_X(t)dt \\
&= \int (m(x-hu) - m(x))\mathbf{K}(u) f_X(x-hu)du \\
&= \int \left[m(x-hu)f_X(x-hu) - m(x)f_X(x-hu)\right]\mathbf{K}(u)du \\
&= \int \left[m(x-hu)f_X(x-hu) - m(x)f_X(x)\right]\mathbf{K}(u)du - m(x)\int \left[f_X(x-hu) - f_X(x)\right]\mathbf{K}(u)du \\
&\approx -h\int u\left(m(x)f_X(x)\right)'\mathbf{K}(u)du + \frac{h^2}{2}\int u^2\left(m(x)f_X(x)\right)''\mathbf{K}(u)du \\
&+ m(x)h\int uf_X'(x)\mathbf{K}(u)du - \frac{h^2}{2}m(x)\int u^2 f_X''(x)\mathbf{K}(u)du.
\end{aligned}$$

posons

$$\lim_{n\to\infty}\frac{1}{h^2}Q_3 = b(x) = \frac{1}{2}(m(x)f_X(x))''\int t^2\mathbf{K}(t)dt - \frac{1}{2}f_X''(x)\int t^2\mathbf{K}(t)dt < \infty.$$

alors

$$\sqrt{nh}Q_3 \approx \lambda b(x).$$

Nous sommes maintenant en mesure d'énoncer le théorème de normalité.

Le cas $\lambda = 0$ donne évidemment le résultat classique

$$\sqrt{nh}\left(\hat{m}(x) - E\left[\hat{m}(x)\right]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_\varepsilon^2(x) f_X(x)\kappa).$$

■

2.1.3 Sélection du noyau et du paramètre de lissage

Comme nous vous le disons l'estimateur de [N W] dépend de deux paramètres : le noyau **K** et la largeur de la fenêtre h.

Sélection du noyau

A cause de la faible influence de la fonction **K** ; la qualité de l'estimateur n'était pas très affectée par le choix des noyaux. Nous utilisons alors les noyaux qu'on a déjà mentionnés.

Sélection du paramètre de lissage

Dans la pratique on aura besoin de décider quelque choix effectuer pour ce paramètre, on se propose à présent de déterminer la fenêtre optimale au sens d'un certain critère d'efficacité asymptotique.

Critère local : MSE L'idée de ce critère de l'erreur quadratique moyenne (MSE) est de recherché la solution à ce probleme :

$$\min_h E\left[\hat{m}(x) - m(x)\right]^2.$$

Sous le équation 2.11, la fenètre optimale au sens du critère MSE est :

$$h_{MSE} = \frac{1}{n^{\frac{1}{5}}} \left\{ \frac{\frac{\bar{\sigma}^2(x)}{f_X(x)}\kappa}{\left(m''(x) + 2m'(x)\frac{f_X{}'(x)}{f_X(x)} \int_{\mathbb{R}} t^2 \mathbf{K}(t)dt \right)^2} \right\}^{\frac{1}{5}}.$$

Critère global : MISE On s'intéresse à l'estimation de la fonction de régression sur un intervalle $I \subseteq \mathbb{R}$, au risque global de l'estimateur [N W]. On introduit pour cela l'erreur quadratique intégrée moyenne (MISE).

$$MISE(h) := E\left[\int_I (\hat{m}(x) - m(x))^2 dx \right],$$

d'après Fubini

$$MISE(h) = \int_I E\left[\hat{m}(x) - m(x)\right]^2 dx.$$

La fenètre optimale au sens de critère MISE :

$$h_{MISE} = \frac{1}{n^{\frac{1}{5}}} \left\{ \frac{\int_I \frac{\bar{\sigma}^2(x)}{f_X(x)}\kappa dx}{\int_I \left(m''(x) + 2m'(x)\frac{f_X{}'(x)}{f_X(x)} \int_{\mathbb{R}} t^2 \mathbf{K}(t)dt \right)^2 dx} \right\}^{\frac{1}{5}}.$$

Remarque 2.1.1 *Les expressions du Biais et de la variance permettent de conclure que :*

* *Une grande valeur de h donne une augmentation du Biais et une diminution de la variance.*
* *Une faible valeur de h donne une diminution du Biais et une augmentation de la variance.*

Chapitre 3

Application sous R

On termine ce mémoire par une étude de simulation ; dont l'objectif est de renforcer les notions que nous avons déjà énumérées dans le chapitre précédent (la grande influence de paramètre de lissage "h", l'importance du noyau "$\mathbf{K}$" et aussi celle de la taille de l'échantillon "n"), utilisant le logiciel d'analyse statistique **R**.

3.1 Présentation des données

Supposons qu'on a observé un échantillon de taille "n" d'un couple de v.a (X, Y), la relation entre x_i et y_i (les valeurs de X et Y respectivement) ; est définie dans le cadre du modèle de régression standard suivant :

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, ..., n.$$

où : ε_i des erreurs centrées et indépendantes de x_i et m est une fonction inconnue que l'on cherche à estimer.

L'estimation non paramétrique d'une fonction de régression par la méthode du noyau [N

W] est définie par la forme suivante :

$$\hat{m}(x) = \sum_{j=1}^{n} \frac{\mathbf{K}(\frac{x-x_j}{h})}{\sum_{i=1}^{n} \mathbf{K}(\frac{x-x_i}{h})} y_j.$$

Le h optimal est défini comme suit : $h_{opt} \simeq Cn^{-\frac{1}{5}}$; pour $C \in \mathbb{R}$.

Afin de montrer graphiquement la convergence de l'estimateur à noyau [N W] de la fonction de régression vers la vraie fonction de régression ; nous supposons que notre modèle à la forme :

$$y = m(x) + \varepsilon, \qquad \varepsilon \rightsquigarrow \mathcal{N}(0,1).$$

et nous avons présenté les résultats obtenus pour les différents jeux de données ainsi que pour les différentes valeurs de h strictement positif (h fixé ou h varié), différents noyaux $\mathbf{K}$ (à support non compact et à support compact), régression linéaire et non linéaire.

Dans les résultats graphique de cette section :

-La droite noire exprime la fonction de régression $m(x)$.

-La droite en rouge exprime l'estimation de la fonction de régression par noyau $\hat{m}(x)$.

3.2 Etude du régression linéaire

On veut estimer le modèle de régression suivant :

$$Y = 7 - 0.5X + \varepsilon, \qquad X \rightsquigarrow \mathcal{N}(1,2).$$

Et effectuant les variations suivantes :

3.2.1 "h" fixée, "K" fixée et "n" variée

Cette partie de simulation consiste à apprécier la qualité de l'estimation du premier modèle (linéaire) de régression pour diférentes tailles d'échantillon "50, 100 et 500".

Choisissant pour cela : $h = Cn^{-\frac{1}{5}}$, $C = 1$.

1. **"K"** fixé, à support non compact.

   $\mathbf{K}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$, (Normale). Effectuant un code **R**, il résulte la figure 3.1.

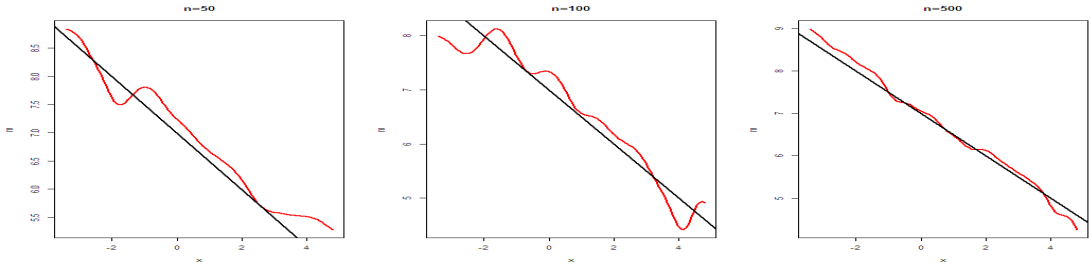

FIG. 3.1 – Régression linéaire : h fixé, K noyau Normale et n varié.

2. **"K"** fixé, à support compact.

   $\mathbf{K}(x) = \frac{3}{4}(1-x^2)\mathbf{1}_{[-1,1[}(t)$, (Epanechnikov). Effectuant un code **R**, il résulte la figure 3.2.

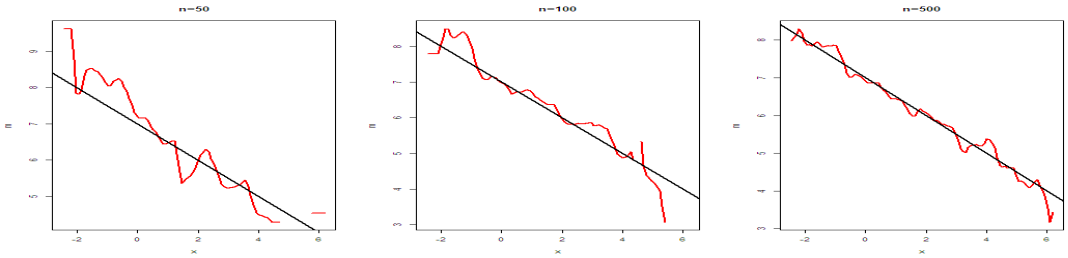

FIG. 3.2 – Régression linéaire : h fixé, K noyau Epanechnikov et n varié.

$\mathbf{K}(x) = \frac{1}{2}\mathbf{1}_{[-1,1[}(x)$., (Uniforme). Effectuant un code **R**, il résulte la figure 3.3.

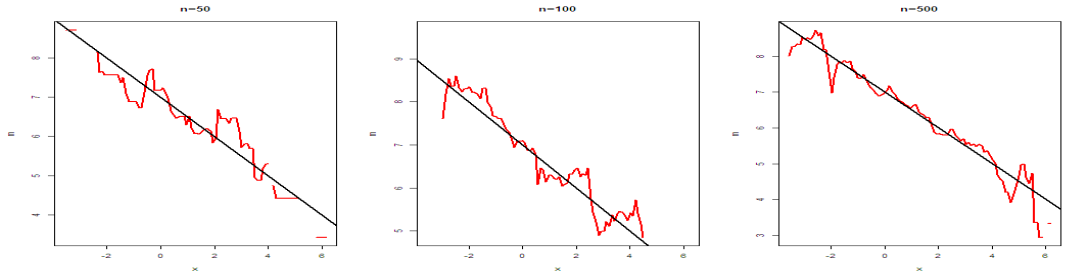

FIG. 3.3 – Régression linéaire : h fixé, K noyau uniforme et n varié.

Remarque 3.2.1 *-On remarque graphiquement que le graphe rouge de $\hat{m}$ est approche beaucoup à la droite noire de m dans le troisième graphe, donc ces graphes expriment la convergence de l'estimateur $\hat{m}$ vers m pour n assez grand et pour n'importe quel noyau utilisé.*

- le changement du noyau n'a pas fait une grande influence sur la convergence de l'estimateur.

-Si nous gardons le mème modèle linéaire ; mais avec X suit un autre loi (exponentielle, uniforme, ...). On arive au mème conclusion de la convergence.

3.2.2 "n" fixéé", K" fixée et "h" variée

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 fixé la taille de l'échantillon $n = 300$.

l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans les figures 3.4 (avec noyau Normale "à support non compact") et 3.5 (avec noyau Triweight "à support compact").

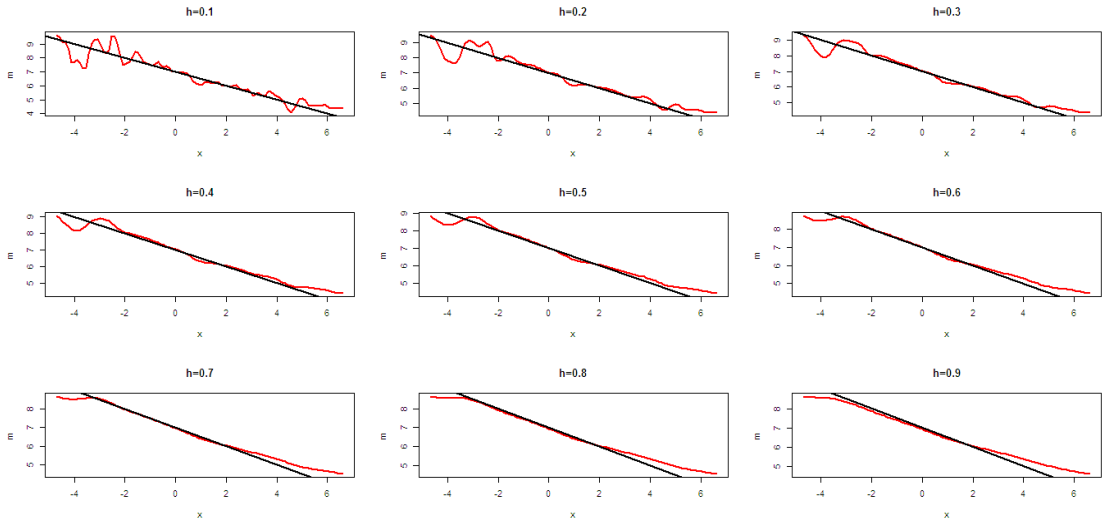

FIG. 3.4 – Régression linéaire : K noyau Normale, n fixée et h varié.

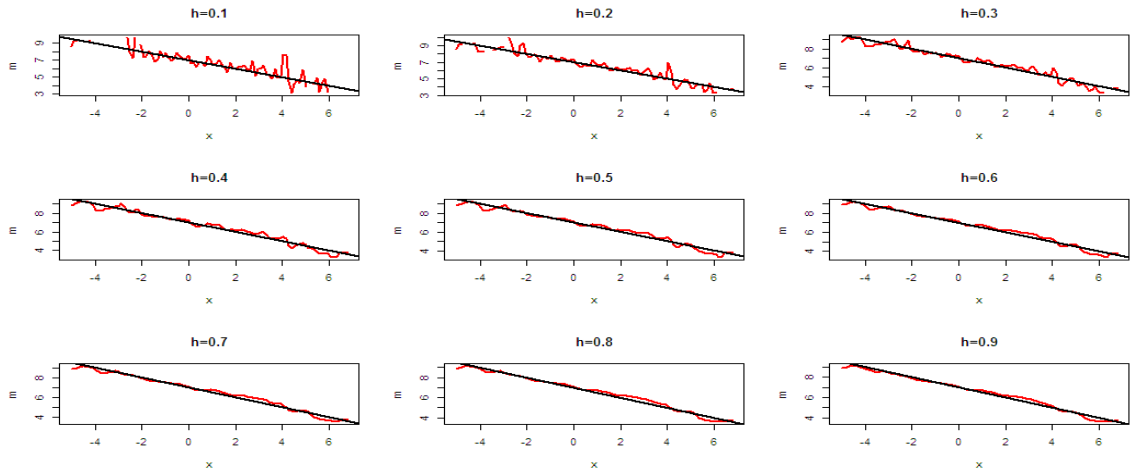

FIG. 3.5 – Régression linéaire : K noyau Triweight, n fixée et h varié.

Remarque 3.2.2 *-La variation du paramètre h a une influence remarquable sur la convergence de l'estimateur de la fonction de régression vers la vraie fonction m.*

-Le h optimal est situé dans l'intervalle $[0.4, 0.6]$ *pour la noyau Normale et dans l'intervale* $[0.6, 0.8]$ *pour la noyau Triweight.*

-Si nous gardons le mème modèle linéaire ; mais avec **K** *la noyau Triangulaire ; Le h optimal est situé dans l'intervalle* $[0.7, 0.9]$.

3.3 Etude du régression non linéaire

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

$$Y = sin(0.2X) + cos(3\pi - X) + \varepsilon, \qquad X \rightsquigarrow \mathcal{U}[-5, 3].$$

Et effectuant les variations suivantes :

3.3.1 "h" fixée, "K" fixée et "n" variée

Dans ce premier cas, le paramètre de lissage ou la fenètre h est fixé $h = n^{-\frac{1}{5}}$ et n varié "50, 100 et 500".

1. **"K"** fixé, à support non compact.

   $\mathbf{K}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$, (Normale). Effectuant un code **R**, il résulte la figure 3.6.

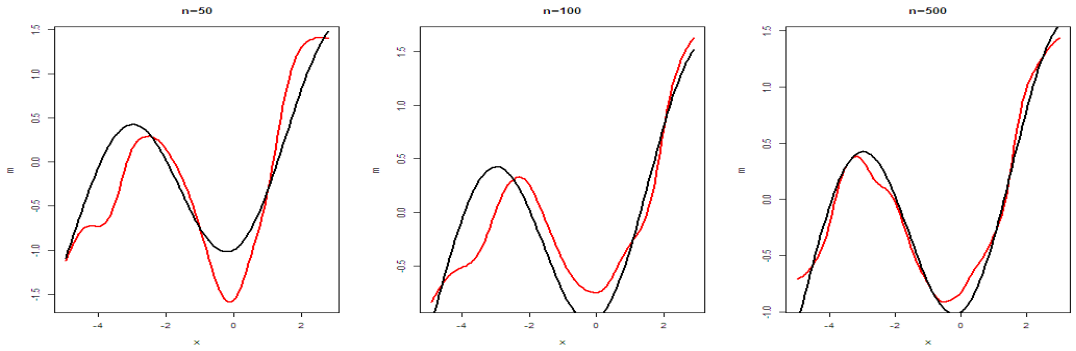

FIG. 3.6 – Régression non linéaire : K noyau Normale, n fixée et h varié.

2. **"K"** fixé, à support compact.

   $\mathbf{K}(x) = (1- \mid x \mid)\mathbf{1}_{[-1,1[}(x)$. (Triangulaire). Effectuant un code **R**, il résulte la figure 3.7.

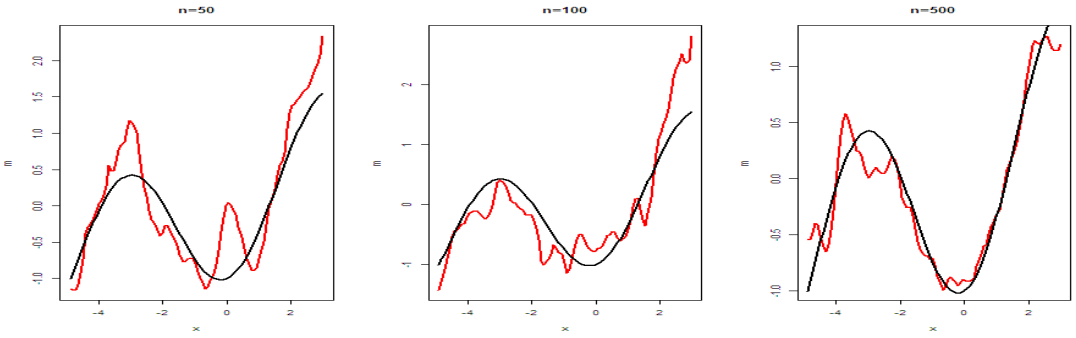

FIG. 3.7 – Régression non linéaire : h fixé, K noyau Triangulaire et n varié.

   $\mathbf{K}(x) = \frac{15}{16}(1-x^2)^2\mathbf{1}_{[-1,1[}(t)$. (Biweight). Effectuant un code **R**, il résulte la figure 3.8.

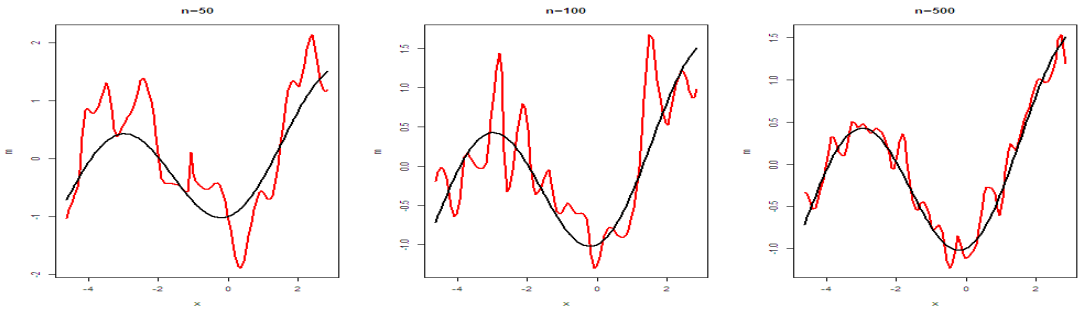

FIG. 3.8 – Régression non linéaire : h fixé, K noyau Biweight et n varié.

Remarque 3.3.1 *-La mème conclusion pour le cas non linéaire que le cas linéaire (c-à-d, convergence de l'estimateur pour n assez grand et pour n'importe quel noyau utilisé).*
-Si nous gardons le mème modèle non linéaire ; mais avec X suit un autre loi (erxponentielle, normale,...) ; on arive au mème conclusion de la convergence de l'estimateur.

3.3.2 "n" fixéé", K" fixée et "h" variée

De mème facon que pour la régression linéaire, 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 fixé la taille de l'échantillon $n = 300$.

L'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans les figures 3.9 (avec noyau Normale "à support non compact") et 3.10 (avec noyau Biweight "à support compact"). Le h optimal est situé dans l'intervalle $[0.4, 0.6]$ pour la noyau Normal et dans l'intervale $[0.6, 0.8]$ pour la noyau Biweight.

Si nous gardons le mème modèle non linéaire ; mais avec **K** la noyau Uniforme ; Le h optimal est situé dans l'intervalle $[0.7, 0.9]$.

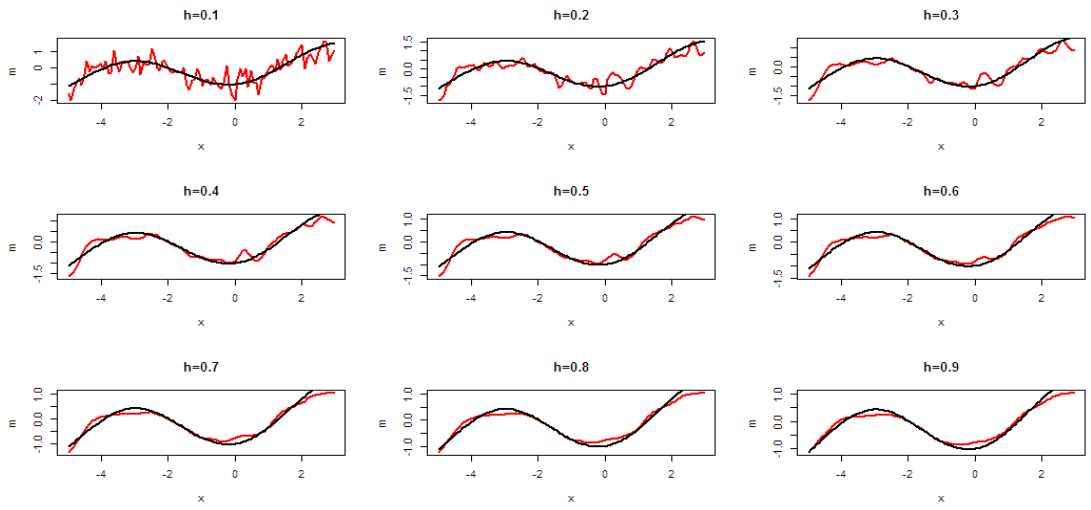

FIG. 3.9 – Régression non linéaire : K noyau Normale, n fixée et h varié.

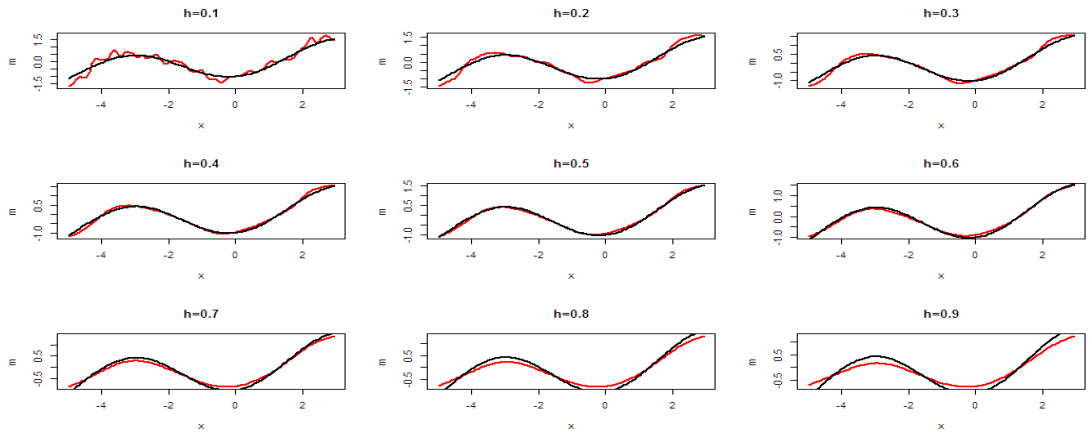

FIG. 3.10 – Régression non linéaire : K noyau Biweight, n fixée et h varié.

conclusion 3.3.1 *Ce chapitre montre l'importance de paramètre de lissage h et du noyau* **K** *dans l'estimation à noyau de la régression linéaire et non linéaire. Mais à noté que le choix de h est plus important que celui de noyau.*

Conclusion

Dans ce mémoire, nous étudierons une méthode d'estimation non paramétrique de la régression nommée "la méthode du noyau", cette méthode est trés pratique lorsque l'on s'intéresse à la relation entre une variable expliquée Y et une variable explicative X, mais que l'on ne veut supposer aucune forme particulière pour la relation entre ces deux variables, laissant ainsi aux données le choix exclusif de cette forme. La méthode du noyau est une méthode qui est communément utilisée pour faire de la régression non paramétrique.

Ce mémoire a démontré que la méthode du noyau vraiment un bon estimateur, convergent et simple à utiliser, et pour conffirmer notre étude nous avons fait des simulations des données par le logiciel **R** qui a effectivement validé nos résultats.

Bibliographie

[1] Banon, G. (1976). Sur un estimateur non paramétrique de la densité de probabilité. Revue de statistique appliquée, 24(4), 61-73.

[2] Bierens, H. J. (1987). Kernel estimators of regression functions. In Advances in econometrics : Fifth world congress (Vol.1, pp.99-144).

[3] Carbon, M. & Francq, C. (1995). Estimation non paramétrique de la densité et de la régression-Prévision non paramétrique. La Revue de Modulad, ISSN 1145-895X, 15, 1, 25.

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

[5] Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2006). A distribution-free theory of nonparametric regression. Springer Science & Business Media.

[6] Härdle, W. (1990). Applied nonparametric regression. Cambridge university press.

[7] Kiessé, T. S. (2008). Approche non-paramétrique par noyaux associés discrets des données de dénombrement (Doctoral dissertation, Université de Pau et des Pays de l'Adour).

[8] Nadaraya, E. A. (1989). Nonparametric estimation of probability densities and regression curves. Springer.

[9] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

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

Annexe A : Logiciel *R*

Le code **R** qui permet de calculer l'estimateur à noyau de la fonction de régression est :

- Dans le modèle linéaire : $X \rightsquigarrow \mathcal{N}(1,2)$, $\varepsilon \rightsquigarrow \mathcal{N}(0,1)$, $n = 50$, $h = n^{-\frac{1}{5}}$, et $K =$ "normale".

  $n = 50\,;\ x = rnorm(n,1,2)\ e = rnorm(n)\,;\ y = -0.5 * x + 7 + e$

  #Génère un échantillon suit la loi $\mathcal{N}$(1,2), un erreur suit la loi $\mathcal{N}$(0,1) de taille n=50 et crée un modèle linéaire.

  $h = n\hat{}(-1/5)$ # Le paramètre de lissage optimale.

  $s = 100\,;\ t = seq(min(x), max(x), length = s)$

  # Génère un intervale de min(x) jusqu'a max(x) de de largeur $s = 100$.

  $k = function(t)\{(1/sqrt(2 * pi)) * exp(-0.5 * t\hat{}2)\}$ # La noyau normale.

  $f = numeric(n)\,;\ fn = numeric(s)\,;\ a = numeric(n)\,;\ an = numeric(s)$

  # Crée des vecteurs de taille n et s.

  $for(j\ in\ 1:s)\{for(i\ in\ 1:n)\{f[i] = k((t[j] - x[i])/h)\}; fn[j] = sum(f)\}$ # L'estimateur de f.

  $for(j\ in\ 1:s)\{for(i\ in\ 1:n)\{a[i] = y[i] * k((t[j] - x[i])/h)\}; an[j] = sum(a)\}$ # L'estimateur de a.

  $mn = an/fn$ # L'estimateur de la fct de régression.

  $plot(t, mn, type = "l", col = 2, lwd = 2, xlab = "x", ylab = "m", main = "n = 50")$

  # Représente graphiquement l'estimateur à noyau de la fonction de régression

  $abline(7, -0.5, lwd = 2)$ # Représente graphiquement la vraie fonction de régression.

- Dans le modèle non linéaire : $X \rightsquigarrow \mathcal{U}[-5,3]$, $\varepsilon \rightsquigarrow \mathcal{N}(0,1)$, $n = 50$, $h = n^{-\frac{1}{5}}$, et $K =$ "normale"

  $n = 50\,;\ x = runif(n,-5,3)\ e = rnorm(n)\,;\ y = sin(0.2 * x) + cos(3 * pi - x) + e$

  # Génère un échantillon suit la loi $\mathcal{U}$[-5,3], un erreur suit la loi $\mathcal{N}$(0,1) de taille n=50 et crée un modèle non linéaire.

  $h = n\hat{}(-1/5)$

  $s = 100\,;\ t = seq(min(x), max(x), length = s)$

  $k = function(t)\{(1/sqrt(2 * pi)) * exp(-0.5 * t\hat{}2)\}$

  $f = numeric(n)\,;\ fn = numeric(s)\,;\ a = numeric(n)\,;\ an = numeric(s)$

  $for(j\ in\ 1:s)\{for(i\ in\ 1:n)\{f[i] = k((t[j] - x[i])/h)\}; fn[j] = sum(f)\}$

  $for(j\ in\ 1:s)\{for(i\ in\ 1:n)\{a[i] = y[i] * k((t[j] - x[i])/h)\}; an[j] = sum(a)\}$

  $mn = an/fn$

  $plot(t, mn, type = "l", col = 2, lwd = 2, xlab = "x", ylab = "m", main = "n = 50")$

  # Le mème code **R.**

  $lines(t, sin(0.2 * t) + cos(3 * pi - t), lwd = 2)$

  # Représente graphiquement la vraie fonction de régression.

- Nous devons jouer avec les donneés pour les différentes valeurs de h, différents noyaux et les différents taille de l'échantillon.

  $par(mfrow = c(3,3))$ # Découpe la fenètre graphique en 9 (3 lines et 3 colomns).

  $k = function(t)\{((3/4) * (1 - t\hat{}2)) * ifelse(abs(t) <= 1, 1, 0)\}$ # La noyau epanechnikov.

  $k = function(t)\{((35/32) * (1 - t\hat{}2)\hat{}3) * ifelse(abs(t) <= 1, 1, 0)\}$ # La noyau triweight.

  $k = function(t)\{(1 - abs(t)) * ifelse(abs(t) <= 1, 1, 0)\}$ # La noyau triangulaire.

  $k = function(t)\{((15/16) * (1 - t\hat{}2)\hat{}2) * ifelse(abs(t) <= 1, 1, 0)\}$ # La noyau biweight.

Annexe B : Abréviations et Notations

Les différentes abréVariations 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.
$\xrightarrow{\mathcal{P}}$	convergence en probabilité.
$\xrightarrow{\mathcal{L}}$	convergence en loi, convergence en distribution.
$\xrightarrow{m.q}$	convergence en moyenne quadratique.
$\xrightarrow{L^2}$	convergence en moyenne quadratique.
$\hat{\theta}$	estimateur de θ.
$MSE_{\hat{\theta}}$	erreur quadratique moyenne de l'estimateur $\hat{\theta}$.
$b_{\hat{\theta}}$	biais de l'estimateur $\hat{\theta}$.
$\doteq$	indique un équivalent asymptotique.
$:=$	indique égale par définition.
$X \rightsquigarrow F_\theta$	X suit la loi F_θ.
$i.i.d$	indépendantes de mème loi.
$c - \grave{a} - d$	c'est-à-dire.
$[N\ W]$	Nadaraya et Watson.