République Algériennne 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 pour obtenir le diplôme de

MASTER en "**Mathématiques Appliquées**"

Option : **statistique**

Par :

BAIA Ikram

Titre :

Erreur quadratique moyenne intégrée asymptotique de l'estimateur à noyau de la densité

Membres du Comité d'Examen

Dr.	SAYAH Abdallah	MCA	UMKB	Président
Dr	KHEIREDDINE Souraya	MCB	UMKB	Rapporteur
Dr.	DHIABI Samra	MAA	UMKB	Examinateur

Juin 2022

Dédicace

Je dédie ce travail à mon papa "**Baia Hocine",**

et ma mamam **"Boularassi Warda",**

à mon mari "**Khmouli Nabil**",

à mon frére "**Yassine",**

à mes soeurs "**Wiam, Farah, Ikhlasse",**

à mon défunt grand-pére "**Boularassi Hamid",**

et au défunt professeur "**khmouli Saad",**

à mes amies "**Chaima,Fatima",**

et à toute ma famille "**Baia, Boularassi, Khmouli**"

IKRAM

Remerciements

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

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

Mes vifs remerciements aux membres du Jury ; Mr. Sayah Abdallah. et Mme. Dhiabi Samra pour l'intérêt qu'ils ont porté à mon mémoire en acceptant de l'examiner.

Je tiens à remercier toute ma famille, mes amies et mes condisciples de la promotion 2022.

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

Résumé

Dans ce mémoire, nous étudions l'estimateur non paramétrique de la fonction de densité par la méthode du noyau. La construction de l'estimateur est basée sur l'utilisation d'une densité K appelée noyau et d'un paramètre de lissage h. Nous rappelons les propriétés de l'estimateur. Nous parlons aussi du choix de noyau et de paramètre de lissage. Finalement, nous donnons des explications graphiques des résultats théoriques appliqués sur des exemples de densité à l'aide du logiciel R.

Notations et Abréviations

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

$i.i.d$	indépendantes et indentiquement distribuées
E, μ	Espérance mathématique (moyenne)
Var	Variance mathématique
F	La fonction de répartition
F_n	La fonction de répartition empirique
Γ	La fonction gamma
$\hat{f}_n$	L'estimateur de f
$Biais$	Le Biais de l'estimateur
$\overline{X}$	La moyenne empirique
σ	L'ecart type
K	Le noyau

h	Le paramétre de lissage
MSE	erreur quadratique moyenne
$MISE$	erreur quadratique moyenne intégrée
$AMISE$	erreur quadratique moyenne intégrée asymptotique
$AMSE$	erreur quadratique moyenne asymptotique
$\sim$	équivalance en loi de probabilité
$I_{\{X_i \leq x\}}$	fonction indicatrice de l'ensemble $\{X_i \leq x\}$

Table des matières

Table des figures

Liste des tableaux

Introduction

Lorsque l'on cherche à étudier une suite de mesures provenant de la répétition d'une expérience, une méthode de modélisation consiste à supposer que ces mesures sont des réalisations des variables aléatoires indépendantes et identiquement distribuées. Comprendre ces mesures et la façon dont elles sont distribuées revient à étudier la loi de probabilité de la variable aléatoire sous-jacente.

Par exemple, en médecine, on cherche à étudier l'assimilation d'un traitement antibiotique. Pour cela, on mesure, pour chaque patient $j = 1, ..., n$*, la concentration de l'antibiotique qui est passée dans le sang du patient après* 6 *heures (temps moyen d'absorption). On modélise le phénomène de la manière suivante :* $x_1, ..., x_n$ *sont les réalisations de* n *variables aléatoires indépendantes* $X_1, ..., X_n$ *ayant même densité* f*. Dans ce contexte médical, comprendre le processus d'assimilation de l'antibiotique dans le sang revient à connaître* f*.*

La densité de probabilité permet de résoudre de nombreux problèmes autres que le calcule de la moyenne et de la variance, les testes d'ajustement,... Elle permette le déplacement à l'estimation de la fonction de répartition, de régression, et des densités et quantiles conditionnelles et elle est appliquée aussi en prévision non paramétrique.

L'estimation à noyau est une méthode non paramétrique basée sur l'utilisation d'une fonction appelée noyau et d'un paramètre de lissage ou fenêtre, ce dernier est beaucoup plus déterminant pour l'obtention des bons estimateurs. Historiquement, les premiers travaux

qui portent sur l'estimation à noyau de la densité sont ceux de Rosenblatt (1956), et Parzen (1962)..

En plus, ce mémoire est rédigé en trois chapitres :

Chapitre un : On commence ce chapitre par un rappel des définitions et on parle sur l'estimation paramétrique et non paramétrique .

Le deuxième chapitre porte sur l'estimation à noyau de la densité (Rosenblatt, 1956 et Parzen, 1962) , et nous présentons les propriétés asymptotiques de l'estimateur, choix optimal du paramètre de lissage et du noyau....

Finalement, le troisième chapitre.illustré des simulations effectuées en utilisant le Logiciel ***R*** *. Les codes* ***R*** *utilisés sont donnés avec les sorties graphiques correspondantes.*

Chapitre 1

Estimation fonctionnelle

On commence ce chapitre par un rappel des définitions et on parle sur l'estimation paramétrique et non paramétrique

1.1 Concepte de base

1.1.1 Variable aléatoire

Définition 1.1.1 *On appele une variable aléatoire, l'application mesirable de* (Ω, A) *dans* U *noté, et selon l'ensemble d'arrivé* U*, on dit que* X *est discréte si* U *est fini ou au plus dénombrable* $(U \subseteq N^*)$ *elle est dit continue si* U *est une partie de* $\mathbb{R}$ *ou* $\mathbb{R}$ $(U \subseteq \mathbb{R})$*.*

1.1.2 Caractéristiques d'une varaible aléatoire

Espérance mathématique

L'esperance d'une variable aléatoire notée $E(X)$ est définie comme suit :

$$E(X) = \int_R x f(x) dx$$

Variance et L'écart type

$Var(X)$ représente la variance de X telle que :

$$\begin{aligned} Var(X) &= E\left[(X - E(X))^2\right] \\ &= E\left(X^2\right) - E(X)^2 \end{aligned}$$

L'écart type σ soit la racine carrée de la variance est une mesure de la v.a et peut être vu comme une mesure de risque. En finance, des rendement avec un plus grand écart type sont souvent vue comme plus risqué.

1.1.3 Echantillonnage

Nous allons étudier comment se comporte un échantillon (élément pris au hasard) dans une population dont on connait les caractéristiques (lois,...) d'une varaible considérée X. Dans ce cas, prendre un échantillon aléatoire de taille n consiste a considérer n réalisation de X ou encore considérer n varaibles aléatoires $X_1, ..., X_n$ indépendantes, de même loi que X.

Soit X une variable aléatoire sur un référentiel Ω. Un échantillon de X de taille n est un n-uplet $(X_1, ..., X_n)$ de varaibles aléatoires indépendantes de même loi que X . La loi de X sera applée loi même. Une réalisation de cet échantillon est un n-uplet de réels $(x_1, ..., x_n)$ ou $X_i(\omega) = x_i$.

1.1.4 Fonction de répartition

Soit X une variable aléatoire à valeurs dans un intervalle I de la forme $[a, b]$, qui suit une loi de probabilité P, on appelle fonction de répartition de X la fonction F, $F(X) = P(X \leq x) = \int_a^x f(t)\, dt$.

1.1.5 Fonction de répartition empirique

Soit $X_1, X_2, ..., X_n$ des variables aléatoires indépendantes suivant la même loi de fonction de répartition F.

On définit la fonction de répartition empirique de $X_1, X_2, ..., X_n$ par :

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

$$= \begin{cases} 0 & \text{si} \quad x \prec X_{(1)} \\ \frac{k}{n} & \text{si} \quad X_{(k)} \leq x \leq X_{(k+1)} \quad k = 1, ..., n-1 \\ 1 & \text{si} \quad x \geq X_{(n)} \end{cases}$$

1.1.6 Densité de probabilité

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

$$\int_I f(t)\, dt = 1$$

$$f(x) = \frac{\partial F(x)}{\partial x}$$

1.2 Quelques Lois de probabilité

Loi de probabilité	support	densité	Espérence	variance
Exponentielle(θ)	$\mathbb{R}_+$	$\frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$	θ	θ^2
Gamma(k,θ)	$\mathbb{R}_+$	$\frac{x^{k-1}\exp\left(-\frac{x}{\theta}\right)}{\theta^k\Gamma(k)}$	$k\theta$	$k\theta^2$
Qui-deux(k)	$\mathbb{R}_+^*$	$\frac{x^{\frac{k}{2}}\exp(-\frac{x}{2})}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}$	k	$2k$
Poisson(λ)	$\mathbb{N}$	$\frac{\lambda}{k!}\exp(-\lambda)$	λ	λ

TAB. 1.1 – Quelques Lois de probabilité

Notation 1 $\Gamma(x) = \int_0^{+\infty} t^{x-1}\exp(-t)$ *c'est la fonction gamma.*

1.3 Estimation paramétrique et Estimation non paramétrique

1.3.1 Estimation paramétrique

Si l'on sait à priori que h appartient à une famille paramétrée $\{h(x,\theta), \theta \in \Theta\}$ ou $\Theta \subset \mathbb{R}^s$ et $h(.,.)$ est une fonction connue, on parle alors d'estimation paramétrique, car éstimer h revient à éstimer le paramétre fini-dimensionnel θ .

1.3.2 Estimation non paramétrique

Si l'on sait seulement que h appartient àappartient à P ensemble des lois de probabilités qui est un éspace de dimension infinie, alors on dit que l'on fait de l'estimation non

paramétrique ou de l'estimation fonctionnelle.

Dans ce qui suit,on suppose que l'on a observé un échantillon $X_1, ..., X_n$ à valeurs dans $\mathbb{R}^s$ muni de sa tribu borélienne β. De plus,on suppose que les $\{X_i, i = 1, ..., n\}$ sont indépendantes et identiquemment distribuées (i.i.d) $\mu \in \rho_0$ une famille de loi sur $(\mathbb{R}^s, \beta)$.

Définition 1.3.1 *On appelle un estimateur de θ toute fonction mesurable f de $X = (X_1, ..., X_n)$ dans Θ (ouvert dans $\mathbb{R}$), autrement dit :*

$$\hat{f}_n : X \to \Theta$$
$$X \to \hat{f}_n(X)$$

$\hat{f}_n(X)$ s'appelle l'estimateur de θ.

Propriétés d'un estimateur :

Soit $\widehat{\theta}$ un estimateur de θ , lorsque la taille n de l'échantillon intervient dans les propriétés, nous noterons $\widehat{\theta}_n$.

la fct $\widehat{\theta}(x)$ est un densité de probabilité.

Démenstration la somme continue de $\hat{f}_n(x)$ sur le support $\mathbb{R}$ est :

$$\begin{aligned}\int_{\mathbb{R}} \hat{f}_n(x)\, dx &= \int_{\mathbb{R}} \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) dx \\ &= \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - X_1}{h}\right) dx,\end{aligned}$$

on pose $t = \frac{x - X_1}{h}$ et donc $dx = hdt$,

alors,

$$\begin{aligned}\int_{\mathbb{R}} \hat{f}_n(x)\,dx &= \int_{\mathbb{R}} \frac{1}{h} K(t)\,h dt \\ &= \int_{\mathbb{R}} K(t)\,dt = 1\end{aligned}$$

Estimateur avec biais : Un estimateur $\hat{f}_n$ de θ est dit avec biais si pour tout $\theta \in \Theta$ (Θ ouvert de $\mathbb{R}$) et tout entier positif n,

$$E(\hat{f}_n) = \theta + Biais(\hat{f}_n)$$

La quantité $Biais(\hat{f}_n)$ est le biais de l'estimateur $\hat{f}_n$.

Par exemple : $Y = \frac{1}{n+1}\sum_{i=1}^{n} X_i$, est un estimateur biaisé de $E(X)$ car :

$$\begin{aligned}E(Y) &= E(\frac{1}{n+1}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n+1} E(\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n+1}\sum_{i=1}^{n} E(X_i) \\ &= \frac{n}{n+1} E(X_i)\end{aligned}$$

Estimateur sans biais : Un estimateur $\hat{f}_n$ de θ et $\theta \in \Theta$ est dit sans biais si :

$$Biais(\hat{f}_n) = 0 \;\text{ alors } E(\hat{f}_n) = \theta$$

Par exemple : la moyenne empirique $\left(\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\right)$ est un estimateur sans biais de la moyenne μ.

Estimateur asymptotiquement sans biais :

Un estimateur $\hat{f}_n$ de θ est dit asymptotiquement sans biais si :

$$\lim_{n\to\infty} Biais(\hat{f}_n) = 0 \text{ alors } \lim_{n\to\infty} E(\hat{f}_n) = \theta$$

Par exemple : $\widetilde{S} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ est un estimateur asymptotiquement sans biaisé de σ^2 .

1.4 Estimateur paramétrique de la fonction de densité

Il y a plusieurs approches pour estimer une densité à partir de données, Il y a l'approche paramétrique, où l'objectif est d'estimer les paramètres d'une distribution connue.

Un estimateur f est une fonction $\hat{f}_n(x, X_1, ..., X_n)$ mesurable par rapport à l'observation $(X_1, ..., X_n)$. Si l'on sait à priori que f appartient à une famille paramétrique $\{f(x,\theta), \theta \in \Theta\}$ ou $\Theta \in \mathbb{R}^d$ et $f(.,.)$ est une fonction continue, on parle alors d'estimation paramétrique,car estimer f revient à estimer le paramétre fini dimensionnel θ.

Nous associons maintenant un modèle statistique P_θ qui dépend d'un paramètre θ. Pour se faire une idée de la valeur inconnue du paramétre θ , à partir des observation $(X_1, ..., X_n)$ qui sont i.i.d, on calcule ensuite une certaine valeur numérique, que l'on considérera comme valeur approchée de θ qu'on appellera une estimateur de θ .

Dans ce cas ou'il n'y a pas d'estimateur évident, on cherche un estimateur par la méthode de vraisemblance, ou par la méthode des moments,...etc

1. Soit un échantillon issu d'une v.a.r X normale de fonction de densité f qui dépend de deux paramètres qui sont inconnus (μ, σ^2) :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Pour estimer la fonction de densité f il se fait d'estimer le paramètre inconnue $\theta = (\mu, \sigma^2)$, où μ est la moyenne de X et σ^2 est sa variance.

On a dans ce cas :

$$\widehat{\mu} = \overline{X} \text{ et } \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

et l'estimateur de $\theta = (\widehat{\mu}, \widehat{\sigma}^2) = \left(\overline{X}, \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)$.

Donc l'estimateur $\hat{f}_n$ de f est donné :

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{\sqrt{2\pi}\sqrt{\widehat{\sigma}^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\widehat{\mu}}{\sqrt{\widehat{\sigma}^2}}\right)\right] \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}} \exp\left[\frac{-1}{2}\left(\frac{x-\overline{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}\right)\right]\end{aligned}$$

2. Soit X un v.a.r suit la loi exponentielle de fonction de densité $f(x) = \lambda \exp(-\lambda x)$ avec $\lambda > 0$ paramètre inconnu. Pour estimer la fonction de densité f il se fait d'estimer le paramètre inconnue λ.

L'estimateur de λ est $\widehat{\lambda} = \frac{1}{\overline{X}}$.

Donc l'estimateur $\hat{f}_n$ de f est donné :

$$\begin{aligned}\hat{f}_n(x) &= \widehat{\lambda} \exp\left(-\widehat{\lambda}x\right) \\ &= \frac{1}{\overline{X}} \exp\left(-\frac{1}{\overline{X}}x\right)\end{aligned}$$

1.5 Théorèmes de convergences de variables aléatoires

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

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

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

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

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

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

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

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

5.Théorème (Lois des grands nombres) : Si $(X_1, X_2, ..., X_n)$ est un échantillon pro-

venant d'une v.a. X telle que $\mathbb{E}|X| < \infty$, alors :

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

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

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

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

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

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

Ces quantités vérifies les assertions :

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

Chapitre 2

Estimation non paramétrique de la fonction de densité

Ce *chapitre porte sur l'estimation à noyau de la densité (Rosenblatt, 1956 et Parzen, 1962) , et nous présentons les proprietés asymptotiques de l'estimateur, choix optimal du paramètre de lissage et du noyau...*

2.1 Estimateur à noyau de la densité

Définition 2.1.1

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

avec

$$W(y) = \begin{cases} 1/2 & si \quad y \in [-1, 1[\\ 0 & sinon\,; \end{cases}$$

la densité de probabilité uniforme sur l'intervalle $[-1, 1[$*. Cet estimateur peut être généralis en remplaçant la fonction de poids* $W(\cdot)$ *(la densité de probabilité uniforme) par une*

fonction de poids plus générale K. Ceci résulte en l'estimateur

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.1}$$

Notation 2 K { *La fonction de poids* ; *Le noyau*

K une densité de probabilité symétrique.

h { *Le paramétre de lissage* ; *La fenetre*

Les estimateurs par noyaux introduits ci-dessous sont très réguliers et on peut exhiber des bornes précises sur leur risque, c'est ce qui va nous conduire à nous concentrer sur eux jusqu'à la fin du texte. On ne considère ici que des noyaux positifs, à savoir, des fonctions $K : \mathbb{R} \to \mathbb{R}^+$ telles que ; $\int K d\lambda = 1$;

dont on supposera en outre qu'elles admettent un moment d'ordre deux (sans que cette condition fas :se partie de la définition d'un noyau),

et

$$\sigma^2 = K^2 d\lambda < \infty \ .$$

ou les noyaux uniformes donnés par des lois uniformes,

$$K_{(a,b)} = \frac{1}{b-a} I_{[a,b]}$$

Soient $X_1, ..., X_n$ des v.a i.i.d de densité de probabilité f 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

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

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

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

quand $n \to \infty$. Donc, F_n est une estimateur convergent (consistant) de F. De plus, la convergence presque sûre est uniforme en $x \in \mathbb{R}$ d'aprés le Théorème de Glivenko-Cantelli .

Comment peut-on estimer ? Une des premières solution intuitives a été proposée par Rosenblatt (1956). pour $h > 0$ assez petit on a

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

En remplaçant ici F par l'estimateur F_n, on obtient

$$\hat{f}_n^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

La fonction $\hat{f}_n^R$ est un estimatuer de f appelé estimatuer de Rosenblatt.

On peut aussi l'écriren sous la forme

$$\begin{aligned}\hat{f}_n^R(x) &= \frac{1}{2nh}\sum_{i=1}^{n} I(x-h < X_i \leq x+h) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),\end{aligned}$$

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

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{2.2}$$

où $K : \mathbb{R} \to \mathbb{R}$ est un fonction intégrable, 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 la paramètre h fenêtre de l'estimteur.

Dans le cadre symptotique où $n \to \infty$ ou supposera que la fenêtre h dépend de n et on la notera h_n, la suite $(h_n)_{n\geq 1}$ tondant vers 0 lorsque $n \to \infty$. La notation h, sans indice n, sera également utilisée afin d'abréges l'écriture lorsqu'il n'y sura pas d'ambiguïté.

Dans la suite, définissons le produit de convolution par

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

L'estimateur de Parzen -Rosenblatt définie en (2.2) apparaît bien comme la densité obtenue en régularisant la mesure empirique

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

par convolution avec $\frac{1}{h} K\left(\frac{\cdot}{h}\right) =: K_h(\cdot)$, où δ est la masse de Dirac. Alors,

$$\hat{f}_n(x) = K_h * \mu_n(x) \qquad pour\ x \in \text{IR}. \tag{2.3}$$

Si le noyau K est positif alors $\hat{f}_n(x)$ est une densité de probabilité.

Définition 2.1.2 *Un noyau est dit de Parzen-Rosenblatt si :*

$$\lim_{\|x\| \to \infty} \|x\| K(x) = 0,$$

où $||\cdot||$ *est la norme euclidienne.*

Lemme 2.1.1 (Bochner) *1) Soit* K *un noyau de Parzen -Rosenblatt et* $f \in L^1$ *alors en tout point* x *de continuité de* f *on a*

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

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

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

2.2 Propriétés de l'estimateur à noyau de la densité

Les propriétés des noyaux :

1. pour tout x "positivité" : $K(x) \geq 0$
2. "densité" $\int K(x)dx = 1$
3. "symétrique" $\int xK(x)dx = 0$
4. "la variance de K finie" $\int t^2 K(t)\, dt < +\infty$
5. "carré intégrable" $\int K^2(t)\, dt < +\infty$

Si $\bullet K$ densité $\Longrightarrow$ $\hat{f}_n$densité.

$\bullet K$ continue $\Longrightarrow$ $\hat{f}_n$ continue.

$\bullet K$ différentiable $\Longrightarrow$ $\hat{f}_n$ différentiable.

$\bullet K$ non négatives $\Longrightarrow$ $\hat{f}_n$ non négatives.

2.2.1 Biais ponctuel

Le biais ponctuel mesure la différence entre la valeur moyenne de l'estimateur $\hat{f}_n$ est la valeur de la fct inconnue f en un point x.

$$Biais\left(\hat{f}_n(x)\right) = E\left(\hat{f}_n(x)\right) - f(x)$$

Soit x fixé dans $\mathbb{R}$. Le biais de l'estimteur à noyau présenté dans (2.2) est

$$Biais\left(\hat{f}_n(x)\right) = \frac{1}{2}h^2 f''(x)\int_{\mathbb{R}} t^2 K(t)\,dt$$

Démonstration. Comme les variables aléatoires $X_1, X_2, ..., X_n$ sont i.i.d., nous avons successivement ■

$$\begin{aligned}
E\left(\hat{f}_n(x)\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}E\left(\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) \\
&= E\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right) \\
&= \int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-X_1}{h}\right)f(x_1)\,dx_1
\end{aligned}$$

Nous effectuons le changement de variables suivant : $-t = \frac{x-x_1}{h}$, d'où $x_1 = ht + x$. De là, en utilisant l'hypothèse $K(-x) = K(x)$, le biais de $\hat{f}_n(x)$ s'exprime ainsi par

$$\begin{aligned}
Biais\left(\hat{f}_n(x)\right) &= \int_{\mathbb{R}} K(-t)\,f(x+ht)\,dt + f(x) \\
&= \int_{\mathbb{R}} K(t)\,f(x+ht)\,dt + f(x)
\end{aligned}$$

Dans le but d'avoir une forme plus simple, qui ne dépend que du paramètre h, nous approximons la formule du biais en utilisant la formule de Taylor-Lagrange :

$$f(x+ht) = f(x) + htf'(x) + \frac{h^2t^2}{2}f''(x) + o\left(h^2t^2\right)$$

Ainsi, nous obtenons

$$Biais\left(\hat{f}_n(x)\right) = f(x)\int_{\mathbb{R}} K(t)\,dt + hf'(x)\int_{\mathbb{R}} tK(t)\,dt + \frac{h^2}{2}f''(x)\int_{\mathbb{R}} t^2K(t)\,dt - f(x) + o\left(h^2\right)$$

D'après les propriétés de noyau K (3) , (4) et (5) nous avons finalement

$$Biais\left(\hat{f}_n(x)\right) = \frac{h^2}{2}f''(x)\int_{\mathbb{R}} t^2K(t)\,dt + o\left(h^2\right).$$

2.2.2 Variance ponctuelle

Propriété Soit x fixé dans $\mathbb{R}$. La variance de l'estimateur $\hat{f}_n$ est

$$Var\left(\hat{f}_n(x)\right) = \frac{1}{nh}f(x)\int_{\mathbb{R}} K(t)^2\,dt + o\left(\frac{1}{h}\right).$$

Démonstration. Partant de l'hypothèse d'indépendance entre les X_i,nous avons

$$\begin{aligned} Var\left(\hat{f}_n(x)\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right) \\ &= \frac{1}{n}Var\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right) \\ &= \frac{1}{n}E\left[\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right)^2\right] - \frac{1}{n}E\left[\left(\frac{1}{h}K\left(\frac{x-X_1}{h}\right)\right)\right]^2 \\ &= \frac{1}{n}\int_{\mathbb{R}}\frac{1}{h^2}K^2\left(\frac{x-x_1}{h}\right)f(x_1)\,dx_1 - \frac{1}{n}\left(\int_{\mathbb{R}}\frac{1}{h}K\left(\frac{x-x_1}{h}\right)f(x_1)\,dx_1\right)^2 \end{aligned}$$

Nous effectuons le changement de varaible $-t = \frac{x-x_1}{h}$.

Nous trouvons

$$\begin{aligned} Var\left(\hat{f}_n(x)\right) &= \frac{1}{nh^2}\int_{\mathbb{R}}K(-t)^2 f(ht+x)\,hdt - \frac{1}{n}\left(\int_{\mathbb{R}}K(-t)\,f(ht+x)\,hdt\right)^2 \\ &= \frac{1}{nh}\int_{\mathbb{R}}K(t)^2 f(ht+x)\,hdt - \frac{1}{n}\left[Biais\left(\hat{f}_n(x)\right) + f(x)\right]^2 \\ &= \frac{1}{nh}\int_{\mathbb{R}}K(t)^2 f(ht+x)\,hdt - \frac{1}{n}\left(o\left(h^2\right) + f(x)\right)^2 \end{aligned}$$

Finalment, sous la condition d'avoir $\int K(t)^2\,dt < \infty$ et pour n grand, nous avons

$$Var\left(\hat{f}_n(x)\right) = \frac{1}{nh}f(x)\int_{\mathbb{R}}K(t)^2\,dt.$$

■

2.3 Erreur Quadratique Moyenne

Lorsqu'on définit l'estimateur à noyau, il y a un certain nombre de critères qui permettent d'évaluer la similarité de l'estimateur $\hat{f}_n$ par rapport à la vraie densité f à estimer. Parmi les nombreux critères proposés dans la littérature, la formule de l'Erreur Quadratique

Moyenne ponctuelle (MSE).

Ainsi, pour l'estimateur $\hat{f}_n$ définie en (2.2) on a :

$$MSE = E\left[\left(\hat{f}_n(x) - f(x)\right)^2\right], \qquad x \in \mathbb{R} \tag{2.4}$$

Il est intéressant de développer cette expression pour faire apparaitre la variance et le biais ponctuels de l'estimateur $\hat{f}_n$ notés $Var\left(\hat{f}_n(x)\right)$ et $Biais\left(\hat{f}_n(x)\right)$ définis respectivement en (2.5) et (2.6) pour tout point $x \in \mathbb{R}$par :

$$Var\left(\hat{f}_n(x)\right) = E\left(\left[\hat{f}_n(x) - E\left(\hat{f}_n(x)\right)\right]^2\right) \tag{2.5}$$

$$Biais\left(\hat{f}_n(x)\right) = E\left(\hat{f}_n(x)\right) - f(x) \tag{2.6}$$

Ceci conduit à une nouvelle expression du $MSE(\hat{f}_n(x))$ comme suit

$$\begin{aligned} MSE\left(\hat{f}_n(x)\right) &= E\left[\left(\hat{f}_n(x)\right)^2\right] - 2E\left(\hat{f}_n(x)\right) f(x) + \left[E\left(f_n(x)\right)\right]^2 \\ &= E\left[\left(\hat{f}_n(x)\right)^2\right] - \left[E\left(\hat{f}_n(x)\right)\right]^2 + \left[E\left(\hat{f}_n(x)\right)\right]^2 - 2f(x)\, E\left(\hat{f}_n(x)\right) + f^2(x) \\ &= Var\left(\hat{f}_n(x)\right) + Biais^2\left(\hat{f}_n(x)\right) \end{aligned} \tag{2.7}$$

Cette dernière expression montre bien le compromis pour la minimisation du MSE entre le biais (erreur systématique) et la variance (erreur aléatoire). A cause des termes $E\left\{\hat{f}_n(x)\right\}$ dans (2.5) et (2.6), on voit qu'une réduction de biais entraîne une augmentation de la variance et vice versa. Cependant, on peut constater que l'espérance et la variance de l'estimateur $\hat{f}_n$ en un point x de R peuvent être exprimées en fonction du noyau classique

K et de la densité à estimer f comme suit :

$$E\left(\hat{f}_n(x)\right) = \frac{1}{h}E\left(K\left(\frac{x-t}{h}\right)\right) = \frac{1}{h}\int_T K\left(\frac{x-t}{h}\right) f(t)\, dt \tag{2.8}$$

et

$$\begin{aligned} Var\left(\hat{f}_n(x)\right) &= \frac{1}{nh^2} Var\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right) \\ &= \frac{1}{nh^2}\left[E\left(K^2\left(\frac{x-X_1}{h}\right)\right) - E^2\left(K\left(\frac{x-X_1}{h}\right)\right)\right] \\ &= \frac{1}{nh^2}\int_T K^2\left(\frac{x-t}{h}\right) f(t)\, dt - \frac{1}{nh^2}E^2\left(K\left(\frac{x-X_1}{h}\right)\right) \end{aligned} \tag{2.9}$$

Il découle de (2.8) que le biais ne dépend pas directement de la taille de l'échantillon, mais plutôt du noyau. Par conséquent, l'obtention d'une estimation asymptotiquement sans biais implique un ajustement du noyau K et de paramètre de lissage h.

Une mesure, globale, de l'efficacité de l'estimateur $\hat{f}_n$ est obtenue en intégrant le MSE sur tout le support $\mathbb{R}$ de f. Il s'agit de "l'Erreur Quadratique Moyenne Intégrée" (MISE). En utilisant (2.7), elle s'écrit :

$$\begin{aligned} MISE\left(\hat{f}_n(x)\right) &= \int_{\mathbb{R}} MSE\left(\hat{f}_n(x)\right) dx \\ &= \int_{\mathbb{R}} Var\left(\hat{f}_n(x)\right) dx + \int_{\mathbb{R}} Biais^2\left(\hat{f}_n(x)\right) dx. \end{aligned} \tag{2.10}$$

Il est possible de reporter dans (2.10), les valeurs de la variance et du biais définies dans (2.5) et (2.6), mais cela va rendre les calculs lourds et conduit à des résultats rarement exploitables. D'où l'intérêt de faire quelques approximations et hypothèses supplémentaires

pour garder les résultats généraux. Ainsi, en dépit des conditions vérifiées par le noyau classique, nous supposons dans la suite que la densité à estimer f admette des dérivées de tous ordres.

En reprenant l'expression de l'espérance de l'estimateur $\hat{f}_n$ dans (2.8) et la variance dans (2.9) puis en posant $t = x - hu$, on a :

$$E\left(\hat{f}_n(x)\right) = \int_{\mathbb{R}} K(u) f(x - hu)\, du \tag{2.11}$$

et

$$Var\left(\hat{f}_n(x)\right) = \frac{1}{nh^2}\int_{\mathbb{R}} K^2(u) f(x - hu)\, du - \frac{1}{nh^2} E^2\left(K\left(\frac{x - X_1}{h}\right)\right) \tag{2.12}$$

Le développement en séries de Taylor de $f(x - hu)$ au voisinage de x est alors

$$f(x - hu) = f(x) - huf'(x) + \frac{1}{2}h^2u^2f''(x) + o\left((hu)^2\right)$$

En l'injectant dans (2.11) et (2.12) on obtient alors :

$$\begin{aligned} E\left(\hat{f}_n(x)\right) &= \int_{\mathbb{R}} K(u)\left(f(x) - huf'(x) + \frac{1}{2}h^2u^2f''(x)\right)dx + o\left((hu)^2\right) \\ &= f(x) + \frac{1}{2}h^2f''(x)\,\sigma_K^2 + o\left(h^2\right) \end{aligned} \tag{2.13}$$

et

$$Var\left(\hat{f}_n(x)\right) = \frac{1}{nh}\int_{\mathbb{R}} K^2(u) f(x)\, du + R_n + o\left(\frac{1}{nh}\right)$$
$$= \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(u)\, du + o\left(\frac{1}{nh}\right) \tag{2.14}$$

avec $R_n = \left(\frac{1}{nh^2}\right)\left(-huf'(x) + \frac{1}{2}h^2u^2f''(x) - E^2\left[K\left\{\frac{x-X_1}{h}\right\}\right]\right) \simeq o\left\{\frac{1}{nh}\right\}$ .De ces deux derniers résultats (2.14) et (2.13), on peut déduire les formes approximées et asymptotiques du MSE et du $MISE$ notées respectivement $AMSE$ et $AMISE$ par :

$$AMSE\left(\hat{f}_n(x)\right) = \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(u)\, du + \frac{1}{2}h^4\sigma_K^4 f''^2(x) \tag{2.15}$$

et

$$AMISE\left(\hat{f}_n(x)\right) = \frac{1}{nh}\int_{\mathbb{R}} K^2(u)\, du + \frac{1}{4}h^4\sigma_K^4 \int_{\mathbb{R}} f''^2(x)\, dx \tag{2.16}$$

2.4 Expressions asymptotiques du MSE/MISE

$$MSE\left[\hat{f}_n(x)\right] = \frac{1}{nh} f(x) R(K) + \frac{h^4}{4}\left(f''(x)\right)^2 \mu_2^2 + 0\left(h^4 + \frac{1}{nh}\right)$$

et

$$MISE\left[\hat{f}_n(x)\right] = \frac{1}{nh} R(K) + \frac{h^4}{4}\int \left(f''(x)\right)^2 dx \mu_2^2 + 0\left(h^4 + \frac{1}{nh}\right)$$

tq : $R(K) = \int K^2(x)\, dx$

2.5 Choix de paramètre de lissage

Ceci implique une possibilité pour choisir le paramétre de lissage h par minimisation du AMSE/AMISE :

Pour l'AMSE, $h = h(x)$, paramétre de lissage variable (locale), si $f''(x) \neq 0$:

$$h_{AMSE}(x) = \left(\frac{f(x) R(k)}{(f''(x))^2 \mu_2^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Pour l'AMISE, h paramètre de lissage constant (globale) :

$$h_{AMISE}(x) = \left(\frac{R(K)}{R(f'') \mu_2^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

avec

$$\lim_{n \to \infty} \frac{h_{AMSE}}{h_{AMISE}} = 1.$$

Insérer la bardwidth $h_{opt} = cn^{-\frac{1}{5}}$ dans le MISE donne un taux de convergence de l'ordre

$$MISE(\hat{f}_n) \sim n^{-\frac{1}{5}}$$

Pour un noyau avec

$$\int uK(u)\, du = 0$$

$$MISE(\hat{f}_n) \sim n^{-\frac{2}{3}}$$

2.6 Choix du noyau

Dans la ittérature, il existe plusieurs fonctions qui jouent le role d'un noyau, les plus usuels sont (voir Silverman, 1986) :

– Noyau Triangulaire : $K(t) = (1 - |t|)\,\mathbf{1}_{\{|t|\leq 1\}}$,

– Noyau Biweight : $K(t) = \frac{15}{16}\left(1 - t^2\right)^2 \mathbf{1}_{\{|t|\leq 1\}}$,

– Noyau Gaussien : $K(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-t^2/2\right\},\ \ t \in \mathbb{R}$

– Noyau d'Epanechnikov : $K(t) = \frac{3}{4}\left(1 - t^2\right)\,\mathbf{1}_{\{|t|\leq 1\}}$,

La figure (Fig.**2.1**) si-après présente l'allure des quatre noyaux cités.

Code R :

```
K1=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
K2=function(t){(15/16)*((1-t^2)^2)*ifelse(abs(t)<=1,1,0)}
K3=function(t){dnorm(t)}
K4=function(t){ifelse(abs(t)<1,(3/4)*(1-t^2),0)}
op=par(mfrow=c(2,2))
curve(K1(x),-1,1,ylab="K(x)",main="Triangulaire")
curve(K2(x),-1,1,ylab="K(x)",main="Biweight")
curve(K3(x),-4,4,ylab="K(x)",main="gaussien")
curve(K4(x),-1,1,ylab="K(x)",main="Epanechikov")
par(op)
```

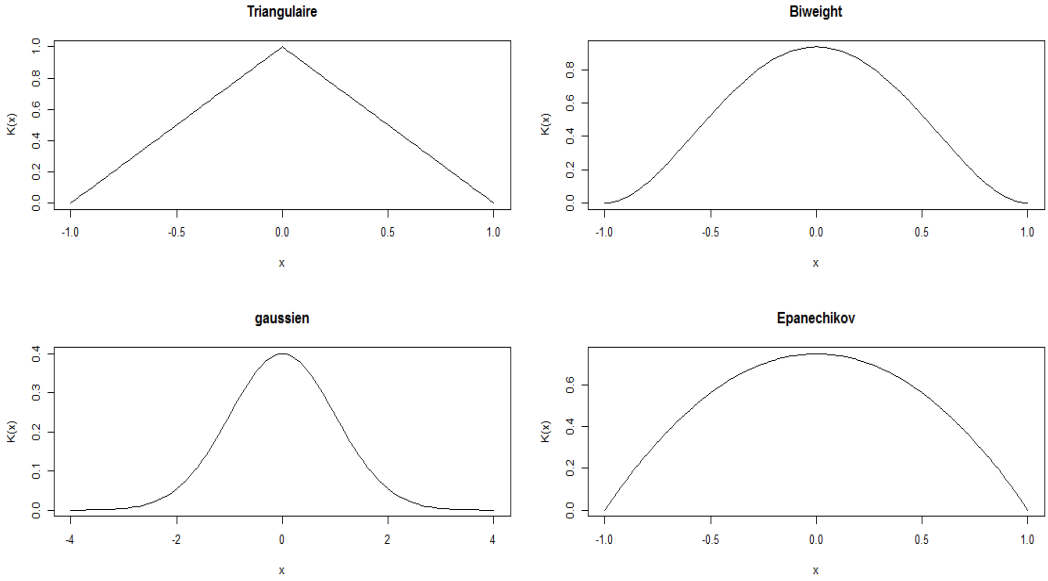

FIG. 2.1 – Allures des noyaux : Triangulaire, Biweight, Gaussien et Epanechnikov.

Chapitre 3

Simulation

Nous terminons notre mémoire par ce troisième chapitre, dont nous essyons par des simulations en utilisant le logiciel **R**, de donner des explications graphiques de différentes notions rencontrer au deuxième chapitre. Nous donnons des exemples qui expriment l'importance de paramètre de lissage h, du noyau et dans l'estimation non paramétrique de la fonction de densité.

L'estimateur non paramétrique de densité est définie par la forme suivante :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{3.1}$$

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 K (noyau gaussien à support non compact, noyau d'Epanechnikov, et noyau uniforme aux supports compacts).

3.1 Présentation des données

on supposons que : X est de loi $\mathcal{N}(0;1)$.

Nous allons donc étudier les cas suivants :

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

Dans les résultats graphique de cette section, on a :

1) Le courbe en noire exprime la fonction de densité f

2) La courbe en rouge exprime l'estimateur à noyau de densité $\hat{f}_n$.

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

3.2 Paramétre de lissage h fixé et n variée

dans cette partie, on choisit le paramètre de lissage $h = n^{-1/5}$ (fixé) et n variée ($n = 40, 200, 600$) , et chaque fois on change le noyau.

3.2.1 K noyau Gaussien (support non compact)

Dans ce premier cas, le paramètre de lissage ou la fenêtre h est fixé $\left(h = n^{-\frac{1}{5}}\right)$ et n varié $(n = 40, 200, 600)$ et K est un noyau normal $K(t) = \exp\left(-\frac{t^2}{2}\right)/\sqrt{2\pi}$, c'est une densité à support non compact.

Code R :

```
n=40
X=rnorm(n)
# Noyau Normal K(t)
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
#La fenetre h
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)}
# Graphes
op=par(mfrow=c(1,3))
plot(x,fn,xlab="x", ylab="fn(x)", main="n=40",type='l',col=10, lwd= 2)
lines(x,dnorm(x),lwd= 2)
####Pour n =200
```

```
n=200
X=rnorm(n)
#La fenetree h
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)}
plot(x,fn,xlab="x", ylab="fn(x)", main="n=200",type='l',col=10, lwd= 2)
lines(x,dnorm(x),lwd= 2)
####Pour n =600
n=600
X=rnorm(n)
#La fenetre h
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)}
plot(x,fn,xlab="x", ylab="fn(x)", main="n=600",type='l',col=10, lwd= 2)
lines(x,dnorm(x),lwd= 2)
par(op)
```

Nous obtenons la figure (3.1) suivante :

Remarque 3.2.1 *On remarque graphiquement que la courbe rouge de f_n est approche beaucoup à la courbe noire de f dans le troisième graphe, donc ce graphe exprime la convergence de l'estimateur f_n vers f.*

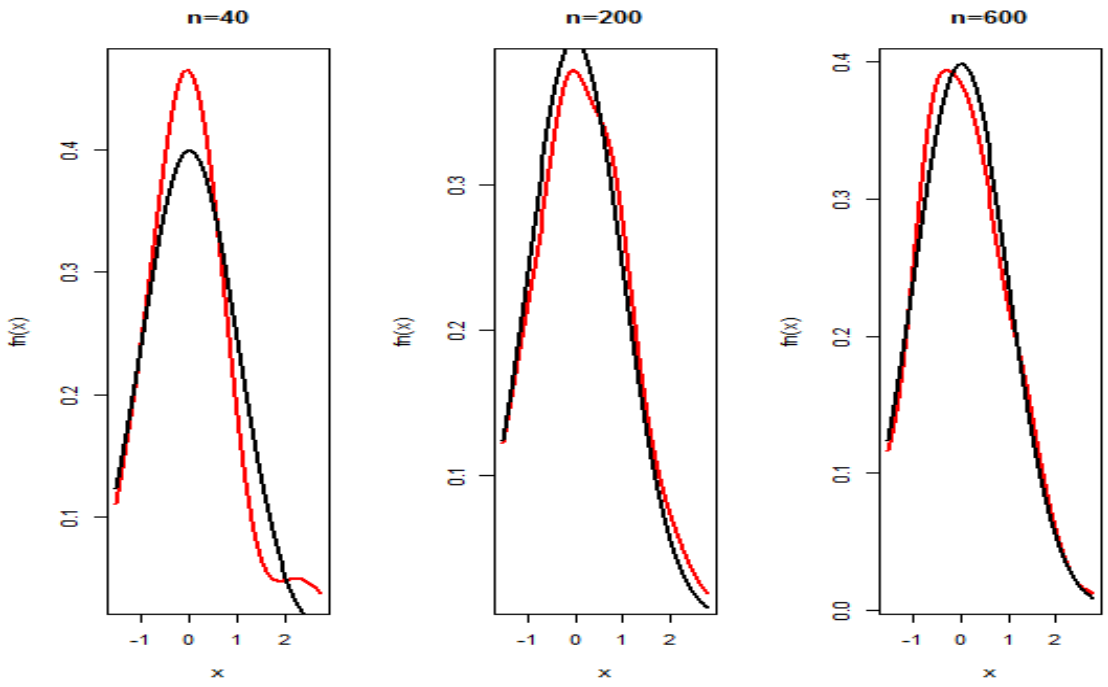

FIG. 3.1 – Estimateur à noyau de la densité : h fixé, n variée et K noyau gaussien

3.2.2 K noyau uniforme (support compact)

Dans ce cas, le paramètre de lissage et toujours fixé $\left(h = n^{-\frac{1}{5}}\right)$ et n variée $(n = 40, 200, 600)$ mais le noyau K est un noyau uniforme $K(t) = \frac{1}{2} I_{(|t| \leq 1)}$ c'est une densité à support compact. On modifie seulement cette partie dans le programme **R** précident :

Noyau Uniforme $K(t)$

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

Nous obtenons la figure (3.2) suivante :

Remarque 3.2.2 *Même conclusion que le graphe (3.1) (i.e., convergence de l'estimateur pour n assez grand) .*

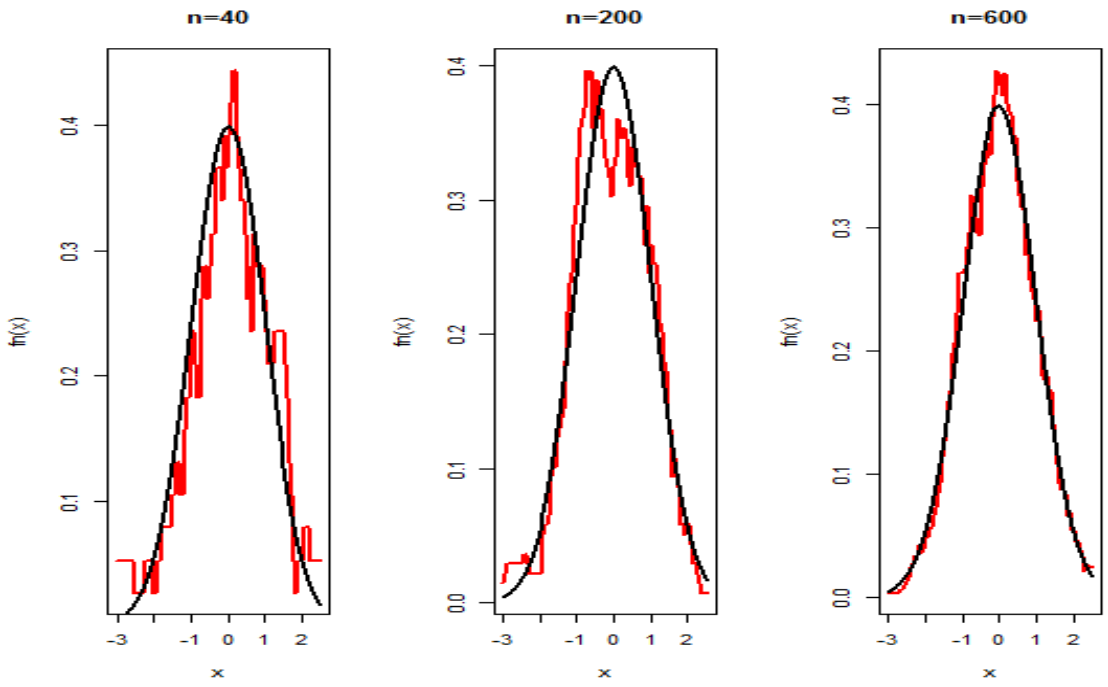

FIG. 3.2 – Estimateur à noyau de la densité : h fixé, n variée et K noyau uniforme.

3.2.3 K noyau d'Epanechnikov

Dans ce cas, le paramètre de lissage et toujours fixé $\left(h = n^{-\frac{1}{5}}\right)$ et n varié $(n = 40, 200, 600)$ mais le noyau K est un noyau d'Epanechnikov $K(t) = \frac{3}{4}(1 - t^2) I_{(|t|\leq 1)}$ c'est une densité à support compact. On modifie seulement cette partie dans le programme **R** précident :

Noyau Epanechenikov $K(t)$

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

Nous obtenons la figure (3.3) suivante :

Remarque 3.2.3 *Même conclusion que les figures (Fig3.1 et Fig 3.2) (i.e., convergence de l'estimateur pour n assez grand) .*

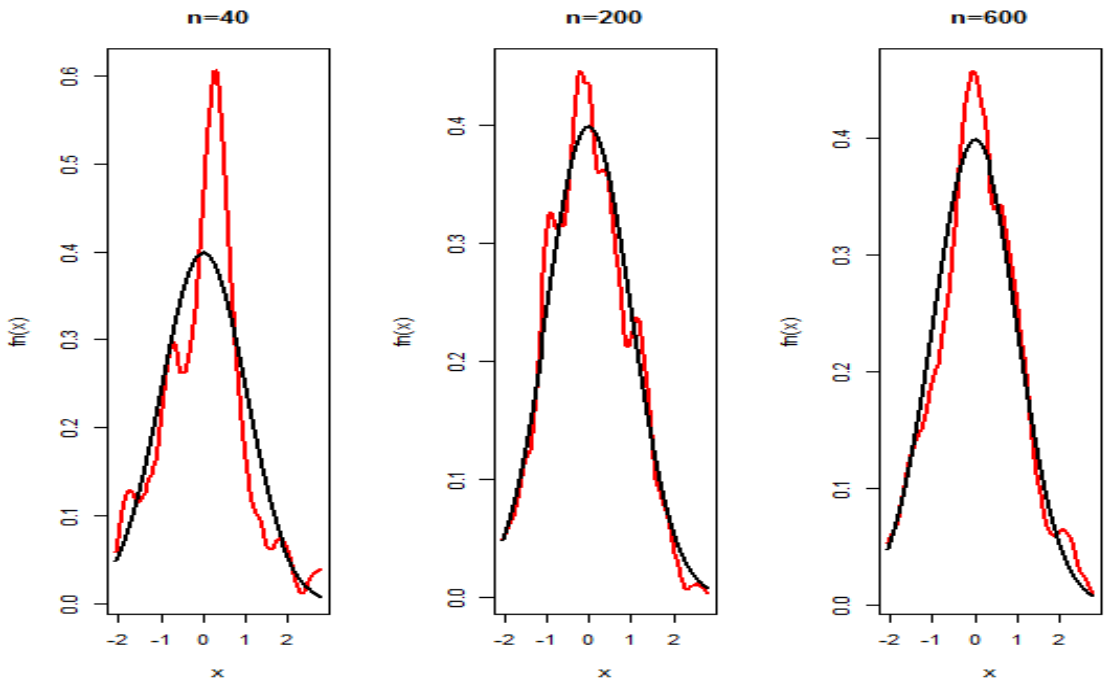

FIG. 3.3 – Estimateur à noyau de la densité :h fixé, n variée et K noyau d'Epanechnikov.

3.3 Choix grafique 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).

3.3.1 K noyau Normal

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 (3.4).

Code R :

```
n=350
X=rnorm(n) # Echantillon X
# Noyau Normal K(t)
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
```

```
#La fenetre h
h=seq(.1,.9,length=9)
# Initiation
s=100 # taille de l'intervalle [a,b]
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=array(dim=c(n,s,9))
fn=array(dim=c(s,9))
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])}}
# Graphes
x11() # nouvelle fenettre graphique
op=par(mfrow=c(3,3))
for(k in 1 :9){
plot(x,fn[,k],xlab="x", ylab="fn(x)", main=" ",type='l',col=4, lwd= 2)
lines(x,dnorm(x),lwd= 2)
}
par(op)
```

Remarque 3.3.1 *Il est clair que la valeur de h optimal est de $h = 0.4$ (ligne 2, colonne 1).*

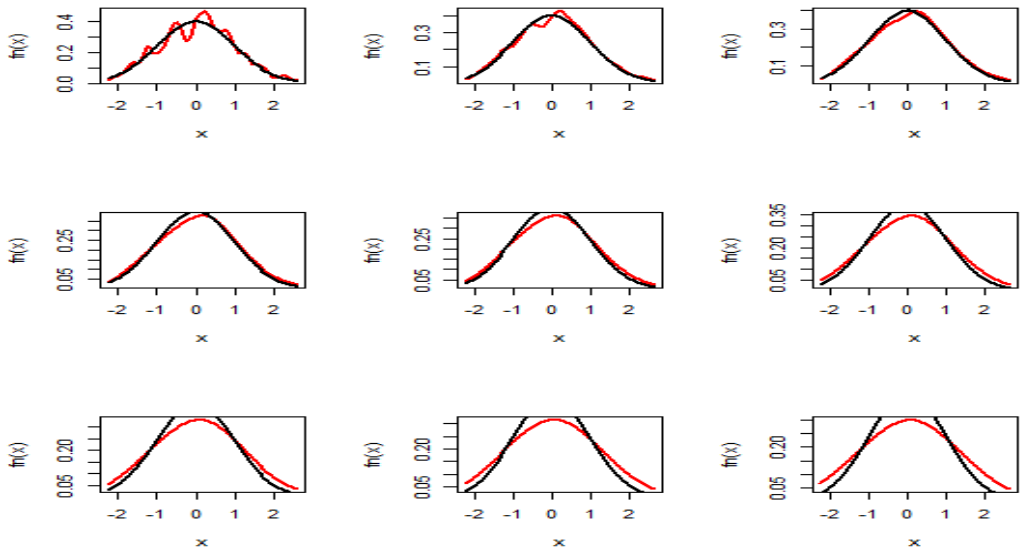

FIG. 3.4 – Estimateur à noyau de la densité : h varié, n fixée et K noyau Normal

3.3.2 K noyau Uniforme

On fixe la taille de l'échantillon $n = 350$ et le noyau K est Uniforme, l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure (3.5).

On modifie seulement cette partie dans le programme **R** précident :

Noyau Uniforme $K(t)$

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

Nous obtenons la figure (3.5) suivante :

Remarque 3.3.2 *Il est clair que la valeur du h optimal est de* $h = 0.6$ *(ligne 2, colonne 3).*

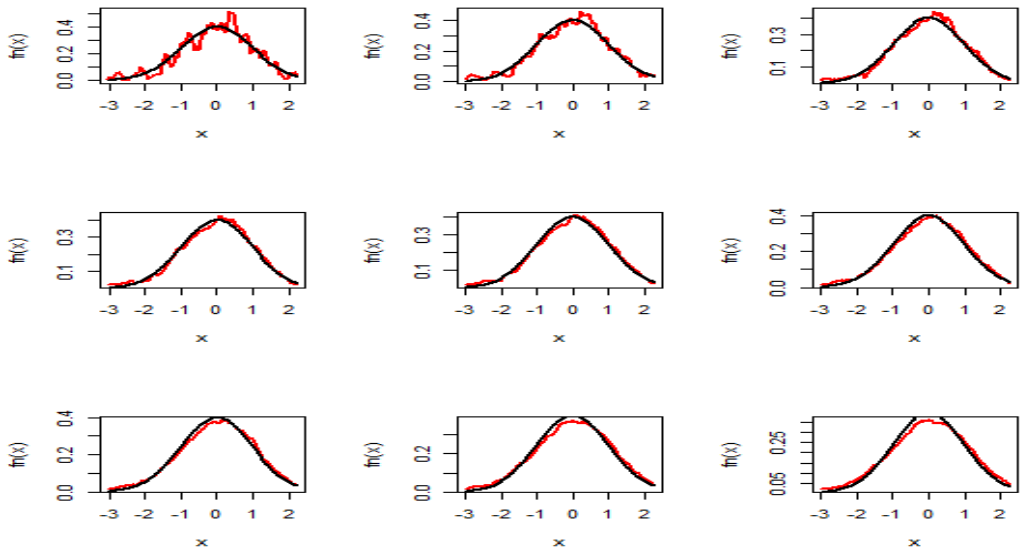

FIG. 3.5 – Estimateur à noyau de la densité : h varié, n fixée et K noyau Uniforme.

3.3.3 K noyau d'Epanechnikov

On fixe la taille de l'échantillon $n = 350$ et le noyau K est d'Epanechenikov, l'estimation obtenue avec les valeurs de h varié de 0.1 à 0.9 sont données dans la figure (3.6).

On modifie seulement cette partie dans le programme **R** précident :

Noyau Epanechnikov $K(t)$

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

Nous obtenons la figure (3.6) suivante :

Remarque 3.3.3 *Il est clair que la valeur du h optimal est de* $h = 0.7$ *(ligne 3, colonne 1).*

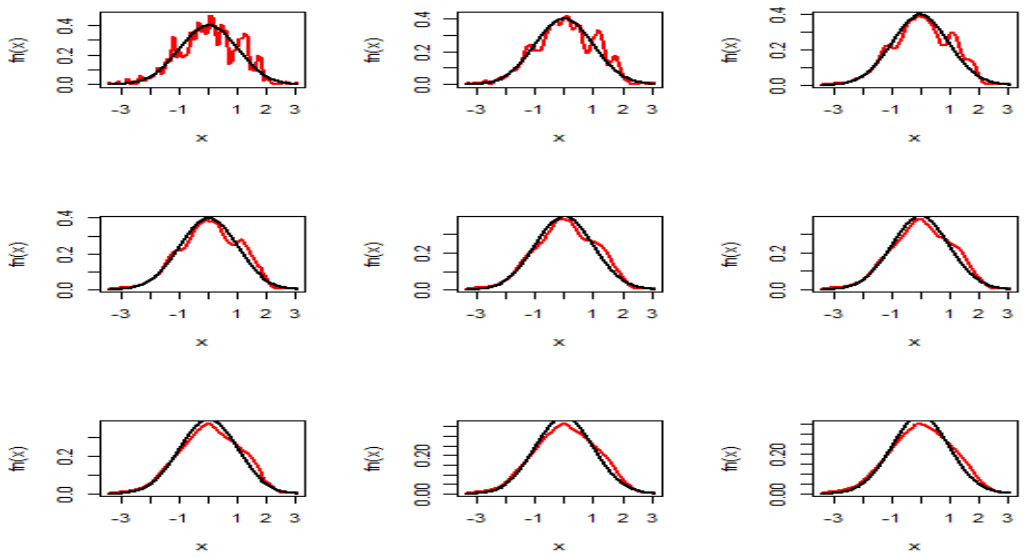

FIG. 3.6 – Estimateur à noyau de la densité : h varié, n fixée et K noyau d'Epanechnikov.

Exemple 3.3.1 *Un exemple de l'impact de la fenêtre sur l'estimateur peut être vu dans la figure (Fig.3.7).*

Code R

```
n=300;m=100
X=rnorm(n)
K=function(t){dnorm(t)}
f1=rep(0,m);f2=rep(0,m);f3=rep(0,m)
x=seq(min(X),max(X),length=m)
h=0.1;for(i in 1:m){f1[i]=sum(K((x[i]-X[1:n])/h))/(n*h)}
h=0.5;for(i in 1:m){f2[i]=sum(K((x[i]-X[1:n])/h))/(n*h)}
h=1;for(i in 1:m){f3[i]=sum(K((x[i]-X[1:n])/h))/(n*h)}
plot(x,dnorm(x),type='s',ylab="f(x)",ylim=c(0,.5))
lines(x,f2,col=2,lty=2,lwd=2)
lines(x,f1,col=3,lty=3,lwd=2)
lines(x,f3,col=4,lty=4,lwd=2)
```

```
legend(-3,.4,c("h=0.0.1","h=0.5","h=1"),col=c(3,2,4),
lty=c(3,2,4), lwd=c(2,2,2), bty = "n", cex=1)
```

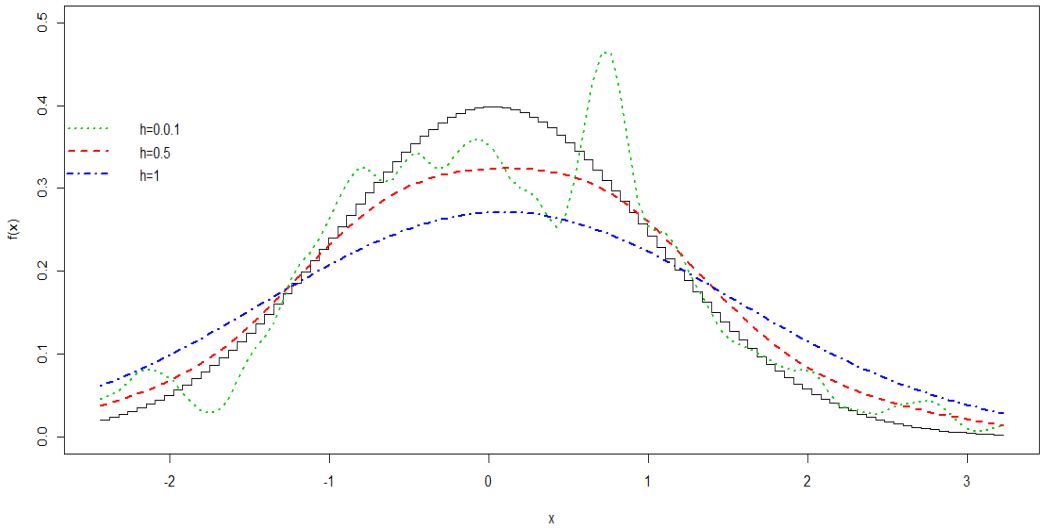

FIG. 3.7 – Estimateur à noyau d'une densité normale, utilisant trois fenêtres différentes.

Exemple 3.3.2 *La figure (Fig.**3.8**) illustre les courbes des estimateurs à noyau de la densité d'une variable normale* $N(0,1)$, *en rouge le noyau triangulaire, en vert le noyau Biweight, en bleu noyau gaussien et en jaune noyau d'Epanechnikov. Comparés à la densité théorique en noire (même valeur de* h $(h_{opt} = cn^{-1/5}, n = 300)$ *est utilisée), les quatre courbes sont identiques ou presque.*

Code R :

```
n=300; m=100
X=rnorm(n); h=1.6*sd(X)*n^-.2
K1=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
K2=function(t){(15/16)*((1-t^2)^2)*ifelse(abs(t)<=1,1,0)}
K3=function(t){dnorm(t)}
```

```
K4=function(t){ifelse(abs(t)<1,(3/4)*(1-t^2),0)
}
a=min(X); b=max(X); x=seq(a,b,length=s)
f1=rep(0,m);f2=rep(0,m);f3=rep(0,m);f4=rep(0,m)
for(i in 1:m){
f1[i]=sum(K1((x[i]-X[1:n])/h))/(n*h)
f2[i]=sum(K2((x[i]-X[1:n])/h))/(n*h)
f3[i]=sum(K3((x[i]-X[1:n])/h))/(n*h)
f4[i]=sum(K4((x[i]-X[1:n])/h))/(n*h)
}
plot(x,dnorm(x),type='s',ylab="f(x)")
lines(x,f1,col=2,lty=2,lwd=2)
lines(x,f2,col=3,lty=3,lwd=2)
lines(x,f3,col=4,lty=4,lwd=2)
lines(x,f4,col=7,lty=5,lwd=2)
legend(-3,.4,c("Triangulaire","Biweight","gaussien","Epanechikov"),
col=c(2,3,4,7), lty=c(2,3,4,5), lwd=c(2,2,2,2), bty = "n", cex=1)
```

Remarque 3.3.4 *On remarque que, le noyau optimal est le noyau d'Epanechnikov et la fenêtre optimale est* $h_{opt} = cn^{-1/5}$.

Conclusion 1 *En conclusion, ce chapitre montre l'importance de paramètre de lissage* h *et du noyau* K *dans l'estimation non paramétrique de la fonction de densité. Mais à noté que le choix de* h *est plus important que celui de noyau.*

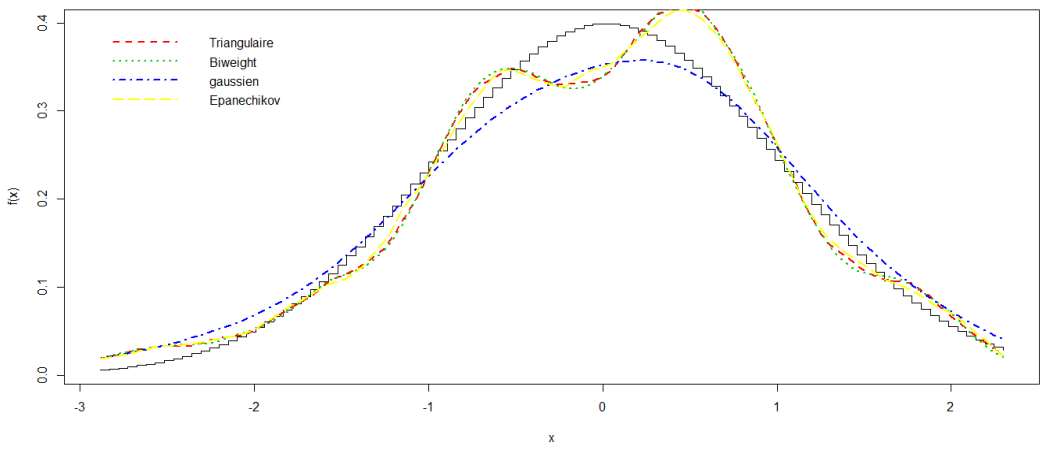

FIG. 3.8 – Estimateurs avec des noyaux : Triangulaire, Biweight, Gaussien et Epanechnikov d'une densité normale.

Conclusion

L'estimation de la fonction de densité joue un rôle central dans l'estimation fonctionnelle. Dans ce mémoire nous avons étudié l'estimation de la densité de probabilité dans le cadre de données complètes.

On a vu dans ce mémoire l'estimateur la fonction de densité : l'estimation paramétrique et l'estimation non-paramétrique, dans cette dernière on parle sur la méthode du noyau qui est proposée par Rosenblatt en 1956 puis améliorée par Parzen en 1962 , cette méthode de noyau est la plus utilisée et basée sur une fonction K appelée noyau et une fenêtre h (paramètre de lissage) qui joue un rôle important dans la qualité de l'estimation. On a vu que l'estimateur à noyau de la densité dépend de la taille de l'échantillon n mais aussi de la fenêtre h et du noyau K.

En conclusion, nous disons que le choix du noyau n'a pas d'influence majeure sur la qualité de l'estimateur, par contre le choix de la fenêtre h est crucial, ceci est illustré par les résultats de simulations.

Bibliographie

[1] Achour, S., 2015 : Le choix du noyau et de la fenêtre dans l'estimation de la densité. Université de Biskra.

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

[3] Cours STAT 2150 "Statistique non paramétrique : Méthodes de ... perso.uclouvain.be › STAT2150 › STAT2150_Transp [pdf].

[4] Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. Revue de Statistique Appliquée, 25(3), 5-42.

[5] Devroye, L. (1985). Nonparametric density estimation. The L_1 View.

[6] Dobele-Kpoka, F. G. B. L. (2013). Méthode non-paramétrique des noyaux associés mixtes et applications (Doctoral dissertation, Université de Franche-Comté ; Université Ouaga 1 Professeur Joseph Ki-Zerbo (Ouagadougou, Burkina Faso)).

[7] density. Theory Probab. Appl. 14, 153-158.

[8] Epanechnikov, V. A., 1969. Nonparametric Estimation of a Multivariate Probability Density. Theory of Probability and its Applications.

[9] Estimation d'une fonction de densité par la méthode des noyaux ...www.ummto.dz › dspace › bitstream › handle › ummt [pdf].

[10] Khalifa, I. B. (2008). Estimation non-paramétrique par noyaux associés et données de panel en marketing. Projet de Fin d'Etude. Université du, 7.

[11] Kokonendji, C., & Kiessé, T. S. (2006). Estimateur à noyau discret standard pour une densité de probabilité discrète.

[12] Le Jan, Y., & Lemaire, S. (2008). Probabilites et statistiques. Cours de licence de Mathématiques fondamentales, Université de Paris-Sud, centre d'Orsay, version.

[13] Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. Theory of Probability & Its Applications, 10(1), 186-190.

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

[15] Rosenblatt, M. (1956). Estimation of a probability density-function and mode. Ann Math Statist, 27, 832-837.

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

[17] Rivoirard, V., & Stoltz, G. (2009). Statistique en action (p. 320). Vuibert.

[18] Rivoirard, V., & Stoltz, G. (2012). Statistique mathématique en action.

[19] STAT., 2413 ; 2002 2003 : Chapitre 3 Estimation non paramétrique d'une fonction de répartition et d'une densité.

[20] Sheather, S. J. (2004). Density estimation. Statistical science, 588-597.

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

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

[23] Silverman, B. W. (2018). Density estimation for statistics and data analysis. Routledge.

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

Annexe : Logiciel R

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

 est un système qui est communément appelé logiciel : R Development Core Team (2010). : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http ://www.R-project.org/.

 permet de réaliser des analyses des statistiques. Plus particulièrement, il comporte des moyennes qui rendent possibles la manipulation des données, les calculs et les représentations graphique, R à aussi la possibilité d'exécuter des programmes stokes dans des fichiers textes. En effet R possède :

1. déférentes opérateurs pour calcul sur tableaux, en particulier les matrices,
2. un grand nombre d'outils pour l'analyse des données et les méthodes statistique,
3. des moyennes graphiques pour visualiser les analyses,
4. un langage de programmation simple et performât comportant,
5. Conditions, boucles, moyennes d'entrées sorties...