الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre Universitaire Salhi Ahmed- Naama
Institut des sciences et technologies
Département de Mathématiques et Informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master
En : Mathématiques

Spécialité : Probabilités, Statistique et Application

Intitulé

L'estimation à noyau d'une densité de probabilité et applications

Présenté par :
BOUGUERNE Zineb

Soutenu : Juillet 2022

Devant le jury composé de :

Dr.BARI Amina	MCB	C-Univ Naâma	Présédent
M.KHELOUATI Hafidha	MAA	C-Univ Naâma	Examinatrice
M .MOUSSAOUI Fatma	MAA	C-Univ Naâma	Encadreur

Année universitaire 2021/2022

Dédicace

A celle qui m'a transmis la vie, l'amour , le courage , à toi chèr maman toutes mes joies,
mon amour et ma reconnaissance.
A mon père pour l'éducation qu'il m'a prodigué , avec tous les moyens et au prix de toutes
les sacrifices qu'il a consentis á mon égard et à mes études depuis mon enfance.
A mon chère ABDALRAHMAN , le meilleur de ce qu'une personne acquiert, ce sont des
frères car il aident dans la vie.
A toute ma famille et tous mes proches.
A ma chère enseignante M FATMA qui a tout le respect et la reconnaissance , à la bonne
âme qui nous a toujours quittés, le père du professeurs, nous demandons à Dieu tout-
Puissant de le couvrir de sa large miséricorde et d'habiter son vaste paradis.

Je dédie ce travail.

BOUGUERNE ZINEB

Remerciements

Je tiens à remercier en premier lieu et avant tout ALLAH le tout puissant, qui nous a donné la force et la patience d'accomplir notre travail dans les meilleures conditions.

Je tiens également à remercier du coeur mon encadreur Melle.MOUSSAOUI Fatma, son aide précieuse et ses conseilles judicieuses.

Je voudrai également remercier les membres de jury Dr.BARI Amina et Madame.KHELOUATI Hafidha maîtres de conférences au centre universitaire Salhi Ahmed de Naâma, qui m'ont fait l'honneur de juger ce modeste travail.

Table des matières

Notations

$X_1, ..., X_n$:	échantillant de taille n.
F :	fonction de répartition.
f :	densité de probabilité.
$v.a$:	variable aléatoire.
$i.i.d$:	indépendantes et identiquement distribuées.
Θ :	espace des paramètres.
F_n :	fonction de répartition empirique.
f_n :	l'estimateur à noyau de la densité de probabilité f.
MSE :	erreur quadratique moyenne.
MISE :	erreur quadratique moyenne intégré.
$\xrightarrow{P}$:	convergence en probabilité.
$\xrightarrow{\mathcal{L}}$:	convergence loi.
$\xrightarrow{p.s}$:	convergence presque sûre.
eff :	effecacité.
$p.co$:	convergence presque complète.

Table des figures

Liste des tableaux

Introduction

L'estimation statistique est un domaine trés important de la statistique mathématique est divisée en deux volets principaux, savoir l'estimation paramétrique et l'estimation non-paramétrique. L'objectif de l'estimation paramétrique est d'estimer les paramèties d'une distibution connue, différentes méthodes existent pour l'estimation notamment : la méthode de maximum de vraisemblance et la méthode des moment. Par opposition, l'estimation non-paramétique est estime la densité directement à partire de l'information disponible sur l'ensemble d'observations. Plus particulièrement, on parle de méthode d'estimation non-paramétique l'orsque celle-ci ne se ramène pas l'estimation d'un nombre fini de paramètres réels associés la loi de l'échantillon. Il existe plusieurs méthodes non-paramétrique pour l'estimation on peut citer la méthode de l'histogramme, la méthode de séries orthogonales, la méthode splines et la méthode du noyan.

L'estimation de la densité par la méthode du noyan se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support. En ce sens, cette méthode généralise la méthode d'estimation par un histogramme, elle est la plus populaire parmi les autres méthodes d'estimation non -paramétiques de la densité. Cette popularité d'estimateur à noyan peut s'expliquer par au moins trois raisons : la simplicité de sa forme, ses modes de convergence multiples et sa flexibilité qui s'interprète pas la liberté de l'utilisateur dans le choix du noyau K.

La méthode d'estimation non-paramitique du noyau fut introduite par Rosenblatt en (1956) [14] pour estimer des densités de probabilité, puis amélioré par Parzen en(1962) [12] pour estimer le mode d'une densité de probabilité et par Nadaraya (1964) [11] et Watson (1964) [20] pour estimer une fonction de regression à support non borné.

Tout au long de ce travail, nous intéressons l'estimation par la méthode du noyau partir d'un échantillon de variables aléatoires indépendantes et identiquement distibuées, et il composé d'une introduction, de trois chapitres et d'une conclusion.

Dans le chpitre un, nous regroupons quelques définitions ainsi que quelques outiles qui seront nécessaires pour l'élaboration des résultats établis dans ce mémoire.

Dans la deuxième chapitre, nous concentrons sur la méthode de noyau pour estimer la densité, et on rappelle ces propriétés fondamentales et aborder le problème de choix

optimal de noyau et de paramètre de lissage.

La dernier chapitre, nous donnons des exemples de simulation par le logiciel R qui expriment l'importance de paramètre de lissage et le noyau.

Chapitre 1

Généralités et rappeles

1.1 Notations et définitions

Définition 1.1.1 ***(Fonction de répartition)*** *Soit X une $v.a$, on appelle fonction de répatition de X, que l'on note F_X, la fonction définie sur $\mathbb{R}$ par :*

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}(X \leq x) = \mathbb{P}_X(]-\infty, x]).$$

Propriété 1.1.1 *Soit F est une fonction de répartition de X, alors :*

1) $0 \leq F(x) \leq 1, \forall x \in \mathbb{R}$.

2) Si $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$.

3) $\lim\limits_{x \to -\infty} F(x) = 0$, $\lim\limits_{x \to +\infty} F(x) = 1$.

4) $\forall a, b \in \mathbb{R}, \mathbb{P}(a < X \leq b) = F(b) - F(a)$.

5) $\forall a \in \mathbb{R}, \mathbb{P}(X > a) = 1 - \mathbb{P}(X \leq a) = 1 - F(a)$.

6) Si X est une $v.a$ discrète alors :$\forall x \in \mathbb{R}$

$$F(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} \mathbb{P}(x_i).$$

7) Si X est une $v.a$ continue de fonction de densité $f(x)$, alors :

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}(X \leq x) = \textstyle\int_{-\infty}^{x} f(t)dt \text{ et } f(x) = \frac{\partial}{\partial x} F(x).$$

Définition 1.1.2 ***(La densité de probabilité)*** *Une fonction $f : \mathbb{R} \to \mathbb{R}$ est appelée densité de probabilité si :*

- *Elle est positive en tout* $x \in \mathbb{R}$ $(f(x) \geq 0)$.
- *Elle est intégrable.*
- $\displaystyle\int_{-\infty}^{+\infty} f(x)dx = 1.$

Définition 1.1.3 ***(Espéravce mathématique)*** *Soit X v.a de densité de probabilité f, alors l'espérance mathématique de X est donnée par*

$$\mathbb{E}(X) = \int x f(x) dx.$$

Définition 1.1.4 ***(Variance mathématique)*** *La variance d'une $v.a X$ est définie par*

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

1.2 Estimation paramétrique

Définition 1.2.1 *Soi $(\Omega, \mathcal{A})$ un espace probabilisable et soit*

$$\mathcal{P} = \{P_\theta, \theta \in \Theta, \Theta \subset \mathbb{R}^k\}$$

une famille des lois de probabilité. Le triplet $(\Omega, \mathcal{A}, \mathcal{P})$ est appelé structure statistique.

Exemple 1 *Le structure gaussienne est $(\mathbb{R}, \mathcal{B}_\mathbb{R}, \mathcal{F}_{m,\sigma})$ avec*

$$\mathcal{F}_{m,\sigma}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-m)^2}{2\sigma^2}}$$

ici $\mathcal{P}$ est la famille des lois définies par la densité $\mathcal{F}_{m,\sigma}, \theta = (m, \sigma), \Theta = \mathbb{R} \times \mathbb{R}_+^*$ et $k = 2$.

Définition 1.2.2 *Soit X une variable aléatoire $(v.a)$ sur $(\Omega, \mathcal{A}, \mathbb{P})$. Une suite de variable aléatoire $X_1, ..., X_n$ est appelé échantillon sur X s'elle sont deux à deux indépandantes et elles ont la même loi probabilité que X.*

Définition 1.2.3 *Soit $X_1, ..., X_n$ un n-échantillon de X de loi P_θ, toute fonction mesurable T de $X_1, ..., X_n$ est appelée statistique.*

1.2.1 Estimateur statistique

Définition 1.2.4 *Soit $X_1, ..., X_n$ un n-échantillon de X de loi P_θ. Un estimateur de θ est une statistique $T(X_1, ..., X_n)$ qui prend ces valeurs dans Θ.*

Exemple 2 *Les paramètres d'une loi normale μ, σ^2 peut être estimés par $\bar{x}$ et s^2 respect.*

Dans tout la suite on prendre $\hat{\theta}_n = T(X_1, ..., X_n)$ estimateur de θ.

1.2.2 Qualité d'un estimateur

Définition 1.2.5 *Le biais d'un estimateur $\hat{\theta}_n$ de θ est donnè par :*

$$biais(\hat{\theta}_n) = b(\hat{\theta_n}) = \mathbb{E}(\hat{\theta}_n - \theta)$$

.

• Estimateur avec biais

Un estimateur $\hat{\theta}_n$ de θ est dit biais si : pour tout entier positif n

$$\mathbb{E}(\hat{\theta}_n) = \theta + b(\hat{\theta}_n)$$

avec $b(\hat{\theta}_n)$ est le biais de l'estimateur $\hat{\theta}_n$.

• Estimateur sans biais

Un estimateur $\hat{\theta}_n$ de θ est dit sans biais si : $\mathbb{E}(\hat{\theta}_n) = \theta$.

• Estimateur asymptotiquement sans biais

Un estimateur $\hat{\theta}_n$ de θ est dit asymptotiquement sans biais si : $\lim_{n\to+\infty} \mathbb{E}(\hat{\theta}_n) = \theta$.

1.3 Estimation non paramétrique

L'estimation paramétrique est que la loi de probabilité inconnue.

1.3.1 Paramètre fonctionnel

Définition 1.3.1 *On dit que θ est un paramètre fonctionnel si : $\theta = T(F)$, òu :*

$$\begin{aligned} T: \quad & \mathcal{F} \to \Theta \\ & F \mapsto T(F) = \theta \end{aligned}$$

$\mathcal{F}$: l'ensemble des fonctions de répertition.

Θ : l'ensemble des paramètres.

T : s'appelle fonctionnelle statistique.

Exemple 3

1. *La densité d'une $v.a$ est un paramètre fonctionnel tel que $\theta = f = T(F) = d(F)$, dans ce cas T est l'opérateur dérivée d.*

2. $\mathbb{E}(X) = \int x \, dF$ *est paramètre foncionnel tel que* $T(F) = \int x \, dF$ *est la fonctionnel statistique.*

1.3.2 Criteurs d'erreurs

a) L'erreur moyenne quadratique

$$
\begin{aligned}
MSE\left(\hat{T}_n(x)\right) &= \mathbb{E}\left[\left(\hat{T}_n(x) - T(x)\right)^2\right] \\
&= \mathbb{E}\left[\hat{T}_n^2(x) + T^2(x) - 2\hat{T}_n(x)T(x)\right] \\
&= \mathbb{E}\left(\hat{T}_n^2(x)\right) + \mathbb{E}\left(T^2(x)\right) - 2\mathbb{E}\left(\hat{T}_n(x)T(x)\right) \\
&= \mathbb{E}\left(\hat{T}_n^2(x)\right) + \mathbb{E}\left(T^2(x)\right) - 2T(x)\mathbb{E}\left(\hat{T}_n(x)\right) \\
&= \mathbb{E}\left(\hat{T}_n^2(x)\right) + T^2(x) - 2T(x)\mathbb{E}\left(\hat{T}_n(x)\right) \\
&= \mathbb{E}\left(\hat{T}_n^2(x)\right) + T^2(x) - 2T(x)\mathbb{E}\left(\hat{T}_n(x)\right) + \mathbb{E}\left(\hat{T}_n(x)\right)^2 - \mathbb{E}\left(\hat{T}_n(x)\right)^2 \\
&= \left[\mathbb{E}\left(\hat{T}_n^2(x)\right) - \mathbb{E}\left(\hat{T}_n(x)\right)^2\right] \\
&+ \left[\mathbb{E}\left(\hat{T}_n(x)\right)^2 - 2T(x)\mathbb{E}\left(\hat{T}_n(x)\right) + T^2(x)\right] \\
&= Var\left(\hat{T}_n(x)\right) + \left[T(x) - \mathbb{E}\left(\hat{T}_n(x)\right)\right]^2 \\
&= Var\left(\hat{T}_n(x)\right) + biais^2\left(\hat{T}_n(x)\right)
\end{aligned}
$$

b) L'erreur moyenne quadratique intégrée

$$\begin{aligned} MISE\left(\hat{T}_n(x)\right) &= \int_{\mathbb{R}} MSE\left(\hat{T}_n(x)\right) dx \\ &= \int_{\mathbb{R}} \mathbb{E}\left[\left(\hat{T}_n(x) - T^2(x)\right)^2\right] dx \\ &= \int_{\mathbb{R}} \left[Var\left(\hat{T}_n(x)\right) + biais^2\left(\hat{T}_n(x)\right)\right] dx \\ &= \int_{\mathbb{R}} Var\left(\hat{T}_n(x)\right) dx + \int_{\mathbb{R}} biais^2\left(\hat{T}_n(x)\right) dx \end{aligned}$$

1.4 Quelques types de convegence

Dans cette section, nous allons introduire différentes notions de convergence pour une suite de variables aléatoires.

Définition 1.4.1 *(Convergence en probabilité) On dit que la suite* $(X_n,\ n \in \mathbb{N})$ *de v.a.r. converge en probabilité vers la variable aléatoire* X*, si*

$$\forall \epsilon > 0 \quad P(|X_n - X| > \epsilon) = 0, \qquad \textit{quand } n \textit{ tend vers } \infty.$$

On note $X_n \xrightarrow{p} X$

Définition 1.4.2 *(Convergence dans* L^p*) Soit* (X_n) *une suite de v.a.r. dans* L^p*. On dit qu'elle converge dans* L^p *vers une v.a.r.* X *si*

$$\|X_n - X\|_p \underset{n\to+\infty}{\longrightarrow} 0$$

.

Définition 1.4.3 *(Convergence presque sûrement) On dit que la suite* (X_n) *de v.a.r. converge presque sûrement vers* X *s'il existe un élément* A *de la tribu* A *tel que* $P(A) = 1$ *et pour tout* $\omega \in A$

$$\lim_{n\to+\infty} X_n(\omega) = X(\omega)$$

On note

$$X_n \xrightarrow{p.s} X$$

Définition 1.4.4 *(Convergence presque complète) On dit que la suite* (X_n) *de v.a.r est*

converge presque complète vers X si

$$\forall \epsilon > 0, \quad \sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty$$

et notée par $\lim_{n\to\infty} X_n = X, \quad p.co$

Proposition 1.4.1 *Si* $\lim\limits_{n\to\infty} X_n = X, p.co$*, alors nous avons*

1. $\lim\limits_{n\to\infty} X_n = X, p.$
2. $\lim\limits_{n\to\infty} X_n = X, p.s.$

1.5 Taux de convergence

le taux de convergence presque sure à 0 pour une suite des variable aléatoire est définie par :

$$\begin{aligned} X_n - X = O_{p.s}(u_n) &\iff \mathbb{P}((X_n - X) = O_{u_n}) = 1 \\ &\iff \mathbb{P}(\exists c < \infty, \exists n, \forall m > n, |X_n - X| \leq c u_m) = 1. \end{aligned}$$

et le taux de convergence en probabilité définie par :

$$X_n - X = O_p(u_n) \iff \lim_{m} \limsup_{n\to\infty} \mathbb{P}(|X_n - X| \leq m u_n) = 1$$

Maintenant on définit le taux de la convergence presque complète :

Définition 1.5.1 *On dit que le taux de convergence presque complète des $(X_n)_{n\in\mathbb{N}}$ vers X est l'ordre u_n telle que $(u_n)_{n\in\mathbb{N}}$ est une suite déterministe de nombres réels positifs qui tend vers zéro si et seulement si :*

$$\exists \epsilon_0 > 0, \sum_{n\in\mathbb{N}} \mathbb{P}(|X_n - X| > \epsilon_0 u_n) < \infty.$$

et on écrit : $X_n - X = O_{p.co}(u_n)$.

Proposition 1.5.1 *[7] Si* $\lim\limits_{n\to\infty} u_n = 0$*,* $X_n = O_{p.co}(u_n)$ *et* $\lim\limits_{n\to\infty} Y_n = l_y p.co$*, où : l_y est une nombre réel déterministe. Alors*

i) $X_n Y_n = O_{p.co}(u_n)$.

ii) $\frac{X_n}{Y_n} = O_{p.co}(u_n)$*, avec :* $l_y \neq 0$.

Chapitre 2

Estimation non paramétrique de la densité

2.1 La construction d'un estimateur à noyau

Soientt $X_1, ..., X_n$ n variables aléatoires réelles indépendantes et identiquement distribuées de fonction de distribution F et de densité f. L'objectif de notre étude est la construction d'un estimateur de f en un point fxe x. Notons $F(x) = \mathbb{P}(X_1 \leq x)$ la fonction de répartition de la loi de X_1. La densité est la dérivée de la fonction de répartition, ce qui permet d'écrire pour tout x :

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

Considérons la fonction de répartition empirique :

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

La loi forte des grands nombres permet d'affirrmer que F_n est un estimateur de F. Ramplace F par sa estimateur F_n, donc :

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h},$$

nous pouvons le réécrire sous la forme

$$f_n(x) \quad = \frac{1}{2nh} \sum_{i=1}^{n} (\mathbf{1}_{\{X_i \leq x+h\}} - \mathbf{1}_{\{X_i \leq x-h\}}) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbf{1}_{\{x-h < X_i \leq x+h\}}$$

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

avec

$$K_0(t) = \frac{1}{2}\mathbf{1}_{\{-1<t\leq 1\}}, \forall t \in \mathbb{R}$$

Cet estimateur appelé estimateur de Rosenblatt, est le premier exemple d'estimateur à noyau construit à l'aide du noyau $K_0(t) = \frac{1}{2}\mathbf{1}_{\{-1<t\leq 1\}}$.
Définissons maintenant plus généralement la notion d'estimateur à noyau :

Définition 2.1.1 *Un estimateur à noyau de la densité f est une fonction définie par :*

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{2.1}$$

avec :

- *Le réel h que l'on appellera* ***la fenêtre****.*
- *La fonction $K : \mathbb{R} \to \mathbb{R}$, intégrable, que l'on appelle* ***le noyau****.*

tel que h et K vérifions les conditions suivantes :

1. *K est une densité :* $\int_{-\infty}^{+\infty} K(t)dt = 1.$
2. *K est carré intégrable :* $\int_{-\infty}^{+\infty} |K(t)|^2 dt < \infty.$
3. *K est symétrique autour de zéro, c.à.d* $K(t) = K(-t) \Longrightarrow \int_{-\infty}^{+\infty} tK(t)dt = 0$
4. *K possède un moment dordre 2 fini, c.à.d* $\int_{-\infty}^{+\infty} t^2 K(t)dt < \infty.$
5. $\int_{-\infty}^{+\infty} t^2|K(t)|dt < \infty$
6. $\sup_t |K(t)| < +\infty.$
7. $K(.) \in L^1(\mathbb{R}), \qquad c-\grave{a}-dire \int_{-\infty}^{+\infty} K(t)dt < \infty.$
8. $\lim_{|t|\to\infty} K(t) = 0.$
9. $\lim_{n\to\infty} h_n = 0, \qquad \lim_{n\to\infty} nh_n = \infty.$
10. $\lim_{n\to\infty} \frac{nh_n}{\ln n} = \infty.$

Proposition 2.1.1 *Si K est positive et $\int_{-\infty}^{+\infty} K(t)dt = 1$, alors $f_n(x)$ est une densité de probabilité. De plus, $f_n(x)$ est continue si K est continue.*

Preuve 2.1.1 *L'estimateur à noyau est positive et continue car la somme des fonctions positives et continues est elle-mème une fonction positive et continue. Il faut donc vérifier*

que l'intégrale de $f_n(x)$ vaut un. En effet,

$$\int_{-\infty}^{+\infty} f_n(x)dx = \int_{-\infty}^{+\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) dx$$

$$= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K\left(\frac{X_i - x}{h}\right)dx.$$

On pose $t = \frac{X_i - x}{h} \Longrightarrow dx = -hdt$

$$\int_{-\infty}^{+\infty} f_n(x)dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{+\infty}^{-\infty} K(t)(-h)dt,$$

$$= -\int_{+\infty}^{-\infty} K(t)dt = \int_{-\infty}^{+\infty} K(t)dt = 1.$$

2.1.1 Noyaux usuelles

Voici quelques exemples de noyaux les plus utilisés dans le tableu suivant :

$K(t) = \frac{1}{2}$	Si $t \in [-1, 1]$	noyau Uniforme (rectangulaire)
$K(t) = 1 - \lvert t\rvert$	Si $t \in [-1, 1]$	noyau de triangulaire
$K(t) = \frac{3}{4}(1 - t^2)$	Si $t \in [-1, 1]$	noyav d'Epanechnikov
$K(t) = \frac{15}{16}(1 - t^2)^2$	Si $t \in [-1, 1]$	noyau de biweight
$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2)$	Si $t \in \mathbb{R}$	noyau de gaussien
$K(t) = \frac{35}{32}(1 - t^2)^3$	Si $t \in [-1, 1]$	noyau de Triweight

TABLE 2.1 – Quelques exemples des noyaux les plus couramment utilisés

La représentation graphique des quelques noyaux définis ci dessus est donnée par la figure (2.1)

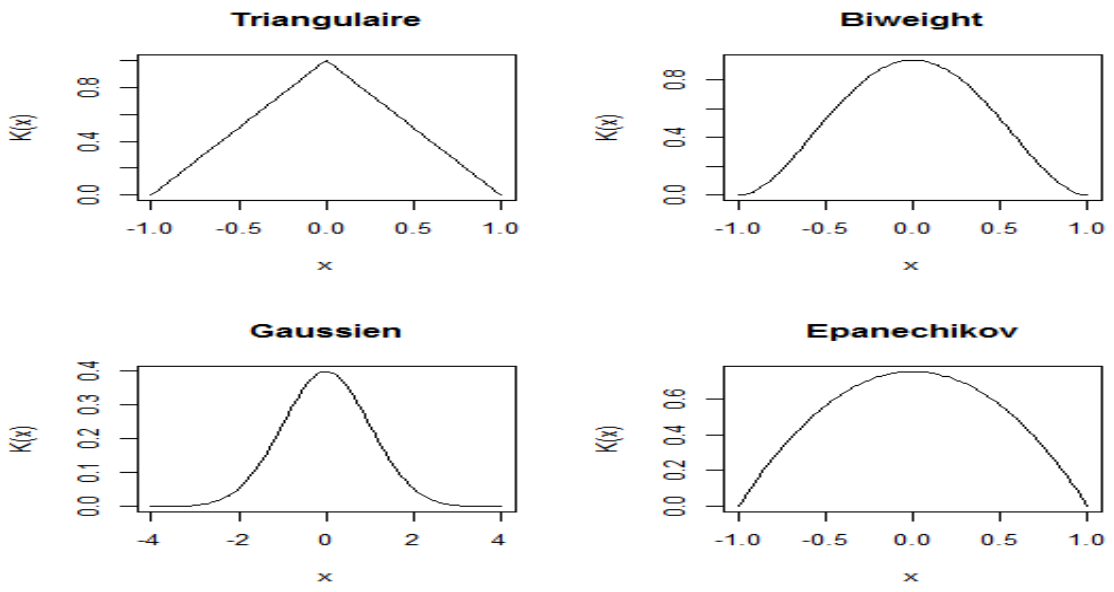

FIGURE 2.1 – Courbes des noyaux : Triangulaire, Biweight, Gaussien, Epanechnikov

2.2 Propriétés d'un estimateur à noyau

Nous allons maintenant donner quelques propriétés statistiques élémentaires de l'estimateur de la densité à noyau f_n défini par l'équation 2.1

2.2.1 Étude du l'espérance

$$E(f_n(x)) = \frac{1}{nh} E\left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right) = \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{y-x}{h}\right) f(y)dy,$$

en posant $t = \frac{y-x}{h} \Longrightarrow dy = hdt$

$$E(f_n(x)) = \int_{-\infty}^{+\infty} K(t) f(x+th)dt,$$

en faisant le développement de Taylor l'ordre 2 de $f(x+th)$ au voisinage de x est alors :

$f(x+th) = f(x) + (th)f^{'}(x) + \frac{(th)^2}{2}f^{''}(x) + o(h^2)$: Il vient

$$\begin{aligned} E(f_n(x)) &= \int_{-\infty}^{+\infty} K(t)f(x+th)dt \\ &= \int_{-\infty}^{+\infty} K(t)\left[f(x) + (th)f^{'}(x) + \frac{(th)^2}{2}f^{''}(x)\right]dt + o(h^2) \\ &= f(x)\int_{-\infty}^{+\infty} K(t)dt + hf^{'}(x)\int_{-\infty}^{+\infty} tK(t)dt + \frac{h^2}{2}\int_{-\infty}^{+\infty} t^2K(t)f^{''}(x)dt \\ &\quad +o(h^2) \end{aligned}$$

d'aprés les conditions $\int_{-\infty}^{+\infty} K(t)dt = 1$, $\int_{-\infty}^{+\infty} tK(t)dt = 0$ et $\int_{-\infty}^{+\infty} t^2K(t)dt < \infty$

$$E(f_n(x)) = f(x) + \frac{h^2}{2}f^{''}(x)\int_{-\infty}^{+\infty} t^2K(t)dt + o(h^2).$$

Alors l'expression finale est donnée par

$$E(f_n(x)) = f(x) + \frac{h^2}{2}f^{''}(x)\mu_2(K) + o(h^2),$$

avec $\mu_2(K) = \int_{-\infty}^{+\infty} t^2K(t)dt$.

2.2.2 Étude du biais

$$\begin{aligned} biais\ (f_n(x)) &= E(f_n(x)) - f(x) \\ &= f(x) + \tfrac{h^2}{2}f^{''}(x)\mu_2(K) + o(h^2) - f(x) \end{aligned}$$

alors

$$biais\ (f_n(x)) = \frac{h^2}{2}f^{''}(x)\mu_2(K) + o(h^2),$$

Le biais est différent de zéro, ceci signifie que l'estimateur à noyau est un estimateur biaisé.

Proposition 2.2.1 *Si la densité f est bornée et f'' existe et bornée. Sous les conditions $\int_{-\infty}^{+\infty} K(t)dt = 1, K(-t) = K(t), \int_{-\infty}^{+\infty} tK(t)dt = 0$ et $\int_{-\infty}^{+\infty} t^2|K(t)|dt < \infty$, alors*

$$|Biais\ (f_n(x))| \leq C_1h^2,$$

où $C_1 = \frac{1}{2}\sup_{z\in\mathbb{R}}\left|f^{''}(z)\right|\int_{\mathbb{R}} t^2|K(t)|dt$.

2.2.3 Étude du variance

$$\begin{aligned} Var(f_n(x)) &= var\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right) = \frac{1}{nh^2} var\left(K\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh^2}\left[E\left(K^2\left(\frac{X_i - x}{h}\right)\right)\right] - \frac{1}{nh^2}\left[E\left(K\left(\frac{X_i - x}{h}\right)\right)\right]^2 \\ &= \frac{1}{nh^2}\int_{-\infty}^{+\infty} K\left(\frac{y-x}{h}\right)^2 f(y)dy - \frac{1}{nh^2}\left(\int_{-\infty}^{+\infty} K\left(\frac{y-x}{h}\right) f(y)dy\right)^2, \end{aligned}$$

en posant $t = \frac{y-x}{h} \Longrightarrow dy = hdt$

$$Var(f_n(x)) = \frac{1}{nh}\int_{-\infty}^{+\infty} K^2(t) f(x+th)dt - \frac{1}{n}\left(\int_{-\infty}^{+\infty} K(t) f(x+th)dt\right)^2.$$

En utilisant le développement de Taylor de f au voisinage de x à lordre 0 alors

$$f(x+th) = f(x) + o(1).$$

Il vient

$$\begin{aligned} Var(f_n(x)) &= \frac{1}{nh}\int_{-\infty}^{+\infty} K^2(t)(f(x)+o(1))dt - \frac{1}{n}\left(\int_{-\infty}^{+\infty} K(t)(f(x)+o(1))dt\right)^2 \\ &= \frac{1}{nh} f(x) R(k) + o(\frac{1}{nh}), \end{aligned}$$

avec $R(K) = \int_{-\infty}^{+\infty} K^2(t)dt$.

Proposition 2.2.2 *Si la densité f est bornée et f'' existe et bornée. Sous les conditions $\int_{-\infty}^{+\infty} K(t)dt = 1, K(-t) = K(t), \int_{-\infty}^{+\infty} tK(t)dt = 0$ et $\int_{-\infty}^{+\infty} t^2|K(t)|dt < \infty$, alors*

$$var\ (f_n(x)) \leq \frac{C_2}{nh},$$

avec $C_2 = \sup_{z\in\mathbb{R}} f(z) \int_{\mathbb{R}} K(t^2)dt$.

Remarque 2.2.1 *Nous remarquons que si $h \to 0$ et $nh_n \to \infty$ quand $n \to \infty$, on a :*

$$\lim_{n\to\infty} \mathbb{E}(f_n(x)) = f(x) \text{ et } \lim_{n\to\infty} var(f_n(x)) = 0.$$

2.2.4 Erreur quadratique moyenne(MSE)

L'erreur quadratique moyenne (en anglais "Mean squared Error") est donne par :

$$MSE(x) = var f_n(x) + Biais^2 f_n(x)$$

Proposition 2.2.3

$$MSE(x) = \frac{1}{nh} f(x) \int_{-\infty}^{+\infty} K^2(u)du + o(\frac{1}{nh})$$

$$+\frac{h^4}{4}\{f''(x)\}^2 \left[\int_{-\infty}^{+\infty} u^2 K(u)du\right]^2 + o(h^4).$$

Démonstration

$$\begin{aligned} MSE(x) &= E\left(f_n(x) - f(x)\right)^2 \\ &= E\left(f_n(x) - E(f_n(x)) + E(f_n(x)) - f(x)\right)^2 \\ &= Var(f_n(x)) + \left(Biais(f_n(x))\right)^2 \\ &= \frac{1}{nh} f(x) \int_{-\infty}^{+\infty} K^2(u)du + o(\frac{1}{nh}) \\ &+\frac{h^4}{4}(f''(x))^2 \left[\int_{-\infty}^{+\infty} u^2 K(u)du\right]^2 + o(h^4) \end{aligned}$$

□

2.2.5 Erreur quadratique moyenne intégrée(MISE)

L'erreur quadratique moyenne intégrée (en anglais "Mean Integrated Squared Error") est donne par :

$$MISE(n,h) = \int_{-\infty}^{+\infty} MSE(x)dx$$

Proposition 2.2.4

$$\begin{aligned} MISE(n,h) &= \frac{1}{nh} \int_{-\infty}^{+\infty} (f(x) \left[\int_{-\infty}^{+\infty} K^2(u)du\right] + o(\frac{1}{nh})dx \\ &+\frac{h^4}{4} \int_{-\infty}^{+\infty} \left((f''(x))^2 \left[\int_{-\infty}^{+\infty} u^2 K(u)du\right]^2 + o(h^4)\right) dx. \end{aligned}$$

2.3 Convergence presque complète

Soient $X_1, ..., X_n$ un n-échantillon $i.i.d$ de X.

Hypothèses

(H1) : La fonction f est de classe $\mathcal{C}^k$ et $f(x) > 0$

(H2) : $\lim_{n\to\infty} h_n = 0, \lim_{n\to\infty} \frac{nh_n}{\log(n)} = \infty.$

(H3) : K est borné, intégrable, à support compacte et d'ordre k.

Théorème 2.3.1 *(La vitesse de convergence presque complète de l'estimateur à noyau de la densité sous la condition de dérivabilité) Supposons que les hypothèse* $(H1), (H2)$ *et* $(H3)$ *soient réalisées, alors, on a :*

$$f_n(x) - f(x) = o(h^k) + O\left(\sqrt{\frac{\log(n)}{nh}}\right) \quad p.co.$$

Preuve de théorème 2.3.1 On peut écrire :

$$f_n(x) - f(x) = \underbrace{f_n(x) - \mathbb{E}\left(f_n(x)\right)}_{La\ partie\ dsperssion} + \underbrace{\mathbb{E}\left(f_n(x)\right) - f(x)}_{La\ partie\ biais}$$

Alors, la preuve du théorème 2.3.1 est une conséquence du lemmes suivants :

Lemme 2.3.1 *Sous les hypothèse de théorème 2.3.1, on a :*

$$f_n(x) - \mathbb{E}\left(f_n(x)\right) = O\left(\sqrt{\frac{\log(n)}{nh}}\right) \quad p.co.$$

Lemme 2.3.2 *Sous les hypothèse de théorème 2.3.1, on a :*

$$\mathbb{E}\left(f_n(x)\right) - f_n(x) = o(h^k).$$

Preuve de lemme 2.3.1 Il faut montrer que : $\forall \epsilon > 0$,

$$\sum_{i\geq 1} \mathbb{P}\left(|f_n(x) - \mathbb{E}\left(f_n(x)\right)| > \epsilon\sqrt{\frac{\log(n)}{nh}}\right) < \infty.$$

c-à-d $\forall \epsilon > 0, \int_{n\to\infty} \mathbb{P}\left(|f_n(x) - \mathbb{E}\left(f_n(x)\right)| > \epsilon\sqrt{\frac{\log(n)}{nh}}\right) = 0.$

c-à-d $\exists u_n \to 0, \quad lorsque\ n \to 0$ t.q $\mathbb{P}\left(|f_n(x) - \mathbb{E}\left(f_n(x)\right)| > \epsilon\sqrt{\frac{\log(n)}{nh}}\right) \leq u_n \to 0.$

Pour montrer le lemme 2.3.1, on utilise le lemme suivant(inégalité exponentielle) :

Lemme 2.3.3 *Soit $\Delta_1, ..., \Delta_n$ une suite de variables aléatoires $i.i.d$, centrées, tel qu'il existe deux réls positifs d et δ^2, telsque : $|\Delta_1| < d$ et $\mathbb{E}(\Delta_i^2) < \delta^2$, alors : pour tout $\epsilon \in]0, \frac{\delta^2}{d}[$, on a :*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \Delta_n\right| > \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{4\delta^2}\right).$$

Suite de la preuve de lemme 2.3.1 On pose :

$$\Delta_i = h^{-1}\left(K\left(\frac{x - X_i}{h}\right) - \mathbb{E}\left(K\left(\frac{x - X_i}{h}\right)\right)\right).$$

donc,

$$f_n(x) - \mathbb{E}\left(f_n(x)\right) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$$

On a : $\Delta_1, ..., \Delta_n$ sont $i.i.d$ et :

$$\begin{aligned} |\Delta_i| &= \left|h^{-1}\left(K\left(\frac{x - X_1}{h}\right) - \mathbb{E}\left(K\left(\frac{x - X_1}{h}\right)\right)\right)\right| \\ &\leq h^{-1}\left|\left(K\left(\frac{x - X_1}{h}\right)\right| + \left|\mathbb{E}\left(K\left(\frac{x - X_1}{h}\right)\right)\right)\right| \end{aligned}$$

D'aprés $(H3), \exists M > 0$, t.q $|K(u)| \leq M \Rightarrow \mathbb{E}(K) \leq M$ Donc, $|\Delta_1| \leq \frac{C}{h}, C = 2M$. D'autre part, on a :

$$\begin{aligned} \mathbb{E}(\Delta_i^2) &= \mathbb{E}\left(h^{-1}\left(K\left(\frac{x - X_i}{h}\right) - \mathbb{E}\left(K\left(\frac{x - X_i}{h}\right)\right)\right)\right)^2 \\ &= \frac{1}{h^2}\left(\mathbb{E}\left(K^2\left(\frac{x - X_i}{h}\right)\right) - \mathbb{E}^2\left(K\left(\frac{x - X_i}{h}\right)\right)\right) \end{aligned}$$

(on fait le changement de variable)

$$= \frac{1}{h}\left(\int_{\mathbb{R}} K^2(z) f(x - hz) dz\right) - \frac{1}{h}\left(\int_{\mathbb{R}} K(z) f(x - hz) dz\right)^2$$

(f continue, $f(x - hz) \to f(x)$)

$$= \frac{1}{h}\left(\int_{\mathbb{R}} K^2(z) f(x - hz) dz - f^2(x)\right)$$

(f et K sont bornés et $\int K^2(u) du < \infty$), alors,

$\mathbb{E}(\Delta_i) \leq \dfrac{C}{h} \Rightarrow \delta^2 = \dfrac{C}{h}$

Donc, on peut appliquer le lemme 2.3.3 sur Δ_i, donc :

$$\mathbb{P}\left(|f_n(x) - \mathbb{E}\left(f_n(x)\right)| > \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\Delta_n\right| > \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{4\delta^2}\right).$$

On pose $\epsilon = \epsilon_0\sqrt{\frac{\log(n)}{nh}}$ on trouve :

$$\begin{aligned}
\mathbb{P}\left(|f_n(x) - \mathbb{E}\left(f_n(x)\right)| > \epsilon_0\sqrt{\frac{\log(n)}{nh}}\right) &\leq 2\exp\left(\frac{-n\epsilon_0^2\frac{\log(n)}{nh}}{4\frac{C}{h}}\right) \\
&\leq 2\exp\left(\frac{-n\epsilon_0^2\log(n)}{C}\right) \\
&\leq 2\exp\left(\log(n^{-\frac{\epsilon_0^2}{C}})\right) \\
&\leq n^{-\frac{\epsilon_0^2}{C}} \to 0 \; lorsque \; n \to 0
\end{aligned}$$

Donc,

$$\sum_{i=1}^{n}\mathbb{P}\left(|f_n(x) - \mathbb{E}\left(f_n(x)\right)| > \epsilon_0\sqrt{\frac{\log(n)}{nh}}\right) < \infty$$

□

Preuve du lemme 2.3.2

$$\mathbb{E}\left(f_n(x)\right) - f_n(x) = \int_{\mathbb{R}} K(z)f(x - hz)dz - f(x)$$

La fonction f est de classe $\mathcal{C}^k$, un dévellopement de Taylore d'ordre k de f au voisinage de x nous permet d'écrire la fonction f est de classe $\mathcal{C}^k$, un dévellopement de Taylore d'ordre k de f au voisinage de x nous permet d'écrire

$$f(x - hz) = f(x) + \sum_{j=1}^{k-1}\frac{(hz)^j(-1)^j}{j!}f^{(j)}(x) + \frac{(-1)^k(zh)^k}{k!}f^{(k)}(\theta_z)$$

Alors,

$$\begin{aligned}\mathbb{E}\left(f_n(x)\right) - f_n(x) &= \int_{\mathbb{R}} K(z)\left(\sum_{j=1}^{k-1}\frac{(hz)^j(-1)^j}{j!}f^{(j)}(x) + \frac{(-1)^k(zh)^k}{k!}f^{(k)}(\theta_z)\right)dz \\ &\quad (or\ K\ est\ d'ordre\ k) \\ &= \int_{\mathbb{R}} K(z)\frac{(-1)^k(zh)^k}{k!}f^{(k)}(\theta_z)dz \quad (f\ continue\ \Rightarrow f(\theta_z)\to f(x)) \\ &= h^k \underbrace{\frac{(-1)^k(z)^k}{k!}f^{(k)}(x)}_{cst}\underbrace{\int_{\mathbb{R}} K(z)dz}_{1} \Longrightarrow biais\left(f_n(x)\right) = o(h^k).\end{aligned}$$

□

Théorème 2.3.2 *(La vitesse de convergence presque complète de l'estimateur à noyau de la densité sous la condtion du continuité) Soit f une fonction continue et supposons que les hypothèses $(H2)$ et $(H3)$ soient réalisées, alors, on a :*

$$f_n(x) - f_n(x) = o(1) + O\left(\sqrt{\frac{\log(n)}{nh}}\right) \quad p.co.$$

Preuve du théorème 2.3.2 La preuve de le théorème est une conséquance de lemmes suivants :

Lemme 2.3.4 *Sous les conditions de théorème 2.3.2, on a :*

$$\lim_{n\to\infty} \mathbb{E}\left(f_n(x)\right) - f_n(x) = 0$$

Lemme 2.3.5 *Sous les conditions de théorème 2.3.2, on a :*

$$\mathbb{E}\left(f_n(x)\right) - f_n(x) = O\left(\sqrt{\frac{\log(n)}{nh}}\right) \quad p.co.$$

Preuve du lemme 2.3.4

$\mathbb{E}\left(f_n(x)\right) - f_n(x) = \int_{\mathbb{R}} K(z)f(x-hz)dz - f(x)$ Puisque f est continué, on a :

$$\lim_{n\to\infty} f(x-hz) = f(x)$$

donc,

$$\lim_{n\to\infty} \mathbb{E}\left(f_n(x)\right) - f_n(x) = 0$$

□

Preuve du lemme 2.3.5 La preuve du lemme 2.3.5 est similaire que la preuve de lemme 2.3.1 □

Théorème 2.3.3 *(La vitesse de convergence presque complète de l'estimateur à noyau de la densité sous la condtion du liptchize) Soit f une fonction β-liptchize* $(|f(x-f(y)| \leq L|x-y|\beta)$, $\int_{\mathbb{R}} K(z)|z\beta|dz < \infty$ *et supposons que les conditions* $(H2)$ *et* $(H3)$ *soient vérifiées, alors, on :*

$$f_n(x) - f_n(x) = o(h^\beta) + o\left(\sqrt{\frac{\log(n)}{nh}}\right) \quad p.co.$$

Preuve du théorème 2.3.3 La preuve de le théorème est une conséquance de lemmes suivants :

Lemme 2.3.6 *Sous les conditions de théorème 2.3.3, on a :*

$$\mathbb{E}\left(f_n(x)\right) - f_n(x) = o(h^\beta)$$

Lemme 2.3.7 *Sous les conditions de théorème 2.3.3, on a :*

$$\mathbb{E}\left(f_n(x)\right) - f_n(x) = o\left(\sqrt{\frac{\log(n)}{nh}}\right) \quad p.co.$$

Preuve du lemme 2.3.6

$$\begin{aligned}
\mathbb{E}\left(f_n(x)\right) - f_n(x) &= \int_{\mathbb{R}} K(z)f(x-hz)dz - f(x) \\
&= K(z)f(x-hz)dz - \int_{\mathbb{R}} K(z)f(x)dz \\
&= \int_{\mathbb{R}} K(z)(f(x-hz) - f(x))dz \\
|\mathbb{E}\left(f_n(x)\right) - f_n(x)| &\leq \int_{\mathbb{R}} |K(z)||f(x-hz) - f(x)|dz \quad (f\ est\ \beta - liptchize,\ on\ a) \\
&\leq \int_{\mathbb{R}} |K(z)|(hz)^\beta dz \\
&\leq h^\beta \underbrace{\int_{\mathbb{R}} K(z)|z^\beta|dz}_{cst} \leq Ch^\beta \quad \Longrightarrow le\ biais\ (f_n(x)) = O(h^\beta) \quad \square
\end{aligned}$$

2.4 Choix du paramètre de lissage h

Le paramètre de lissage h a un impact important sur les performances de l'estimateur $f_n(x)$. Il existe essentiellement deux façons de trouver la bande passante optimale. La première consiste à trouver les paramètres qui minimisent l'erreur quadratique moyenne de $f_n(x)$, c'est-à-dire

$$\arg\min \left[\mathbb{E}\left(f_n(x)\right) f(x)\right]^2$$

Ainsi, on obtient un paramètre de lissage optimal qui varie en fonction des x pour lesquels on veut estimer la fonction de densité f. La deuxième approche nous donne un paramètre de lissage globalement optimal qui ne dépend pas de x. Pour ce faire, on cherche à minimiser h de l'erreur quadratique moyenne intégrale (MISE), c'est-à-dire

$$h_{optimal} = \arg\min[MISE(n,h)] = \arg\min\left[\int_{\mathbb{R}} \mathbb{E}\{f_n(x) - f(x)\}^2 dx\right]$$

On suppose que la densité à estimer f et le noyau K sont des fonctions de carré intégrable, de sorte que le MISE est finie.

$$MISE(n,h) = \frac{h^4}{4} R(f'')(\mu_2(K))^2 + \frac{1}{nh} R(K) + o(\frac{1}{nh} + h^4)$$

l'approximation asymptôtique de la MISE est donné par

$$MISE(n,h) = \frac{h^4}{4} R(f'')(\mu_2(K))^2 + \frac{1}{nh} R(K)$$

On dérive ce AMISE par rapport à h et on égale à 0 ,on obtient

$$\frac{\partial}{\partial h} AMISE(n,h) = h^3 R(f'')(\mu_2(K))^2 - \frac{R(K)}{nh^2}$$

Si $\frac{\partial}{\partial h} AMISE(n,h) = 0$

$$\Rightarrow h_{optimal} = \left[\frac{R(K)}{R(f'')(\mu_2(K))^2}\right]^{\frac{1}{5}} n^{-1/5}$$

Cas particuliers

Soit $X_1, X_2, ..., Xn$ une suite de variables aléatoires de densité de probabilité f, supposons que f appartient à une famille de distributions normales $\mathcal{N}(\mu, \sigma^2)$

$$Si \begin{cases} X_i & \sim \mathcal{N}(\mu, \sigma^2) \\ K & \sim \mathcal{N}(0,1) \end{cases} \qquad Alors\ h_{optimal} = 1.06\hat{\sigma} n^{-1/5}$$

Si $f \sim \mathcal{N}(\mu, \sigma^2)$ alors $f(x) = \frac{1}{\sigma}\varphi(\frac{x-\mu}{\sigma})$, avec $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x/2}$, et

$$f^{''}(x) = \frac{1}{\sigma^3}\varphi^{''}\left(\frac{x-\mu}{\sigma}\right), \varphi^{''}(x) = \frac{1}{\sqrt{2\pi}}(x^2-1)e^{-x/2}$$

La quantité inconnue $R(f^{''})$ sécrit alors

$$\begin{aligned} R(f^{''}) &= \int_{-\infty}^{+\infty}[f^{''}(x)]^2dx \\ &= \frac{1}{\sigma^6}\int_{-\infty}^{+\infty}\left\{\varphi^{''}\left(\frac{x-\mu}{\sigma}\right)\right\}^2 dx \\ &= \frac{1}{\sigma^5}\int_{-\infty}^{+\infty}\{\varphi^{''}(v)\}^2dv \end{aligned}$$

Nous avons :

$$\begin{aligned} \varphi(v) &= \tfrac{1}{\sqrt{2\pi}}e^{-v/2} \\ &\Rightarrow \varphi^{'}(v) = \tfrac{v}{\sqrt{2\pi}}e^{-v/2} \\ &\Rightarrow \varphi^{''}(x) = \tfrac{1}{\sqrt{2\pi}}(v^2-1)e^{-v/2}. \end{aligned}$$

$$\begin{aligned} R(f^{''}) &= \frac{1}{\sigma^5}\int_{-\infty}^{+\infty}\left\{\frac{1}{\sqrt{2\pi}}(v^2-1)e^{-v/2}\right\}^2 dv \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{\int_{-\infty}^{+\infty}(v^4 - e^{-v/2}dv - 2\int_{-\infty}^{+\infty}v^2e^{-v^2}dv + 2\int_{-\infty}^{+\infty}v^2e^{-v^2}dv\right\} \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty}v^2e^{-v^2}dv + \int_{-\infty}^{+\infty}e^{-v^2}dv\right\} \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty}\frac{u^2}{2}e^{-u^2/2}\frac{1}{\sqrt{2}}du + \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2}}e^{-u^2/2}du\right\} \quad avec\ \ u = \sqrt{2}v \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{4}\sqrt{\pi} + \sqrt{\pi}\right\} \\ &= \frac{1}{\sigma^5}\frac{3}{8\sqrt{\pi}}. \end{aligned}$$

Donc,l'expression du paramètre de lissage optimal devient

$$h_{optimal} = \left[\frac{8\sqrt{\pi}R(K)}{3[\mu_2(K)]^2}\right]^{\frac{1}{5}}\hat{\sigma}_n^{-1/5}$$

Où $\hat{\sigma}$ est un estimateur de σ,tel que

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

On a $K \sim \mathcal{N}(0,1)$ alors

$$\begin{aligned} R(K) &= \int_{-\infty}^{+\infty} [K(u)]^2 du \\ &= \int_{-\infty}^{+\infty} \left[\frac{1}{\sqrt{2\pi}} e^{u^2/2}\right]^2 du \\ &= \int_{-\infty}^{+\infty} \frac{1}{2\pi} e^{-u^2} du \\ &= \frac{1}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}} \end{aligned}$$

Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires de densité de probabilité f, supposons que f appartient à une famille de distributions normales $\mathcal{N}(\mu, \sigma^2)$, soitt K est un noyau d'Epanechnikov.

$$Si \begin{cases} X_i & \sim \mathcal{N}(\mu, \sigma^2) \\ K(u) & = \frac{3}{4}(1-u^2), \mathbf{1}_{|u|\leq 1} \end{cases} \qquad Alors\ h_{optimal} = 2.34\hat{\sigma} n^{-1/5}$$

2.5 Choix du Noyaux

Le problème du choix optimal de K, consiste à chercher un noyau optimal sous la contrainte de posivité, $K \geq 0$. On fait le rappel de l'expression asymptotique de l'erreur quadratique intégrée AMISE

$$AMISE(n, h, K, f) = \frac{1}{nh}\int_{\mathbb{R}} [K(t)]^2 dt + \frac{1}{4}h^4 [V(K)]^2 \int_{\mathbb{R}} [f''(x)]^2 dx$$

on remarque que la dépendance du AMISE par rapport au noyau K s'exprime par l'intervention de sa variance $V(K)$. Un noyau optimal K^* est donc un noyau qui minimise la fonctionnelle $V(K)$, soit

$$V(K^*) = \min_{K \in \mathcal{K}} V(K) \tag{2.2}$$

où $\mathcal{K}$ désigne l'ensemble des noyaux positifs d'ordre 1 satisfaisant aux conditions

$$\int_{\mathbb{R}} [K(t)]^2 dt < +\infty, \int_{\mathbb{R}} t^2 K(t) dt < +\infty.$$

La solution du problème est donnée par la pro suivante :

Proposition 2.5.1 *Une solution du problème de minimisation* (2.2) *est donnée par le noyau d'Epanechnikov*

$$K^*(u) = \frac{3}{4}(1-u^2)$$

qui fournit la valeur minimale $V(K^*) = 3^{4/5}/5^{-6/5}$

On peut considérer l'eficacité de chacun des noyaux symétrique présenté dans le tableau (2.1),en comparant avec le noyau d'Epanechnikov.On défini l'éfficacité par :

$$\begin{aligned} eff(K) &= \left\{ \frac{C(K_e)}{C(K)} \right\}^{5/4} \\ &= \frac{3}{5\sqrt{5}} \frac{1}{\sqrt{\int u^2 K(u)du \int K(u)^2 d}} \end{aligned} \tag{2.3}$$

avec $C(K) = (V(K))^{2/5} \left\{ \int (K(u))^2 du \right\}^{4/5}$ la raison de la puissance $5/4$ dans (2.3) est que pour n grand l'erreur quadratique moyenne intégrée sera la même si on utilise n observations et le noyau K ou si on utise $n\, eff(K)$ observations et le noyau d'Epanechnikov K_e.
Le tableau (2.2) présente les valeurs d'efficacité de quelques noyaux continues symétriques.

Noyau	Efficacité
d'Epanechnikov	≈ 1.000
Biweight	≈ 0.9939
triangulaire	≈ 0.9859
Gaussien	≈ 0.9512
Rectangulaire	≈ 0.9295

TABLE 2.2 – Efficacité des noyaux continus symétriques.

On remarque que les valeurs d'efficacité obtenus sont très proches de 1 et qu'il ya très peu de différence entre les différents noyaux sur la base de l'erreur quadratique moyenne intégrée.

Chapitre 3

Application

3.1 Le paramètre de lissage h fixe, et n varié

3.1.1 Noyau à support non compact

Dans le premier cas, le paramètre de lissage ou fenêtre h est fixé $(h = n^{-1/5})$, on prend différentes valeurs de taille d'échantillon $(n = 70, n = 700)$, et K est un noyau normal $K(t) = \frac{e^{-t^2/2}}{\sqrt{2\pi}} t \in \mathbb{R}$, c'est une densité de support non compact.

Code R utilisé

```
n=70
X=rnorm(n)
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=n^-.2
# Initiation
s=100
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=numeric(n)
fn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
# Graphes
op=par(mfrow=c(1,3))
plot(x,fn,xlab="x", ylab="fn(x)",
main="n=70",type='l',col="red", lwd= 2)
```

```
lines(x,dnorm(x),lwd= 2)
#####Pour n =700
n=700
X=rnorm(n)
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=700",type='l',col="red", lwd= 2)
lines(x,dnorm(x),lwd= 2)
par(op)
```

La figure (3.1) représente la densité théorique en noir et l'estimateur à noyau de la densité en rouge.

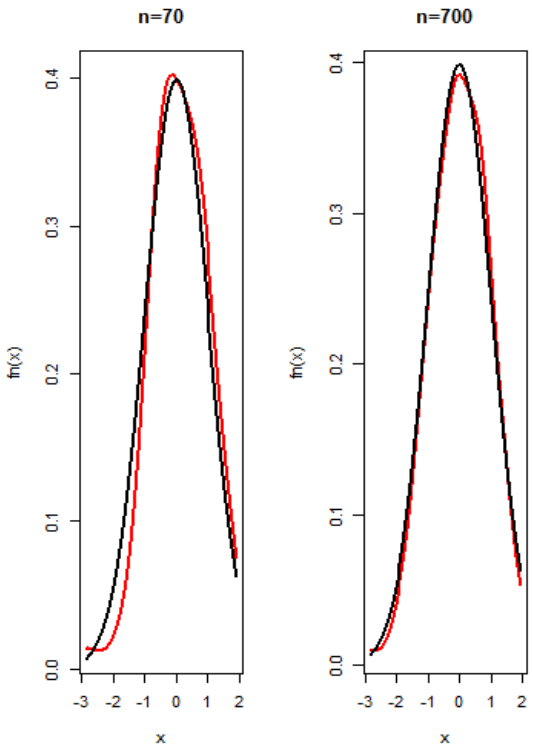

FIGURE 3.1 – Estimateur à noyau de la densité : h fixé, n varié et K noyau normal

Remarque 3.1.1 *Nous remarquons sur le graphe (3.1) que quand n est grand l'estimateur f_n est plus proche de la fonction f (estimateur lisse), ce qui implique que la convergence de l'estimateur.*

3.1.2 Noyau à support compact

On va refaire le même travail précédent, remplaçant seulement le noyau normal par le noyau de Triweight : $K(t) = \frac{35}{32}(1-t^2)^3\mathbf{1}_{(|t|<1)}$. (noyau à support compact).

Code R utilisé

```
n=70
X=rnorm(n)
K=function(t){ifelse(abs(t)<1,(35/32)*(1-t^2)^3,0)}
h=n^-.2
# Initiation
s=100
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=numeric(n)
fn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
# Graphes
op=par(mfrow=c(1,3))
plot(x,fn,xlab="x", ylab="fn(x)",
main="n=70",type='l',col="red", lwd= 2)
lines(x,dnorm(x),lwd= 2)
#####Pour n =700
n=700
X=rnorm(n)
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=700",type='l',col="red", lwd= 2)
lines(x,dnorm(x),lwd= 2)
par(op)
```

On obtient la figure (3.2) suivante :

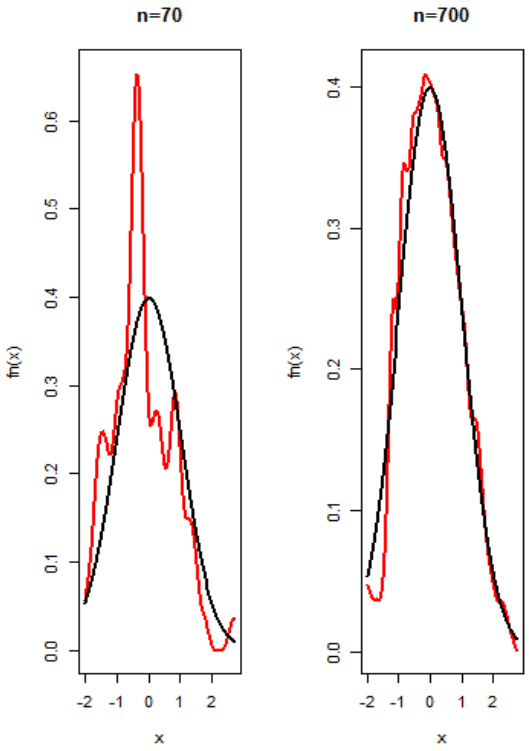

FIGURE 3.2 – Estimateur à noyau de la densité : h fixé, n varié et K noyau de Triweight

Remarque 3.1.2 *Nous remarquons ici que il n'y a pas beaucoup de variabilité sur le cas précédent, telle que l'estimateur lisse dès que n est grand (n = 700).*

3.2 Choix du paramètre de lissage

3.2.1 Noyau à support non compact

Considérons un échantillon de taille $n = 500$ et le noyau K est gaussien. Nous choisissons h variée dans l'intervalle $[0.1, 0.9]$.
La comparaison graphique entre les graphes de la densité théorique et empirique permet trouver une valeur h optimal (au sens graphique).

Code R utilisé :

```
n=500
X=rnorm(n)
# Noyau Normal
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=seq(.1,.9,length=9)
# Initiation
s=100 # taille de lintervalle [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="red", lwd= 2)
lines(x,dnorm(x),lwd= 2)}
par(op)
```

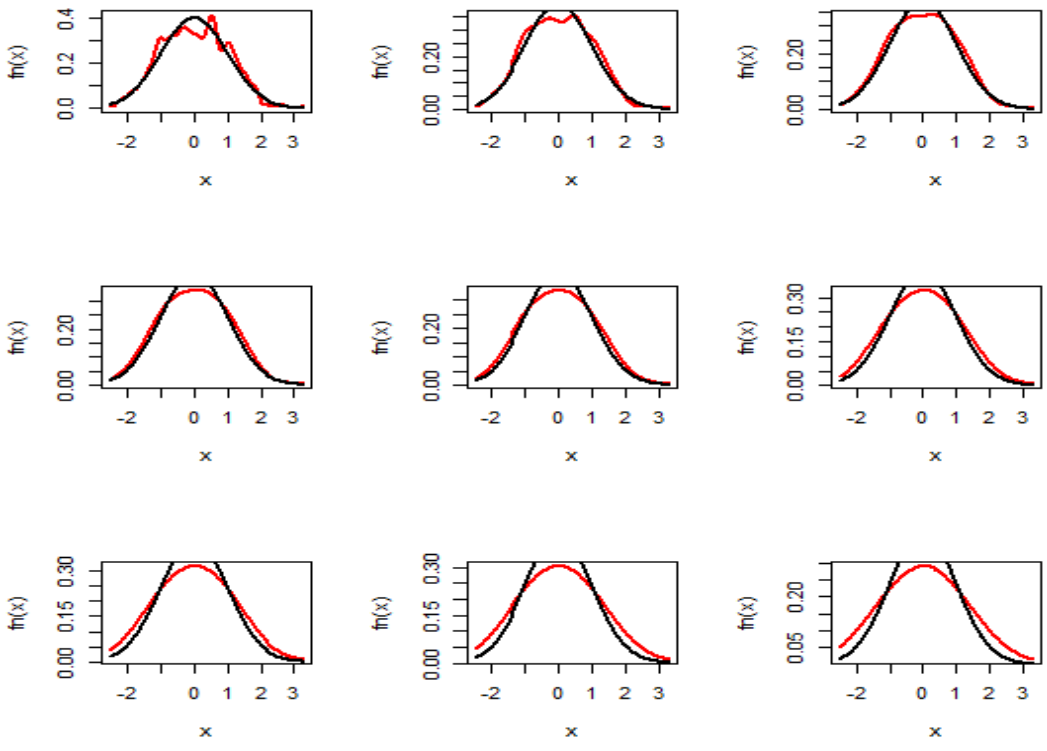

FIGURE 3.3 – Estimateur à noyau de la densité : h varié, n fixé et K noyau gaussien

Évidemment, la valeur optimale de h est $h = 0.4$ (ligne 2, colonne 1)

3.2.2 Noyau à support compact

A noté, que pour les mêmes données, pour le noyau de Triweight :
$K(t) = \frac{35}{32}(1-t^2)^3\mathbf{1}_{(|t|<1)}$.

Code R utilisé :

```
n=500
X=rnorm(n)
# Noyau Normal
K=function(t){ifelse(abs(t)<1,(35/32)*(1-t^2)^3,0)}
h=seq(.1,.9,length=9)
# Initiation
s=100 # taille de lintervalle [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="red", lwd= 2)
lines(x,dnorm(x),lwd= 2)}
par(op)
```

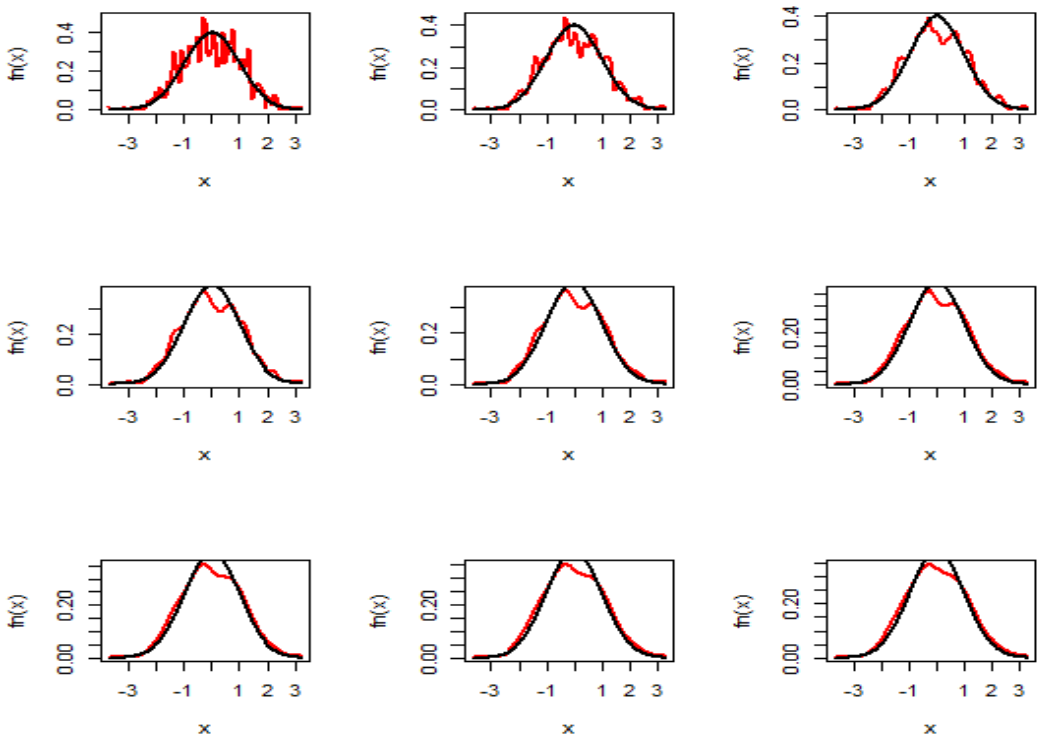

FIGURE 3.4 – Estimateur à noyau de la densité : h varié, n fixé et K noyau de Triweight

C'est clair La meilleure valeur pour h est $h = 0.9$ (ligne 3, colonne 3)

En conclusion, le type du noyau n'est pas trés inuent sur la qualité de l'estimation contrairement à la valeur de h, c'est le choix de la fenêtre h, qui est très important, par rapport au choix du noyau.

Conclusion

En conclusion, l'avantage de l'estimation par la méthode du noyau est qu'un densité continue (noyau) à partir d'une variable aléatoire, cette méthode dépend de nombre d'observations n et certains paramètres (paramètre de lissage h et le noyau K).

L'estimation du noyau est une méthode basée sur l'utilisation d'une méthode non paramètrique appelée fonction du noyau et un paramètre ou une fenêtre de lissage.

On remarque que le choix sur le noyau qui n'a pas une grande influence pour cette estimation, par contre le choix du paramètre de lissage a un impact important, et qui est en effet, beaucoup plus déterminant pour l'obtention des bons estimateurs.

Bibliographie

[1] Bosq, D. (2009). Estimation fonctionnelle. Techniques de lingénieur. Sciences fondamentales, (AF603).

[2] Chen S,X (1999). Beta kernel estimators for density functions. Comput Statist. Data Anal,31 (1999) 131-145.

[3] Devroye, L., Gyr, L., (1985) Nonparametric density estimation. The L1 view. Wiley, New York.

[4] Devroye, L. (1983). The equivalence of weak, strong and complete convergence in $ L_1 $ for kernel density estimates. The Annals of Statistics, 11(3), 896-904.

[5] Duin, R. P. W. (1976). On the choice of smoothing parameters of Parzen estimators of probability density function IEEE Transactions on Computers, C-25, 11751179.

[6] Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1), 153-158.

[7] Ferraty, F and Vieu, P. (2006). Nonparametric functional data analysis theory and pratice. Springer-Verlag.

[8] Gramacki, A. (2018). Nonparametric kernel density estimation and its computational aspects. Berlin : Springer International Publishing.

[9] Jones, M. C. (1990). The performance of kernel density functions in kernel distribution function estimation. Statistics & Probability Letters, 9(2), 129-132.

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

[11] Nadaraya, E.A. (1964). on estimateng regression, theorey of Probability and its Applications 9, 141-142.

[12] Parzen, E. (1962). on estimation of a probabiletey density fonction and mode, Annals of Mathematical statistics 33, 1065-1076.

[13] Rao, P. B. (1983). Non-parametric functional estimation Academic Press.

[14] Rosenblatt, M. (1956). Remarks on some nom parametrice estimates of a denssitey function, Annals of mathematical statistics 27, 832-837.

[15] Scoutt D.W(1979). On Optimal and Data-Based Histograms. Biometrika, Vol 66, No , (Dec,1979),pp 605-610.

[16] Scott, D. W., & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. Journal of the american Statistical association, 82(400), 1131-1146.

[17] Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.

[18] SONG XI CHEN (2000). Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics 54, 471-480.

[19] Tsybakov, A. B. (2008). Introduction to nonparametric estimation. Springer Science & Business Media.

[20] Watson , G.S (1964). smooth regression analyes, sankhey senes A 26, 359-372.

[21] Zhang B (1996). Some Asymptotic Results for Kernel Density Estimation under Random Censorship. Author(s) : Biao Source : Bernoulli, Vol. 2, No. 2 (Jun., 1996), pp. 183-198.

Résumé

Dans ce travail, nous étudions l'estimation non paramétrique de densité de probabilité par la méthode du noyau. La construction de l'estimateur est basée sur l'utilisation d'une densité K appelé noyau et d'un paramètre de lissage h.

Nous rappelons les propriétés asymptotiques de l'estimateur : la convergence presque complète et les propriétés de l'estimateur. Nous parlons aussi du choix de noyau et de paramètre de lissage.

Finalement, nous donnons des explications graphiques à l'aide du logiciel R qui nous permet d'observer l'influence de la taille de l'échantillon et les valeurs choisies de paramètres de ce l'estimateur.

Abstract

In this work, we study the nonparametric estimation of probability density by the kernel method. The construction of the estimator is based on the use of a density K called kernel and a smoothing parameter h.

We recall the asymptotic properties of the estimator: the almost complete convergence and the properties of the estimator. We also talk about the choice of kernel and smoothing parameter.

Finally, we give graphical explanations using R software which allows us to observe the influence of the sample size and the chosen values of parameters of this estimator.

الملخص

في هذا العمل درسنا التقدير الغير براميتري لدالة الاحتمال بطريقة النواة. يعتمد بناء المقدر على استخدام كثافة K تسمى نواة مع معامل تجانس h.

نتذكر الخصائص المقاربة للمقدر : التقارب شبه الكامل وخصائص المقدر. نتحدث أيضا عن اختيار النوا ة ومعامل التعتيم .

أخيرا، نقدم تفسيرات بيانية باستخدام برنامج R الذي يسمح لنا بمراقبة تأثير حجم العينة والقيم المختارة لمعاملات هذا المقدر.