MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDELHAMID BEN BADIS DE MOSTAGANEM
FACULTÉ DES SCIENCES EXACTES ET DE L'INFORMATIQUE
DÉPARTEMENT DE MATHÉMATIQUES ET INFORMATIQUE

MÉMOIRE

Master Académique

pour obtenir le diplôme de Master délivré par

Université de Mostaganem

Spécialité "Modélisation, Contrôle et Optimisation"

présenté et soutenu publiquement par

Mustapha REFAI

le 13 Juin 2018

Estimation Fonctionnelle de la Densité Conditionnelle

Encadreur : **Mustapha MOHAMMEDI (UNIVERSITÉ DE MOSTAGANEM, ALGÉRIE)**

Jury

ANDASMAS. Maamar. ,	Professeur	Président (Université de Mostaganem, Algérie)
LATREUCH. Zinelaabidine. ,	Professeur	Examinateur (Université de Mostaganem, Algérie)

LABORATOIRE DE MATHÉMATIQUES PURES ET APPLIQUÉES
FACULTÉ DES SCIENCES EXACTES ET DE L'INFORMATIQUE (FSEI)
Chemin des Crêtes (Ex-INES), 27000 Mostaganem, Algérie

MASTER

Résumé

Dans ce mémoire, nous proposons d'étudier quelques propriétés asymptotiques d'estimateurs non paramétriques d'une classe de fonction de répartition.

Dans ce cadre, Nous commençons par rappeler d'abord les notions essentielles d'estimation par noyaux. Nous examinons par la suite les propriétés des estimateurs plus précisément le biais, la variance et les erreurs quadratiques moyennes. Par ailleurs, Nous étudions l'estimation non paramétrique de la densité conditionnelle d'une variable réelle Y réponse n'est pas nécessairement bornée quand la variable explicative X est fonctionnelle.
Le but de ce travail est d'établir la convergence presque complète de l'estimateur à noyau sous certain nombre d'hypothèses, lorsque les observations sont i.i.d.
Par la suite, Nous présentons dans le dernier chapitre le travail des simulations sont faites pour illustrer les résultats théoriques établis sur la densité de probabilité. En dernier, Les résultats obtenus sont écrits sous la forme d'une conclusionn.

Mots clés : *Estimation non paramétrique, estimateur à noyau, propriétés asymptotiques d'estimateurs, erreur quadratique moyenne, La convergence presque complète.*

Remerciements

Je veux tout d'abord à remercier ALLAH le tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce Modeste travail.

En second lieu, Je veux remercier mon encadreur Mr : Mustapha Mohammedi, son précieux conseil et son aide durant toute la période du travail.

Mes plus profonds remerciements vont à mes parents. Tout au long de mon cursus, ils m'ont toujours soutenu, encouragé et aidé. Ils m'ont donner toutes les chances pour réussir. Qu'ils trouvent, dans la réalisation de ce travail, l'aboutissement de leurs efforts ainsi que l'expression de mon plus affectueuse gratitude.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à mon recherche en acceptant d'examiner mon travail et de l'enrichir par leurs propositions.

Enfin, je veux également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Sans oublier tout mes professeurs de l'université de Mostaganem.

Table des matières

Résumé **1**

Remerciements **2**

Introduction **6**

1 Estimation de la fonction de répartition **8**
1 Définition de l'estimateur (Statistique) 8
2 Estimateur empirique 9
3 Propriétés de l'estimateur 11
4 Généralisation 15

2 Estimation de la densité de probabilité **18**
1 Estimateur simple 18
2 Estimateur à noyau 20

3 Estimation de la densité conditionnelle **24**
1 Présentation des modèles non paramétriques conditionnels 24
2 Estimation de la fonction de répartition conditionnelle 25
3 L'estimateur à noyau de la densité conditionnelle 26

4 Simulation **30**
1 Présentation du logiciel R 30
2 Plan de simulation 31
3 Algorithme de simulation 32
4 Simulations et Résultats 32
5 Interprétation des résultats 37

Conclusion **38**

Liste des tableaux

4.1 Résultats de la simulation. 35
4.2 Résultats de la simulation par noyau. 36

Liste des figures

2.1 La représentation graphiques de ces noyaux 21

4.1 Démarrage de R pour Windows 31
4.2 Représentation graphique $f(x)$ 31
4.3 Représentation des courbes X_i 32
4.4 Estimations par noyau Gaussien 33
4.5 Densité théorique et empirique. 34
4.6 Représentation de la densité estimée avec la méthode du noyau 35
4.7 Représentation de la densité estimée avec autres noyaux 36

Introduction

L'objet principal de la statistique est de faire, à partir d'observations d'un phénomène aléatoire, une inférence au sujet de la loi générant ces observations en vue d'analyser le phénomène ou de prévoir un événement futur. Pour réduire la complexité du phénomène étudié, nous pouvons utilisé deux approches statistiques : non-paramétrique et paramétrique.

Dans le premier approche, un probléme récurrent en statistique est celui de l'estimation d'une densité f ou d'une fonction de répartition F à partir d'un échantillon de variables aléatoires réelles $X_1, X_2, ..., X_n$ indépendantes et de même loi inconnue. Les fonctions f et F , tout comme la fonction caractéristique, décrivent complètement la loi de probabilité des observations et en connaître une estimation convenable permet de résoudre nombre de problèmes statistiques. Cette estimation tient donc naturellement une place importante dans l'étude de nombreux phénomènes de nature aléatoire.

Pour estimer n'importe quel paramètre fonctionnel il suffit d'estimer la fonction de répartition F par la fonction de répartition empirique F_n , et par conséquent l'estimateur de θ_n est $T(F_n)$ où T est la fonctionnelle statistique.
La fonction de répartition empirique donc joue un rôle fondamental dans l'estimation fonctionnelle plus précisément dans l'estimation de la densité f, pour qu'on puisse tirer plus d'information sur la loi parente. La connaissance de l'estimateur de F et f mènent à résoudre un autre problème fondamental de la statistique non paramétrique, c'est le problème de la régression.

Les estimateurs non-paramétriques classiques ont été introduit par Rosemblatt (1956) pour estimer la fonction de densité, a été reprise simultanément par Waston (1964) et Nadaraja (1964) pour estimer une fonction de régression. Le comportement asymptotique de ces estimateurs a été étudié par de nombreux auteurs tel que Tsybakov (2004). Ainsi, le but de ce travail est de définir les estimateurs à noyau associé et d'établir les propriétés relatives.

Avant de présenter les résultats de façon détaillée, nous en donnons tout d'abord les grandes lignes.

Dans le premier chapitre on introduit le modèl non paramétrique et on présente une inégalité qui s'appelle « l'inégalité de Bernshtein .Frechet » , fondamentale dans l'étude de la vitesse de convergence ponctuelle des estimateurs fonctionnels. Nous étudions également l'estimateur à noyau pour la fonction de répartition en fin de chapitre.

Dans le deuxième chapitre, nous intéressons à l'estimateur à noyau de la densité f.

Dans le troisième chapitre, nous représentons pareillement les modèles non paramétriques conditionnels. ces modèles conduit a estimer la densité conditionnelle.

Dans le dernier chapitre, nous appliquons une partie de ces estimateurs à noyaux associés sur des données aléatoires.

Nous terminons ce rapport par une conclusion générale.

Nous présentons maintenant de manière plus développée le contenu des quatres chapitres de ce mémoire.

Chapitre 1

Estimation de la fonction de répartition

Dans ce premier chapitre, nous donnons la définition d'un estimateur. Nous présentons à partir de cette définition l'estimateur empirique de la fonction de répartition, ensuite nous étudions également les différents propriétés fondamentales de cet estimateur tell que biais, variance, erreur quadratique moyenne. Nous détaillons par la suite l'inégalité de Bernshtein Frechet qui nous aideront à estimer une fonction de répartition par la méthode du noyau, c'est ce que nous intéresse.

1 Définition de l'estimateur (Statistique)

Un estimateur est une statistique (variable aléatoire) permettent d'évaluer un paramètre inconnu relatif à une loi de probabilité (comme les caractéristiques de dispersions et de positions). Il peut par exemple servir à estimer certaines caractéristiques d'une population à partir de données obtenues sur un échantillon.

Définition 1.1 *Soit* $(\Omega, \mathrm{A}, \mathrm{P}_\theta)$ *est une structure statistique , où* $\theta \in \varphi$ *et* $\varphi \subset \mathbb{R}^k$.
Ω *:espace fondamental*
A *:tribu*
P *:ensemble de probabilité*
φ *:ensemble des paramètres*
• *Si* $k < \infty$ *, on dit que la statistique est paramétrique.*
• *Si* $k = \infty$ *, on dit que la statistique est fonctionnelle.*

Définition 1.2 *On appelle fonctionnelle statistique toute application* T *:*

$$\begin{array}{rl} \mathrm{T}: & \digamma \longrightarrow \Phi \\ & \mathrm{F} \longrightarrow \mathrm{T}(\mathrm{F}) \quad = \theta \end{array}$$

où ;
$\digamma$ *: l'espace des fonctions de répartition*
Φ *: l'espace des paramétres.*
F *: la fonction de répartition.*
On dit que θ *un paramètre fonctionnel.*

Exemple 1.1

1. La densité est un paramètre fonctionnel.
   f : paramètre fonctionnel car $d\mathrm{F} = f$.

2. L'espérence de X

$$\mathrm{E}(\mathrm{X}) = \int \mathrm{X}\, d\mathrm{F} = \mathrm{T}(\mathrm{F})$$

2 Estimateur empirique

2.1 La fonction de répartition empirique

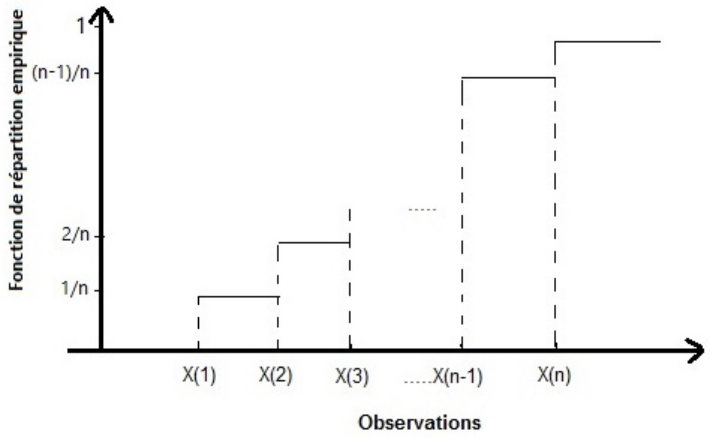
Soit $\mathrm{F}(x) = \mathrm{P}(\mathrm{X} \leq x)$ la fonction de répartition de X.
Soit $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$ un échantillon i.i.d. de F (i.i.d.= indépendantes et identiquement distribuées) et

$$\mathrm{X}_{(1)} \leq \mathrm{X}_{(2)} \leq ... \leq \mathrm{X}_{(n)}$$

les observations ordonnées.

Supposons que F soit complétement inconnue.
Comment estimer F, en se basant sur les observations $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$?
Un bon estimateur pour F est la fonction de répartition empirique, notée F_n, et définie par

$$\begin{aligned}
\mathrm{F}_n(x) &= \frac{nombre\ d'observations \leq x}{n} \\
&= \frac{card\{i : \mathrm{X}_i \leq x\}}{card\ \Omega} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{]-\infty,x]}(\mathrm{X}_i) \\
&= \begin{cases} 0 & si\ \ x < \mathrm{X}_{(1)} \\ \frac{k}{n} & si\ \ \mathrm{X}_{(k)} \leq x \leq \mathrm{X}_{(k+1)} \qquad k = 1, ...n-1 \\ 1 & si\ \ x \geq \mathrm{X}_{(n)} \end{cases}
\end{aligned}$$



2.2 Propriétés élémentaires de la fonction de répartition empirique

Biais de l'estimateur $\mathrm{F}_n(x)$

Théorème 1.1 $\mathrm{F}_n(x)$ *est un estimateur sans biais de* $\mathrm{F}(x)$

Preuve. Soit $X_1, X_2, ..., X_n$ un n_échantillon de X.

$$F_n(x) \in \{0,1\} \Longrightarrow nF_n(x) \in \{0,n\}$$

$$\begin{aligned} P(nF_n(x)=0) &= P(x < X_{(1)}) \\ &= P(x < X_i\ , \forall i) \\ &= P(\bigcap_{i=1}^{n} \{\ x < X_i\ \}) \\ &= \prod_{i=1}^{n} P(x < X_i) \\ &= \prod_{i=1}^{n} (1 - P(X_i \leq x)) \\ &= (1 - F(x))^n \end{aligned}$$

Donc $nF_n(x)$ est une variable aléatoire de loi binomiale des paramètres $(n, F(x))$.
c.à.d

$$nF_n(x) = \sum_{i=1}^{n} \mathbf{I}_{]-\infty,x]}(X_i) \sim Bin\ (n, F(x))$$

Alors

$$\begin{aligned} E[nF_n(x)] = nF(x) &\Longrightarrow nE[F_n(x)] = nF(x) \\ &\Longrightarrow E[F_n(x)] = F(x) \end{aligned}$$

Donc , $F_n(x)$ est un estimateur sans biais de $F(x)$. ■

Variance de l'estimateur $F_n(x)$

Il est facile de montrer que, pour tout x, la variance de l'estimateur $F_n(x)$ est donnée par :

$$Var[nF_n(x)] = nF(x)(1-F(x)) \Longrightarrow Var[F_n(x)] = \frac{F(x)(1-F(x))}{n}$$

Remarquons que , si $n \to \infty$, alors $Var[F_n(x)]$ converge vers 0.

L'erreur quadratique moyenne de l'estimateur $F_n(x)$

$$\begin{aligned} E[F_n(x) - F(x)]^2 &= E[F_n(x) - E[F_n(x)] + E[F_n(x)] - F(x)]^2 \\ &= Var[F_n(x)] + [Biais\{F_n(x)\}]^2 \\ &= \frac{1}{n}F(x)(1-F(x)) \end{aligned}$$

Donc , quand $n \to \infty$, on a que

$$E[F_n(x) - F(x)] \to 0$$

pour tout point x. L'estimateur $F_n(x)$ est alors un estimateur consistent de $F(x)$.

3 Propriétés de l'estimateur

La qualité des estimateurs s'exprime par leur convergence, leur biais, leur efficacité. Diverses méthodes permettent d'obtenir des estimateurs de qualités différentes.

3.1 Inégalité de Bernshtein Fréchet

Lemme 1.1 *Soit* $X_1, X_2, ..., X_n$ *une suite de variables aléatoires indépendantes , tel que,* $\alpha_i \leq X_i \leq \beta_i$ *et* $\alpha_i, \beta_i \in \mathbb{R}$
Alors, $\forall\, t > 0$ *on a :*

$$P\left[\left|\sum_{i=1}^{n}(X_i - E(X_i))\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(\beta_i - \alpha_i)^2}\right)$$

Preuve.
On montre que :

$\forall\, h > 0$

$$\mathbf{I}_{A=\left(\sum_{i=1}^{n}(X_i - E(X_i)) - t \geq 0\right)} \leq \exp\left(h\sum_{i=1}^{n}(X_i - E(X_i)) - t\right) \tag{1.1}$$

et en déduire l'Inégalité de Bernshtein Fréchet pour $\forall\, t > 0$

$$P\left[\left|\sum_{i=1}^{n}(X_i - E(X_i))\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(\beta_i - \alpha_i)^2}\right)$$

$$\mathbf{I}_A(w) = \begin{cases} 1 & \text{si} \quad w \in A \\ 0 & \text{sinon} \end{cases}$$

$\star$ Si $\mathbf{I}_A = 0$

$$\sum_{i=1}^{n}(X_i - E(X_i)) - t \leq 0$$

$\star$ Si $\mathbf{I}_A = 1$

$$\begin{aligned} \sum_{i=1}^{n}(X_i - E(X_i)) - t &\geq 0 \\ &\quad et \\ \exp\left(h\sum_{i=1}^{n}(X_i - E(X_i)) - t\right) &\geq 1 \end{aligned}$$

Alors (1.1)est vrais pour les deux cas.
On sait que $E[\mathbf{I}_A(X)] = P(A)$

$$\begin{aligned}
P(A) &= P\left(\sum_{i=1}^{n}(X_i - E(X_i)) - t \geq 0\right) \\
&\leq E\left[\exp\left(h\sum_{i=1}^{n}(X_i - E(X_i)) - t\right)\right] \\
&\leq \exp(-h\,t).E\left[\exp\left(h\sum_{i=1}^{n}(X_i - E(X_i))\right)\right] \\
&\leq \exp(-h\,t).\prod_{i=1}^{n}E\left[\exp(h(X_i - E(X_i)))\right] \\
&\leq \prod_{i=1}^{n}E\left[\exp(-h\,E(X_i))\right].E\left[\exp(h\,X_i)\right].\exp(-h\,t)
\end{aligned}$$

Pour $(h\,X_i)$ on va utiliser le fait que cette fonction est convexe, on pose

$$\varphi(X_i) = \exp(h\,X_i)$$

où φ est convexe vérifie

$$\varphi(\alpha x + \beta y) \leq \alpha\varphi(x) + \beta\varphi(y) \qquad où\ \alpha, \beta \in \mathbb{R}$$

Posons :

$$\alpha = \frac{\beta_i - X_i}{\beta_i - \alpha_i} \qquad \text{et} \qquad \beta = \frac{X_i - \alpha_i}{\beta_i - \alpha_i}$$

Il est claire que $\alpha + \beta = 1$
On pose

$$\begin{aligned}
X_i &= \alpha x + \beta y \\
x &= \alpha_i \\
y &= \beta_i
\end{aligned}$$

$$\begin{aligned}
\varphi(\alpha x + \beta y) &= \exp(h\,(\alpha x + \beta y)) \\
&\leq \alpha\varphi(x) + \beta\varphi(y)
\end{aligned}$$

$$\Longrightarrow \quad \exp(h\,X_i) \leq \frac{\beta_i - X_i}{\beta_i - \alpha_i}\exp(h\,\alpha_i) + \frac{X_i - \alpha_i}{\beta_i - \alpha_i}\exp(h\,\beta_i)$$

$$\Longrightarrow \quad E\left[\exp(h\,X_i)\right] \leq \frac{\beta_i - E[X_i]}{\beta_i - \alpha_i}\exp(h\,\alpha_i) + \frac{E[X_i] - \alpha_i}{\beta_i - \alpha_i}\exp(h\,\beta_i)$$

Alors

$$E\left[\exp(-h\,E(X_i))\right].E\left[\exp(h\,X_i)\right] \leq E\left[\exp(-h\,E(X_i))\right].\left[\frac{\beta_i - E[X_i]}{\beta_i - \alpha_i}\exp(h\,\alpha_i) + \frac{E[X_i] - \alpha_i}{\beta_i - \alpha_i}\exp(h\,\beta_i)\right] \tag{1.2}$$

On éssaye de mettre (1.2) sous la forme $\exp(\psi(h_i))$ tel que : $h_i = h(\beta_i - \alpha_i)$
c'est à dire ;

$$(1.2)(h) = \exp(\psi(h_i))$$

D'aprés le develeppement limitée de $\psi(h_i)$ on a :

$$\psi(h_i) = \psi(0) + \psi'(0)\,h_i + \frac{1}{2}\psi''(\xi)\,h_i{}^2 \qquad où\ \xi \in [0, h_i]$$

On trouve que :

$$\psi(0) = 0 \quad , \quad \psi'(0) = 0 \quad \text{et} \quad \psi''(h_i) \leq \frac{1}{4}$$

Donc :

$$\left|\psi(h_i)\right| \leq \frac{1}{8} h_i{}^2 = \frac{h^2 (\beta_i - \alpha_i)^2}{8}$$

On peut dire que :

$$(1.2)(h_i) = \exp\left(\frac{h^2 (\beta_i - \alpha_i)^2}{8}\right)$$

celà veut dire

$$\begin{aligned} P\left(\sum_{i=1}^{n} (X_i - E(X_i)) - t \geq 0\right) &\leq \exp(-h\, t) . \prod_{i=1}^{n} E\left[\exp(h(X_i - E(X_i)))\right] \\ &\leq \exp(-h\, t) . \exp\left[\frac{h^2 \sum_{i=1}^{n} (\beta_i - \alpha_i)^2}{8}\right] \\ &\leq \exp\left(-h\, t + \frac{h^2 \sum_{i=1}^{n} (\beta_i - \alpha_i)^2}{8}\right) \end{aligned}$$

La relation est vraie pour $h \geq 0$.

On pose : $h = \frac{4t}{\sum_{i=1}^{n} (\beta_i - \alpha_i)^2} \geq 0$

$$\begin{aligned} P\left(\sum_{i=1}^{n} (X_i - E(X_i)) - t \geq 0\right) &\leq \exp\left(\frac{-4t^2}{\sum_{i=1}^{n} (\beta_i - \alpha_i)^2} + \frac{16t^2}{8\sum_{i=1}^{n} (\beta_i - \alpha_i)^2}\right) \\ &\leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (\beta_i - \alpha_i)^2}\right) \end{aligned}$$

On trouve le mème résultat pour :

$$A = \left(\sum_{i=1}^{n} (X_i - E(X_i)) - t \leq 0\right)$$

Enfin on conclut :

$$P\left[\left|\sum_{i=1}^{n} (X_i - E(X_i))\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (\beta_i - \alpha_i)^2}\right)$$

■

3.2 La vitesse de convergence

Définition 1.3 *La convergence presque complète.*
Soit $(X_n)_n$, $n \in \mathbb{N}$, *une suite de variables aléatoires. On dit que* X_n *converge presque complètement vers* X *si :*

$\forall \varepsilon > 0$,

$$\sum_{n=0}^{\infty} P\ [|X_n - X| > \varepsilon] < \infty$$

ça veut dire,

$$X_n \to X \quad quand \ \ n \to \infty$$

Définition 1.4 *vitesse de convergence en p.c.o.*
On dit que $X_n = 0(Y_n)$ *en p.c.o* $\left(X_n \text{ converge vers 0 pour une vitesse } Y_n\right)$
Si $\exists\, \varepsilon > 0$,

$$\sum_{n=0}^{\infty} P\,[|X_n| > \varepsilon\,|Y_n|] < \infty$$

Théorème 1.2 *Soit* $X_1, X_2, ..., X_n$ *un n_échantillon de* X *de fonction de répartition* F *et* F_n *la fonction empirique.*
Alors, pour tout x on a :

$$F_n(x) - F(x) = 0\left(\sqrt{\frac{\log n}{n}}\right) \ \ p.c.o$$

Preuve.
D'aprés la définition précédent de vitesse de convergence, il suffit de montrer que :
$\exists\, \varepsilon > 0$,

$$\sum_{n=0}^{\infty} P\left[|F_n(x) - F(x)| > \varepsilon\sqrt{\frac{\log n}{n}}\right] < \infty$$

on note :

$$A = \sum_{n=0}^{\infty} P\left[|F_n(x) - F(x)| > \varepsilon\sqrt{\frac{\log n}{n}}\right]$$

on a :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{]-\infty,x]}(X_i) \quad \text{et} \ \ F(x) = E\,[\,F_n(x)]$$

Alors A devient :

$$\begin{aligned} A &= \sum_{n=0}^{\infty} P\left[\left|\tfrac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{]-\infty,x]}(X_i) - \tfrac{1}{n}E\left[\sum_{i=1}^{n} \mathbf{I}_{]-\infty,x]}(X_i)\right]\right| > \varepsilon\sqrt{\tfrac{\log n}{n}}\right] \\ &= \sum_{n=0}^{\infty} P\left[\sum_{i=1}^{n}\left|\mathbf{I}_{]-\infty,x]}(X_i) - E\,\left[\,\mathbf{I}_{]-\infty,x]}(X_i)\right]\right| > \varepsilon\sqrt{n\log n}\right] \end{aligned}$$

On utilise l'inégalité de Bernshtein Fréchet .
On sait que :

$$0 \leq \mathbf{I}_{]-\infty,x]}(\mathrm{X}_i) \leq 1$$

Possons $t = \varepsilon\sqrt{n\log n}$. donc, par identification $\alpha_i = 0$ et $\beta_i = 1$.
On trouve :

$$\begin{aligned} \mathrm{A} = \sum_{n=0}^{\infty} \mathrm{P}\left(\sum_{i=1}^{n} \left| \mathbf{I}_{]-\infty,x]}(\mathrm{X}_i) - \mathrm{E}\left[\mathbf{I}_{]-\infty,x]}(\mathrm{X}_i) \right] \right| > t \right) &\leq \sum_{n=0}^{\infty} 2\exp\left(\frac{-2\varepsilon^2 n\log n}{n} \right) \\ &\leq \sum_{n=0}^{\infty} 2\exp\left(\log n^{-2\varepsilon^2} \right) \\ &\leq \sum_{n=0}^{\infty} 2n^{-2\varepsilon^2} \\ &\leq \sum_{n=0}^{\infty} \frac{2}{n^{2\varepsilon^2}} \end{aligned}$$

Donc, $\exists\, \varepsilon > 0$ tel que :

$$2\varepsilon^2 > 1 \Longrightarrow \varepsilon > \frac{1}{\sqrt{2}}$$

Pour que la série converge.
D'où

$$\sum_{n=0}^{\infty} \mathrm{P}\left[|\mathrm{F}_n(x) - \mathrm{F}(x)| > \varepsilon\sqrt{\frac{\log n}{n}} \right] < \infty$$

De cela, nous concluons que

$$\mathrm{F}_n(x) - \mathrm{F}(x) = 0\left(\sqrt{\frac{\log n}{n}} \right) \text{ p.c.o}$$

■

4 Généralisation

4.1 Estimateur à noyau

Soit $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$ un n_échantillon de X de fonction de répartition F.
On appelle estimateur à noyau pour F noté :

$$\widetilde{\mathrm{F}}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}\left(\frac{x - \mathrm{X}_i}{h_n} \right)$$

Ou H est une fonction de répartition quelquanque et h_n est une suite des nombres réels positifs.

Proposition 1.1 *Soit* $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$ *un n_échantillon de* X *de fonction de répartition* F. *Soit* $\widetilde{\mathrm{F}}_n(x)$ *un estimateur à noyau* H *vérifiant*

$\int_{\mathbb{R}} y\,\mathrm{H}'(y)\,dy < \infty \Longrightarrow \widetilde{\mathrm{F}}_n(x)$ *est un estimateur asymptotiquement sans biais de* $\mathrm{F}(x)$

Preuve. Il suffit de montrer que

$$\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] \to \mathrm{F}(x)\ , \quad \text{si } n \to \infty$$

$$\begin{aligned}
\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{H}\left(\frac{x-\mathrm{X}_i}{h_n}\right)\right] \\
&= \mathrm{E}\left[\mathrm{H}\left(\frac{x-\mathrm{X}_1}{h_n}\right)\right] \\
&= \int_{\mathbb{R}}\mathrm{H}\left(\frac{x-z}{h_n}\right) f(z)\,dz \\
&= \left[\mathrm{H}\left(\frac{x-z}{h_n}\right)\mathrm{F}(z)\right]_{-\infty}^{+\infty} + \frac{1}{h_n}\int_{-\infty}^{+\infty}\mathrm{H}'\left(\frac{x-z}{h_n}\right)\mathrm{F}(z)\,dz
\end{aligned}$$

Par un changement de variables, on pose :

$$\begin{aligned}
\frac{x-z}{h_n} = y &\Longrightarrow z = x - yh_n \\
&\Longrightarrow dz = -h_n\,dy
\end{aligned}$$

On voit que la première terme est nulle car elle est composée de deux fonctions de répartitions.
Alors, $\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right]$ devient :

$$\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] = \int_{-\infty}^{+\infty}\mathrm{H}'(y)\,\mathrm{F}(x-yh_n)\,dy$$

D'aprés le developpement limité de $\mathrm{F}(x-yh_n)$ on a :

$$\mathrm{F}(x-yh_n) = \mathrm{F}(x) - h_n\,y\,\mathrm{F}'(x) + o\left({h_n}^2\right)$$

Donc ;

$$\begin{aligned}
\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] &= \int_{-\infty}^{+\infty}\mathrm{H}'(y)\left(\mathrm{F}(x) - h_n\,y\,\mathrm{F}'(x) + o\left({h_n}^2\right)\right) \\
&= \mathrm{F}(x)\int_{-\infty}^{+\infty}\mathrm{H}'(y)\,dy - h_n\,f(x)\int_{-\infty}^{+\infty}y\,\mathrm{H}'(y)\,dy + o\left({h_n}^2\right)\int_{-\infty}^{+\infty}\mathrm{H}'(y)\,dy \\
&= \mathrm{F}(x) - h_n\int_{-\infty}^{+\infty}y\,\mathrm{H}'(y)\,dy + o\left({h_n}^2\right)
\end{aligned}$$

On pose :

$$\mathrm{M} = \int_{-\infty}^{+\infty}y\mathrm{H}'(y)\,dy$$

Si $\mathrm{M} < +\infty$ Alors ;

$$\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] - \mathrm{F}(x) = -h_n\,\mathrm{M} + o\left({h_n}^2\right) \longrightarrow 0$$

■

4.2 Expressions du variance et de L'erreur quadratique moyenne

Considérons l'estimateur à noyau

$$\widetilde{\mathrm{F}}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}\left(\frac{x-\mathrm{X}_i}{h_n}\right)$$

La variance de l'estimateur à noyau est donnée par :

$$\begin{aligned}
\mathrm{Var}\left[\widetilde{\mathrm{F}}_n(x)\right] &= \mathrm{Var}\left[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathrm{H}\left(\tfrac{x-\mathrm{X}_i}{h_n}\right)\right] \\
&= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} \mathrm{Var}\left[\mathrm{H}\left(\tfrac{x-\mathrm{X}_i}{h_n}\right)\right] \\
&= \tfrac{1}{n}\mathrm{Var}\left[\mathrm{H}\left(\tfrac{x-\mathrm{X}_1}{h_n}\right)\right] \qquad \text{car les } \mathrm{X}_i \text{ sont identiquement distribuées} \\
&= \tfrac{1}{n}\left\{\mathrm{E}\left[\ \mathrm{H}^2\left(\tfrac{x-\mathrm{X}_1}{h_n}\right)\right] - \left(\mathrm{E}\left[\ \mathrm{H}\left(\tfrac{x-\mathrm{X}_1}{h_n}\right)\right]\right)^2\right\} \\
&= \tfrac{1}{n}\left\{\textstyle\int_{\mathbb{R}} \mathrm{H}^2\left(\tfrac{x-z}{h_n}\right) f(z)\ dz - \mathrm{F}^2(x)\right\} \qquad \text{car } \mathrm{E}\left[\ \mathrm{H}\left(\tfrac{x-\mathrm{X}_1}{h_n}\right)\right] \to \mathrm{F}(x)\ , \quad \text{si } n\to\infty \\
&= \tfrac{1}{n}\left\{\left[\mathrm{H}^2\left(\tfrac{x-z}{h_n}\right).\mathrm{F}(z)\right]_{-\infty}^{+\infty} + \textstyle\int_{\mathbb{R}} \left(\mathrm{H}^2\left(\tfrac{x-z}{h_n}\right)\right)' \mathrm{F}(z)\ dz - \mathrm{F}^2(x)\right\} \qquad \text{On pose } y = \tfrac{x-z}{h_n} \\
&= \tfrac{1}{n}\left\{\mathrm{F}(x)\ \textstyle\int_{\mathbb{R}} y\left(\mathrm{H}^2(\ y)\right)'\ dy - \mathrm{F}^2(x)\right\}
\end{aligned}$$

On sait que

$$\int_{\mathbb{R}} y\,\mathrm{H}'(y)\ dy < \infty \Longrightarrow \int_{\mathbb{R}} y\ \left(\mathrm{H}^2(y)\right)'\ dy < \infty$$

D'où, $\mathrm{Var}\left[\widetilde{\mathrm{F}}_n(x)\right] \to 0$ quand $n \to \infty$

Pour L'erreur quadratique moyenne on calcul :

$$\begin{aligned}
\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x) - \mathrm{F}(x)\right]^2 &= \mathrm{E}\left[\widetilde{\mathrm{F}}_n(x) - \mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] + \mathrm{E}\left[\widetilde{\mathrm{F}}_n(x)\right] - \mathrm{F}(x)\right]^2 \\
&= \mathrm{Var}\ \left[\widetilde{\mathrm{F}}_n(x)\right] + \left[\mathrm{B}ias\left\{\widetilde{\mathrm{F}}_n(x)\right\}\right]^2
\end{aligned}$$

D'aprés les résultats précédent on a montrer que :

$\mathrm{B}ias\left\{\widetilde{\mathrm{F}}_n(x)\right\} \to 0$ et $\mathrm{Var}\ \left\{\widetilde{\mathrm{F}}_n(x)\right\} \to 0$ si $n \to \infty$

Alors ;

$$\mathrm{E}\left[\widetilde{\mathrm{F}}_n(x) - \mathrm{F}(x)\right]^2 \to 0$$

Chapitre 2

Estimation de la densité de probabilité

Comment estimer non-paramétriquement la densité de probabilité f , en se basant sur les observations $X_1, X_2, ..., X_n$? Il existe plusieurs méthodes d'estimation non-paramétrique d'une densité. L'objectif de notre étude dans ce chapitre est la construction d'un estimateur de f , c'est-à-dire une fonction $\widehat{f}_n(x) = f(x, X_1, X_2, ..., X_n)$ par la méthode du noyau.

1 Estimateur simple

Rappelons que la densité de probabilité f est égale à la dérivée de la fonction de répartition F (si cette dérivée existe). On peut donc écrire, quand $h \to 0$

$$\begin{aligned} f(x) = F'(x) &= \frac{F(x+h)-F(x-h)}{2h} \\ &= \frac{P[\,x-h < X \leq x+h]}{2h} \end{aligned}$$

Un estimateur de $f(x)$ est alors

$$\begin{aligned} f_n(x) &= \frac{F_n(x+h)-F_n(x-h)}{2h} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbf{I}_{\{X_i \in [x-h\,,\,x+h]\}}(X_i) \\ &= \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2}\mathbf{I}_{\left\{-1 \leq \frac{x-X_i}{h} < 1\right\}}(X_i) \end{aligned}$$

Notons que cette estimateur peut encore s'écrire comme

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

Où

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

Cet estimateur, appelé estimateur de Rosenblatt (1956), est le premier exemple d'estimateur à noyau construit à l'aide du noyau $W(y) = \frac{1}{2}\mathbf{I}_{\{-1 \leq y < 1\}}$, notion que nous allons étudier plus tard.

Quelles sont les propriétés de l'estimateur simple $f_n(x)$?

Remarquons que

$$f_n(x) = \frac{\mathrm{F}_n(x+h) - \mathrm{F}_n(x-h)}{2h}$$

avec F_n la fonction de répartition empirique. Le paramètre de lissage h dépend de la taille de l'échantillon n, c'est-à-dire $h = h_n$.

Nous savons que

$$n\mathrm{F}_n(x) = \sum_{i=1}^{n} \mathbf{I}_{\{\mathrm{X}_i \leq x\}}(\mathrm{X}_i) \sim \mathrm{B}in\ (n, \mathrm{F}(x))$$

et

$$2nh_n\ f_n(x) = n\mathrm{F}_n(x+h_n) - n\mathrm{F}_n(x-h_n) \sim \mathrm{B}in\ (n, \mathrm{F}(x+h_n) - \mathrm{F}(x-h_n))$$

$$\Longrightarrow \mathrm{E}\left[2nh_n\ f_n(x)\right] = n\ [\mathrm{F}(x+h_n) - \mathrm{F}(x-h_n)]$$

$$\Longrightarrow \mathrm{E}\left[\ f_n(x)\right] = \tfrac{1}{2h_n}\ [\mathrm{F}(x+h_n) - \mathrm{F}(x-h_n)]$$

Pour la variance nous trouvons

$$\begin{aligned} \mathrm{Var}\left[2nh_n\ f_n(x)\right] &= n\ [\mathrm{F}(x+h_n) - \mathrm{F}(x-h_n)]\ [1 - \mathrm{F}(x+h_n) - \mathrm{F}(x-h_n)] \\ \Longrightarrow \mathrm{Var}\left[\ f_n(x)\right] &= \tfrac{1}{4nh_n^2}\ [\mathrm{F}(x+h_n) - \mathrm{F}(x-h_n)]\ [1 - \mathrm{F}(x+h_n) - \mathrm{F}(x-h_n)] \end{aligned}$$

Remarquons que , si $n \to \infty$ et $h_n \to 0$, alors

$$\mathrm{E}\left[\ f_n(x)\right] = f(x)$$

et

$$nh_n\ \mathrm{Var}\left[\ f_n(x)\right] \to \frac{1}{2}\ f(x)$$

Remarque 2.1 *Quand $nh_n \to \infty$, l'exprèssion de la variance devient*

$$Var\left[\ f_n(x)\right] = \frac{1}{2nh_n}\ f(x)$$

Donc,

$$Var\left[\ f_n(x)\right] \to 0$$

L'erreur quadratique moyen de l'estimateur $f_n(x)$ de $f(x)$ est donné par :

$$\mathrm{E}\left[\ f_n(x) - \ f(x)\right]^2 = \mathrm{Var}\left[\ f_n(x)\right] + \left[\ \mathrm{B}ias\left\{f_n(x)\right\}\right]^2$$

Donc, si $h_n \to 0$ et $nh_n \to \infty$ quand $n \to \infty$, on a que

$$\mathrm{E}\left[\ f_n(x) - \ f(x)\right]^2 \to 0$$

pour tout point x. L'estimateur simple $f_n(x)$ est alors un estimateur consistent de $f(x)$.

2 Estimateur à noyau

L'estimateur $f_n(x)$ peut être généralisé en remplaçant la fonction de poids $\mathrm{W}(y)$ (la densité de probabilité uniforme) par une fonction de poids plus générale K (par exemple une densité de probabilité quelconque). Ceci donne le résultat qui suivre.

2.1 Définition et construction

Définition 2.1 *Soit* $(\Omega, \mathrm{A}, \mathrm{P}_\Omega)$ *un espace de probabilité. Soit* $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$ *un échantillon i.i.d. de f.d.r* F *et d'une densité* f.

L'estimateur à noyau de la fonction de densité, notée $\widehat{f}_n(x)$ *est définie par*

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{K}\left(\frac{x - \mathrm{X}_i}{h}\right)$$

Où K *est appelé fonction de poids (weight function) ou noyau (kernel function), et h est appelé paramètre de lissage (smoothing parameter) ou fenêtre (window width).*

2.2 Propriétés

Il est facile de voir que l'estimateur à noyau

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{K}\left(\frac{x - \mathrm{X}_i}{h}\right)$$

possède les propriétés suivantes :

• Les fonctions noyaux sont symétriques par rapport à l'axe O_y, $\int_{-\infty}^{+\infty} \mathrm{K}(u)\,du = 1$ et $\mathrm{K}(u) \geq 0$.

• L'estimateur par noyau est une fonction de densité.

• $\widehat{f}_n$ a les mêmes propriétés de continuité et de différentiabilité que K :

  •• Si K est continue, $\widehat{f}_n$ sera une fonction continue.

  •• Si K est différentiable, $\widehat{f}_n$ sera une fonction différentiable.

  •• Si K peut prendre des valeurs négatives,alors $\widehat{f}_n$ pourra aussi prendre des valeurs négatives.

• $\widehat{f}_n$ converge en presque complète vers f

2.3 Exemples de noyaux

Selon la définition précédent, toute fonction K peut servir comme noyau d'estimation d'une densité f. Les noyaux les plus couramment utilisés en pratique sont

– le noyau rectangulaire :

$$\mathrm{K}(u) = \frac{1}{2}\mathbf{I}_{[-1,1]}(u),$$

– le noyau triangulaire :

$$\mathrm{K}(u) = (1 - |u|)\ \mathbf{I}_{[-1,1]}(u),$$

– le noyau d'Epanechnikov :

$$\mathrm{K}(u) = \frac{3}{4}\left(1 - u^2\right)\ \mathbf{I}_{[-1,1]}(u),$$

– le noyau gaussien :

$$\mathrm{K}(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

.

Les courbes de ces noyaux sont présentées ci-dessous :

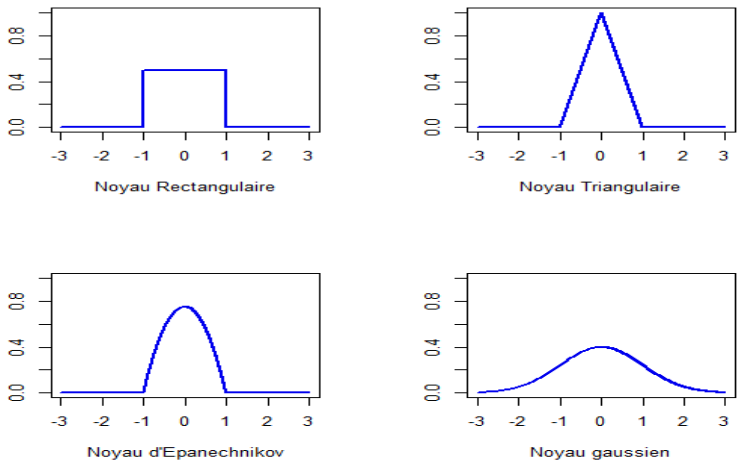


FIGURE 2.1 – La représentation graphiques de ces noyaux

2.4 Etude du biais et de la variance

Lorsqu'on définit un estimateur à noyau, on a non-seulement le choix de la fenêtre $h > 0$ mais aussi celui du noyau K. Il y a un certain nombre de conditions qui sont considérées comme usuelles pour les noyaux et qui permettent d'analyser le risque de l'estimateur à noyau qui en résulte.

HYPOTHÈSE K : On suppose que K vérifie les 4 conditions suivantes :

1. $\int_{\mathbb{R}} \mathrm{K}(u)\, du = 1$
2. K est une fonction paire ou, plus généralement, $\int_{\mathbb{R}} u\mathrm{K}(u)\, du = 0$
3. $\int_{\mathbb{R}} u^2 |\mathrm{K}(u)|\, du < \infty$
4. $\int_{\mathbb{R}} \mathrm{K}(u)^2\, du < \infty$

Proposition 2.1 *Si les trois premières conditions de l'hypothèse* K *sont remplies, alors*

$$Biais[\,\widehat{f}_n(x)] = \frac{1}{2} f''(x)\, \mu_2 h^2 + o\left(h^2\right)$$

où $\mu_2 = \int_{\mathbb{R}} \mathrm{K}(u)\, u^2 du$
Si, de plus, la condition 4 de l'hypothèse K *est satisfaite, alors*

$$Var[\,\widehat{f}_n(x)] = \frac{1}{nh} f(x)\, \mathrm{R}(x) + o\left(\frac{1}{nh}\right)$$

où $\mathrm{R}(x) = \int_{\mathbb{R}} \mathrm{K}^2(u)\, du$

Preuve. Commençons par calculer le biais :
Considérons l'estimateur à noyau

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{K}\left(\frac{x - \mathrm{X}_i}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n} \mathrm{K}_h(x - \mathrm{X}_i)$$

où nous avons introduit la notation

$$\mathrm{K}_h(\cdot) = \frac{1}{h}\mathrm{K}\left(\frac{\cdot}{h}\right)$$

pour une version transformée de K.

$$\begin{aligned}
\mathrm{E}\left[\widehat{f}_n(x)\right] &= \mathrm{E}\left[\mathrm{K}_h(x - \mathrm{X})\right] \qquad \text{car les } \mathrm{X}_i \text{ sont i.i.d.} \\
&= \int_{\mathbb{R}} \mathrm{K}_h(x-y)\, f(y)\, dy \\
&= \int_{\mathbb{R}} \mathrm{K}(u)\, f(x-uh)\, du \qquad \text{avec } u = \tfrac{x-y}{h} \text{ et } du = -\tfrac{1}{h}dy \\
&= \int_{\mathbb{R}} \mathrm{K}(u)\left[f(x) - f'(x)\,uh + \tfrac{1}{2}f''(x)\,u^2h^2 + ...\right] du \quad \text{par Taylor} \\
&= f(x)\int_{\mathbb{R}} \mathrm{K}(u)du - f'(x)\,h\int_{\mathbb{R}} \mathrm{K}(u)\,u\,du + \tfrac{1}{2}f''(x)\,h^2\int_{\mathbb{R}} \mathrm{K}(u)u^2\,du + o\left(h^2\right)
\end{aligned}$$

Alors

$$\begin{aligned}
\mathrm{B}iais\left[\widehat{f}_n(x)\right] &= \mathrm{E}\left[\widehat{f}_n(x)\right] - f(x) \\
&= \frac{1}{2}f''(x)\,h^2\int_{\mathbb{R}} \mathrm{K}(u)u^2\,du + o\left(h^2\right) \\
&= \frac{1}{2}f''(x)\,\mu_2 h^2\,du + o\left(h^2\right) \qquad (2.1)
\end{aligned}$$

Avec $\mu_2 = \int_{\mathbb{R}} \mathrm{K}(u)\,u^2 du$.
d'où la première assertion de la proposition.
Pour la variance on calcule :

$$\begin{aligned}
\mathrm{Var}\left[\widehat{f}_n(x)\right] &= \mathrm{E}\left[\widehat{f}_n^2(x)\right] - \left\{\mathrm{E}\left[\widehat{f}_n(x)\right]\right\}^2 \\
&= \tfrac{1}{n}\left\{\mathrm{E}\left[\mathrm{K}_h^2(x - \mathrm{X})\right] - (\mathrm{E}\left[\mathrm{K}_h(x - \mathrm{X})\right])^2\right\} \\
&= \tfrac{1}{nh^2}\int_{\mathbb{R}} \mathrm{K}^2\left(\tfrac{x-y}{h}\right) f(y)\, dy \\
&= \tfrac{1}{nh}\int_{\mathbb{R}} \mathrm{K}^2(u)\, f(x-uh)\, du \qquad avec \quad u = \tfrac{x-y}{h} \\
&= \tfrac{1}{nh}\int_{\mathbb{R}} \mathrm{K}^2(u)\, \left[f(x) - f'(x)\,hu + ...\right] du \\
&= \tfrac{1}{nh}\, f(x)\int_{\mathbb{R}} \mathrm{K}^2(u)\, du - f'(x)\int_{\mathbb{R}} \mathrm{K}^2(u)\,u\,du + o(1)
\end{aligned}$$

Donc nous trouvons que :

$$\begin{aligned}
\mathrm{Var}\left[\widehat{f}_n(x)\right] &= \frac{1}{nh}\, f(x)\int_{\mathbb{R}} \mathrm{K}^2(u)\, du + o\left(\frac{1}{nh}\right) \\
&= \frac{1}{nh}\, f(x)\, \mathrm{R}(x) + o\left(\frac{1}{nh}\right) \qquad (2.2)
\end{aligned}$$

avec $\mathrm{R}(x) = \int_{\mathbb{R}} \mathrm{K}^2(u)\, du$ ■

Remarque 2.2
Si $h = h_n \to 0$ *quand* $n \to \infty$, *alors*

$$\mathrm{B}iais\left[\widehat{f}_n(x)\right] \to 0$$

Si $h = h_n \to 0$ *et* $nh_n \to \infty$ *quand* $n \to \infty$, *alors*

$$Var\left[\widehat{f}_n(x)\right] \to 0$$

Conclusion 2.1

- *Si h décroît alors le* $(\mathrm{B}iais)^2 \searrow$ *et la variance* $\nearrow$
- *Si h augmente alors le* $(\mathrm{B}iais)^2 \nearrow$ *et la variance* $\searrow$

Il faut donc essayer de choisir un h qui fasse un compromis entre le $(\mathrm{B}iais)^2$ *et la variance.*

2.5 Expression d'erreur quadratique moyenne (MSE)

Les expressions asymptotiques du biais et de la variance nous premettent de trouver l'expression asymptotique pour la (MSE) et l'erreur quadratique moyenne intégrée (MISE), Ces expressions ont été obtenues sous certaines conditions sur K et en supposant que la densité de probabilité f avait toutes les dérivées (continues) nécessaires.
A partir de (2.1) et (2.2) on peut obtenir facilement l'expression suivante pour la MSE et la MISE.

$$\mathrm{MSE}\left[\widehat{f}_n(x)\right] = \frac{1}{4}h^4\ \mu_2^2\ \left(f''(x)\right)^2 + \frac{1}{nh}\ f(x)\ \mathrm{R}(x) + o\left(h^4 + \frac{1}{nh}\right) \tag{2.3}$$

$$\mathrm{MISE}\left[\widehat{f}_n(x)\right] = \frac{1}{4}h^4\ \mu_2^2 \int \left(f''(x)\right)^2 dx + \frac{1}{nh}\ f(x)\ \mathrm{R}(x) + o\left(h^4 + \frac{1}{nh}\right) \tag{2.4}$$

sous des conditions appropriées d'intégrabilité de f et ses dérivées.

On note l'approximation asymptotique de la MSE par

$$\mathrm{AMSE}\left[\widehat{f}_n(x)\right] = \frac{1}{4}h^4\ \mu_2^2\ \left(f''(x)\right)^2 + \frac{1}{nh}\ f(x)\ \mathrm{R}(x) \tag{2.5}$$

et l'approximation asymptotique de la MISE par

$$\mathrm{AMISE}\left[\widehat{f}_n(.)\right] = \frac{1}{4}h^4\ \mu_2^2\ \mathrm{R}\left(f''\right) + \frac{1}{nh}\ f(x)\ \mathrm{R}(x) \tag{2.6}$$

Chapitre 3

Estimation de la densité conditionnelle

Dans ce chapitre en premier lieu nous allons faire un rappel sur les modèles non paramétriques conditionnels : Quelques résultats théoriques de base et l'Estimation de la fonction de répartition conditionnelle.
En deuxième lieu nous allons présenter l'estimateur à noyau de la densité conditionnelle quand la variable explicative est fonctionnelle.
Il existe plusieurs estimateurs de la densité conditionnelle tels que : l'estimateur des "points les plus proches", l'estimateur "histogramme", et l'estimateur à noyau.
Pour notre travail nous nous sommes concentrés sur l'estimation par la méthode du noyau, car l'estimateur à noyau d'une densité est l'un des estimateurs les plus étudiés et les plus performants . Ce travail est basé sur les résultats de Ferraty et Vieu.

1 Présentation des modèles non paramétriques conditionnels

1.1 Type de noyau

Nous allons considérer deux sortes de noyaux : noyaux de type I et noyau de type II. La famille du noyau de type I contient les noyaux usuels discontinus, tandis que la seconde famille contient les noyaux standards continus.

Définition 3.1 *Une fonction* $\mathrm{K} : \mathbb{R} \to \mathbb{R}^*$ *telle que* $\int \mathrm{K} = 1$ *est dite noyau de type I si il existe deux constantes réelles* $0 < \mathrm{C}_1 < \mathrm{C}_2 < \infty$ *telleque*

$$\mathrm{C}_1 \mathbf{I}_{[0,1]} \leq \mathrm{K} \leq \mathrm{C}_2 \mathbf{I}_{[0,1]}$$

Définition 3.2 *Une fonction* $\mathrm{K} : \mathbb{R} \to \mathbb{R}^*$ *telle que* $\int \mathrm{K} = 1$ *est dite noyau de type II si son support est* $[0,1]$ *et si sa dérivé* K' *existe sur l'intérval* $[0,1]$ *et cette dérivé vérifie la condition suivante :*

si il existe deux constante réelles C_1 *et* C_2 *telle que* $-\infty < \mathrm{C}_2 < \mathrm{C}_1 < 0$ *alors*

$$\mathrm{C}_2 \leq \mathrm{K}' \leq \mathrm{C}_1$$

Afin de simplifier notre objectif, pour la pondération locale des variables aléatoires réelles nous allons définir le noyau suivant :

Définition 3.3 *Une fonction* $\mathrm{K} : \mathbb{R} \to \mathbb{R}^*$ *telle que* $\int \mathrm{K} = 1$ *sur un support compact* $[-1,1]$ *et* $\forall u \in (0,1), \mathrm{K}(u) > 0$ *est appelé noyau de type 0.*

1.2 Probabilité des petites boules

Soit X une v.a.f dans $\mathbb{E}$, x un élément fixe dans $\mathbb{E}$, et pour mieux fixé les idées, on utilise un noyau asymétrique simple de type I. La relation qui lie la pondération locale et la notion de probabilités petite boule est donnée comme suit :

$$\mathrm{E}\left[1_{[0,1]}\left(\frac{d(x,\mathrm{X})}{h}\right)\right] = \mathrm{E}\left[1_{\beta(x,h)}(\mathrm{X})\right] = \mathrm{P}\left(\mathrm{X} \in \beta(x,h)\right)$$

Dans la suite de notre travail nous utiliserons, pour tout x dans $\mathbb{E}$ et pour tout h réel positif, la notation suivante :

$$\varphi_x(h) = \mathrm{P}\left(\mathrm{X} \in \beta(x,h)\right)$$

1.3 Quelques résultats théoriques de base

Comme l'idée du noyau du poids local fonctionnel est le foyer de toutes les méthodes non paramétriques fonctionnelles qu'on va étudier, on utilise les deux résultats suivants :

Lemme 3.1 *Si* K *est un noyau de type I, alors il existe deux constantes réelles non négatives* C *et* C′ *telles que :*

$$\mathrm{C}\varphi_x(h) \leq \mathrm{E}\left(\mathrm{K}\left(\frac{d(x,\mathrm{X})}{h}\right)\right) \leq \mathrm{C}'\varphi_x(h) \tag{3.1}$$

Lemme 3.2 *Si* K *est un noyau de type II, et si* $\varphi_x(h)$ *satisfait*

$$\exists \mathrm{C} > 0, \exists \epsilon_0, \forall \epsilon < \epsilon_0, \int_0^{\epsilon} \varphi_x(u)\,du > \mathrm{C}\epsilon\varphi_x(\epsilon) \tag{3.2}$$

et si il existe deux constantes réelles non négatives C *et* C *alors :*

$$\mathrm{C}\varphi_x(h) \leq \mathrm{E}\left(\mathrm{K}\left(\frac{d(x,\mathrm{X})}{h}\right)\right) \leq \mathrm{C}'\varphi_x(h) \tag{3.3}$$

2 Estimation de la fonction de répartition conditionnelle

Commençons par proposer pour l'opérateur non linéaire r , définie par :

$$r(x) = \mathrm{E}\left[\mathrm{Y} \mid \mathrm{X} = x\right]$$

L'estimateur de régression du noyau fonctionnel suivant :

$$\widehat{r}(x) = \frac{\sum_{i=1}^{n} \mathrm{K}\left(h^{-1} d(x,\mathrm{X}_i)\right)\mathrm{Y}_i}{\sum_{i=1}^{n} \mathrm{K}\left(h^{-1} d(x,\mathrm{X}_i)\right)} \quad si \quad \sum_{i=1}^{n} \mathrm{K}\left(h^{-1} d(x,\mathrm{X}_i)\right) \neq 0$$

Où k est un noyau asymétrique et h (selon n) est un réel strictement positif. c'est une extension fonctionnelle de l'estimation familière de Nadaraya-Watson introduite par Ferraty et Vieu (2000).

nous concentrons maintenant sur l'estimateur $\widehat{\mathrm{F}}_{\mathrm{Y}}^{\mathrm{X}}$ de la fonction de répartition conditionnelle $\mathrm{F}_{\mathrm{Y}}^{\mathrm{X}}$, mais expliquons d'abord comment nous pouvons étendre l'idée utilisée pour la construction de l'estimateur de régression du noyau. Clairement, $\mathrm{F}_{\mathrm{Y}}^{\mathrm{X}}(x,y) = \mathrm{P}\left(\mathrm{Y} \leq y \mid \mathrm{X} = x\right)$ peut être exprimé en termes d'espérance conditionnelle :

$$\mathrm{F}_{\mathrm{Y}}^{\mathrm{X}}(x,y) = \mathrm{E}\left[\mathbf{1}_{(-\infty,y]}(\mathrm{Y}) \mid \mathrm{X} = x\right].$$

Et par analogie avec le contexte de régression fonctionnelle, un noyau naïve conditionnel c.d.f. L'estimateur pourrait être défini comme suit :

$$\widehat{\mathrm{F}}_{\mathrm{Y}}^{\mathrm{X}}(x,y)=\frac{\sum_{i=1}^{n}\mathrm{K}\left(h^{-1}d(x,\mathrm{X}_i)\right)\mathbf{1}_{(-\infty,y]}\mathrm{Y}_i}{\sum_{i=1}^{n}\mathrm{K}\left(h^{-1}d(x,\mathrm{X}_i)\right)}$$

Enfin, les idées précédemment développées par Roussas, Samanta et Ferraty et Vieu , l'estimateur de la fonction de répartition conditionnelle est donné par :

$$\widehat{\mathrm{F}}_{\mathrm{Y}}^{\mathrm{X}}(x,y)=\frac{\sum_{i=1}^{n}\mathrm{K}\left(h^{-1}d(x,\mathrm{X}_i)\right)\mathrm{H}\left(g^{-1}\left(y-\mathrm{Y}_i\right)\right)}{\sum_{i=1}^{n}\mathrm{K}\left(h^{-1}d(x,\mathrm{X}_i)\right)}\qquad\forall y\in\mathbb{R}$$

Soit K_0 un noyau symétrique habituel, soit H défini comme suit :

$$\forall u\in\mathbb{R}\quad \mathrm{H}(u)=\int_{-\infty}^{u}\mathrm{K}_0(v)\,dv$$

considérons K_0 comme un noyau de type 0
de plus, nous pouvons écrire :

$$\mathrm{H}\left(g^{-1}\left(y-\mathrm{Y}_i\right)\right)=\begin{cases}0\Longleftrightarrow y\leq \mathrm{Y}_i-g,\\ 1\Longleftrightarrow y\geq \mathrm{Y}_i+g.\end{cases}$$

3 L'estimateur à noyau de la densité conditionnelle

Soit $(\mathrm{X}_1,\mathrm{Y}_1)$, ..., $(\mathrm{X}_n,\mathrm{Y}_n)$ un échantillon aléatoire du couple (X,Y) indépendant, identiquement distribué qui est à valeurs dans $\mathscr{H}\times\mathbb{R}$, où $\mathscr{H}$ est un espace de Hilbert muni du produit scalaire $<.,.>$, et dont la norme associée est notée $||.||$.
l'estimateur à noyau de la densité conditionnelle $f(y\mid x)$ noté $\widehat{f}(y\mid x)$ est défini par :

$$\widehat{f}(y\mid x)=\frac{\sum_{i=1}^{n}\mathrm{K}\left(\frac{d(\mathrm{X}_i,x)}{h}\right)\frac{\partial}{\partial y}\mathrm{H}\left(\frac{y-\mathrm{Y}_i}{g}\right)}{\sum_{i=1}^{n}\mathrm{K}\left(\frac{d(\mathrm{X}_i,x)}{h}\right)},\quad\forall y\in\mathbb{R}\tag{3.4}$$

où H est définit par

$$\forall u\in\mathbb{R},\mathrm{H}(u)=\int_{-\infty}^{u}\mathrm{K}_0(v)\,dv$$

et

$$d(\mathrm{X}_i,x)=\|\mathrm{X}_i-x\|.$$

La fonction K est un noyau de type I ou de type II et la fonction K_0 est un noyau de type 0 et $h=h(n)(resp.g=g(n))$ est une suite de nombres réels positifs qui tend vers zéro lorsque n tend vers l'infinie. Il est aussi appelé le paramètre de lissage ou largeur de fenêtre.
Tout au long de notre travail, nous noterons par C et C′ deux constantes génériques et strictement positives. Afin d'établir la convergence presque complète (p.co.) de notre estimateur on considère les hypothèses suivantes. Soient x (resp. y) un élément de $\mathscr{H}$ (resp.de $\mathbb{R}$), $\mathbb{N}_x\subset\mathscr{H}$ un voisinage de x et S un sous ensemble compact de $\mathbb{R}$ tels que :

$$\mathrm{P}(d(\mathrm{X},x)<h)=\varphi_x(h)>0,\tag{3.5}$$

$$\exists \mathrm{C}>0,\forall(x,x')\in\mathbb{R}\times\mathbb{R},\left|\mathrm{K}_0(x)-\mathrm{K}_0(x')\right|\leq \mathrm{C}\left|x-x'\right|,$$

$$\lim_{n\to\infty}\frac{\log n}{ng\varphi_x(h)}=0 \quad \text{et} \quad \exists\alpha>0,\lim_{n\to\infty}gn^{\alpha}=\infty \tag{3.6}$$

$$\exists C>0,\exists\epsilon_0,\forall\epsilon<\epsilon_0,\int_0^{\epsilon}\varphi_x(u)du>C\epsilon\varphi_x(\epsilon), \tag{3.7}$$

$$\begin{cases} \exists C_x>0 \ \text{ tel que } \ \forall(y_1,y_2)\in S^2,\forall(x_1,x_2)\in\mathbb{N}_x\times\mathbb{N}_x, \\ \left|f(y_1\mid x_1)-f(y_2\mid x_2)\right|\le C_x\left(d^{\beta_1}(x_1,x_2)+\left|y_1-y_2\right|^{\beta_2}\right),\beta_1>0,\beta_2>0. \end{cases} \tag{3.8}$$

Théorème 3.1 *Sous les conditions (3.4), (3.5), (3.6), (3,7) et (3.8), nous avons pour tout nombre réel fixé y :*

$$\widehat{f}(y\mid x)-f(y\mid x)=0\left(h^{\beta_1}\right)+0\left(g^{\beta_2}\right)+0_{p.co.}\left(\sqrt{\frac{\log n}{ng\varphi_x(h)}}\right) \tag{3.9}$$

Preuve. La preuve est basée sur la décomposition qui suit

$$\widehat{f}(y\mid x)-f(y\mid x)=\frac{\left(\widehat{r}_3(x,y)-E\widehat{r}_3(x,y)\right)-\left(f(y\mid x)-E\widehat{r}_3(x,y)\right)}{\widehat{r}_1(x)}-\frac{f(y\mid x)}{\widehat{r}_1(x)}\left[\widehat{r}_1(x)-1\right] \tag{3.10}$$

où $\widehat{r}_1$ est définit par

$$\widehat{r}_1(x)=\frac{1}{n}\sum_{i=1}^{n}\Delta_i \tag{3.11}$$

avec

$$\Delta_i=\frac{K\left(\frac{d(x,X_i)}{h}\right)}{E\left(K\left(\frac{d(x,X_i)}{h}\right)\right)}$$

et où

$$\widehat{r}_3(x,y)=\widehat{r}_1(x)\,\widehat{f}(y\mid x)=\frac{1}{n}\sum_{i=1}^{n}\Delta_i\Gamma_i(y) \tag{3.12}$$

avec

$$\Gamma_i(y)=\frac{1}{g}K_0\left(\frac{y-Y_i}{g}\right) \tag{3.13}$$

Ainsi, la preuve est une conséquence directe des résultats qui suivent.

Lemme 3.3 *Sous les hypothèses (3.5) et (3.8), lorsque n tend vers l'infini, nous avons :*

$$E\widehat{r}_3(x,y)-f(y\mid x)=0\left(h^{\beta_1}\right)+0\left(g^{\beta_2}\right). \tag{3.14}$$

Preuve. Puisque $E\Delta_i=1$ et puisque K_0 est une fonction intégrable, nous avons :

$$\begin{aligned}
E\widehat{r}_3(x,y)-f(y\mid x) &= E\Delta_1\Gamma_1(y)-f(y\mid x)\\
&= E\left(\Delta_1\left(E\left(\Gamma_1(y)\mid X\right)-f(y\mid x)\right)\right)\\
&= E\left(\Delta_1\int_{\mathbb{R}}\frac{1}{g}K_0\left(\frac{y-u}{g}\right)f(u\mid X)du-f(y\mid x)\right)\\
&= E\left(\Delta_1\int_{\mathbb{R}}\frac{1}{g}K_0\left(\frac{y-u}{g}\right)\left(f(u\mid X)-f(y\mid x)\right)du\right)\\
&= E\left(\Delta_1\int_{\mathbb{R}}\frac{1}{g}K_0(v)\left(f(y-vg\mid X)-f(y\mid x)\right)dv\right)\\
&= E\left(1_{\beta(x,h)}\Delta_1\int_{\mathbb{R}}\frac{1}{g}K_0(v)\left(f(y-vg\mid X)-f(y\mid x)\right)dv\right)
\end{aligned} \tag{3.15}$$

Compte tenu du fait que le support $K_0 = [-1,1]$ et puisque h et g tendent vers zéro, la condition de Hölder (3.8) permet d'écrire que :

$$\sup_{v\in[-1,1]} \left| f(y-vg \mid X) - f(y \mid x) \right| = 0\left(h^{\beta_1}\right) + \left| 0\left(g^{\beta_2}\right)\right|^{\beta_2} \tag{3.16}$$

Ainsi, le résultat (3,14) découle directement en combinant (3,15), (3,16) et le fait que $E\Delta_1 = 1$. ■

Lemme 3.4 *Sous les hypothèses du théorème, lorsque n tends vers l'infini, nous avons :*

$$\widehat{r}_3(x,y) - E\widehat{r}_3(x,y) = 0_{p.co.}\left(\sqrt{\frac{\log n}{ng\varphi_x(h)}}\right) \tag{3.17}$$

Preuve. Pour cela, nous utilisons la décomposition suivante :

$$\widehat{r}_3(x,y) - E\widehat{r}_3(x,y) = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{3.18}$$

Où

$$Z_i = (T_i - ET_i) \qquad et \qquad T_i = \Delta_i \Gamma_i(y). \tag{3.19}$$

Afin d'appliquer une inégalité de type-Bernstein, nous commençons par montrer

$$|T_i| \leq \frac{C}{g\varphi_x(h)} \qquad et \qquad ET_i^2 \leq \frac{C}{g\varphi_x(h)} \tag{3.20}$$

En utilisant le Lemme 3.1 ou le Lemme 3.2 et en tenant compte de l'hypothèse (3.7) et du fait que K est de type I ou de type II, nous avons :

$$C\varphi_x(h) \leq E\left(K\left(\frac{d(X,x)}{h}\right)\right) \leq C'\varphi_x(h) \tag{3.21}$$

et en utilisant le dernier résultat et puisque K_0 est borné, nous obtenons

$$|T_i| \leq \frac{C}{g\varphi_x(h)}$$

Le second moment des variables T_i peut être calculé en utilisant l'intégration par changement de variable :

$$\begin{aligned} ET_i^2 = E\left(\Delta_i^2 \frac{1}{g^2} K_0^2\left(\frac{y-Y_i}{g}\right)\right) &= E\left(E\left(\Delta_i^2 \frac{1}{g^2} K_0^2\left(\frac{y-Y_i}{g}\right) \mid X = x\right)\right) \\ &= \frac{1}{g^2} E\left(\Delta_i^2 \int_{\mathbb{R}} K_0^2\left(\frac{y-u}{g}\right) f(u \mid X)\, du\right) \\ &= \frac{g}{g^2} E\left(\Delta_i^2 \int_{\mathbb{R}} K_0^2(v)\, f(y-vg \mid X)\, dv\right) \\ &= \frac{1}{g} E\Delta_i^2. \end{aligned} \tag{3.22}$$

Puisque $0 < \int K^2 < \infty$, si K est de type I (resp.II) alors $\frac{K^2}{\int K^2}$ est aussi de type I (resp.II). Ainsi, en appliquant le Lemme 3.1 ou le Lemme 3.2 on trouve

$$C\varphi_x(h) \leq E\left(K^2\left(\frac{d(X_i,x)}{h}\right)\right) \leq C'\varphi_x(h) \tag{3.23}$$

et en utilisant le dernier résultat, on écrit que

$$\frac{C}{\varphi_x(h)} \leq E\Delta_i^2 \leq \frac{C'}{\varphi_x(h)} \tag{3.24}$$

ce qui implique que

$$ET_i^2 \leq \frac{C}{g\varphi_x(h)} \tag{3.25}$$

En tenant compte de (3.20), on peut appliquer l'inégalité de type-Bernstein-type donné par le corollaire A.9 (voir Ferraty [2]), et on obtient :

$$\forall \epsilon \geq 0,\ P\left[\left|\widehat{r}_3(x,y) - E\widehat{r}_3(x,y)\right| > \epsilon\right] \leq 2\exp\frac{\epsilon^2 ng\varphi_x(h)}{2C'(1+\epsilon)} \tag{3.26}$$

Puisque la suite $\frac{\log n}{ng\varphi_x(h)}$ tend vers zéro, en choisissant $\epsilon = \epsilon_0\sqrt{\frac{\log n}{ng\varphi_x(h)}}$ dans le résultat (3.26) nous obtenons directement

$$\begin{aligned} P\left[\left|\widehat{r}_3(x,y) - E\widehat{r}_3(x,y)\right| > \epsilon_0\sqrt{\frac{\log n}{ng\varphi_x(h)}}\right] &\leq 2\exp\frac{\epsilon_0^2 \log n}{2C'\left(1+\epsilon_0\sqrt{\frac{\log n}{ng\varphi_x(h)}}\right)} \\ &\leq 2n^{-C\epsilon_0^2}, \end{aligned}$$

et il en résulte que pour ϵ_0 assez large $\left(\epsilon_0 > \frac{1}{\sqrt{C}}\right)$:

$$\sum_{i=1}^{n} P\left[\left|\widehat{r}_3(x,y) - E\widehat{r}_3(x,y)\right| > \epsilon_0\sqrt{\frac{\log n}{ng\varphi_x(h)}}\right] < +\infty$$

Ainsi, la preuve de (3.17) est maintenant achevée. ■

Lemme 3.5 *Sous les hypothèses du théorème, lorsque n tend vers l'infini, nous avons*

$$\widehat{r}_1(x) - 1 = 0_{p.co.}\left(\sqrt{\frac{\log n}{n\varphi_x(h)}}\right) \tag{3.27}$$

Preuve. Ce résultat peut directement se déduire du lemme 3.4 en prenant $\Gamma_i(y) = 1$. ■

Lemme 3.6 *Sous les hypothèses du théorème, lorsque n tend vers l'infini, nous avons*

$$\widehat{r}_1(x) \nrightarrow 0_{p.co.}$$

Preuve. Notons que les dénominateurs introduits dans la décomposition (2.9) sont directement traités en utilisant les lemmes ci-dessus et la proposition A.6. (voir Ferraty [2]). ■

■

Chapitre 4

Simulation

Nous présentons dans ce chapitre le travail de simulation effectué pour étayer les différents aspects théoriques abordés dans notre étude. L'expérimentation numérique nous servira en particulier à :

— Étudier la performance de la méthode du noyau.

— Étudier l'influence de la taille de l'échantillon sur les résultats.

— Nous changeons le noyau et étudions les résultats obtenus pour chaque changement.

1 Présentation du logiciel R

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. **R** a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets.
Il a été créé, en 1996, par Robert Gentleman et Ross Ihaka du département de statistique de l'Université d'Auckland en Nouvelle Zélande.
Le logiciel **R** est disponible sur le site

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

Il existe des versions

- Windows
- MacOS
- Linux.

Outils disponibles :

- un langage de programmation orienté objet
- des fonctions de "base"
- des librairies complémentaires (1800 sur le site CRAN)

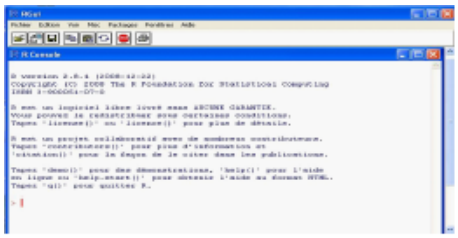


FIGURE 4.1 – Démarrage de R pour Windows

Objets :

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

1. Fonctions
2. Vecteurs, Matrices, etc
3. Listes : C'est une structure qui regroupe des objets (pas nécessairement de même type).
4. Boucles et calculs vectoriels
5. Graphiques

2 Plan de simulation

Nous nous contenterons de faire des simulations et d'observer le comportement asymptotique de l'estimateur à noyau calculé à partir d'échantillons simulés. Ceci nous permettra de savoir si l'estimateur f_n converge vers f.
Nous utiliserons pour les simulations, des échantillons de lois connues de taille de plus en plus grande n = (100, 500, 1000 et 5000).
Nous prenons l'exemple où la densité suit une loi normale $\mathcal{N}(0,1)$

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

Le graphique de $f(x)$ est représenté dans la figure suivante :

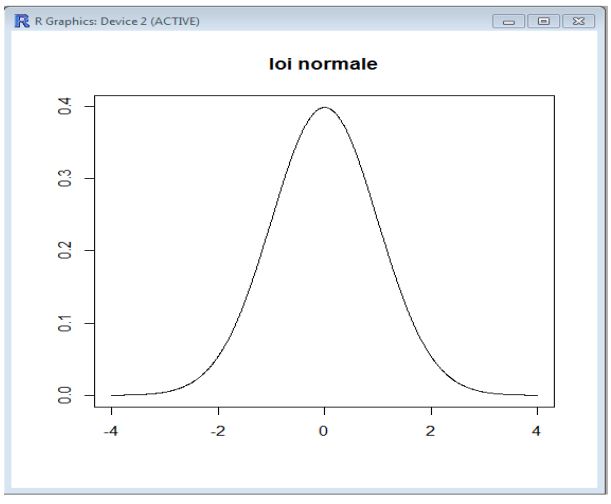


FIGURE 4.2 – Représentation graphique $f(x)$

3 Algorithme de simulation

L'algorithme de simulation que nous avons utilisé comporte quatre phases :

- Simuler un échantillon de taille n.
- Calculer le paramètre de lissage h qu'on fait varier sur un intervalle $[0,1]$ et qui minimise $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f}_n(x) - f(x)\right)^2$.
- Construire l'estimateur par la méthode du noyau à partir des observations. Le choix du noyau n'a pas d'impact très significatif sur la qualité d'estimation, dans le sens où la fenêtre est bien choisie.
- Tracer les deux courbes : la densité théorique f et la densité estimée $\widehat{f}_n$.
- Utiliser d'autres noyaux et tracer les deux courbes.

.
Les simulations et les graphes ont été réalisés à l'aide du logiciel **R**. Nous avons utilisé la version 3.0.0 pour la programmation.

4 Simulations et Résultats

On donne les résultats de simulation sous forme de tableaux et de graphes.
Les graphes ci-dessous représentent les courbes X_i pour $n = 100, 500, 1000$ et 5000.

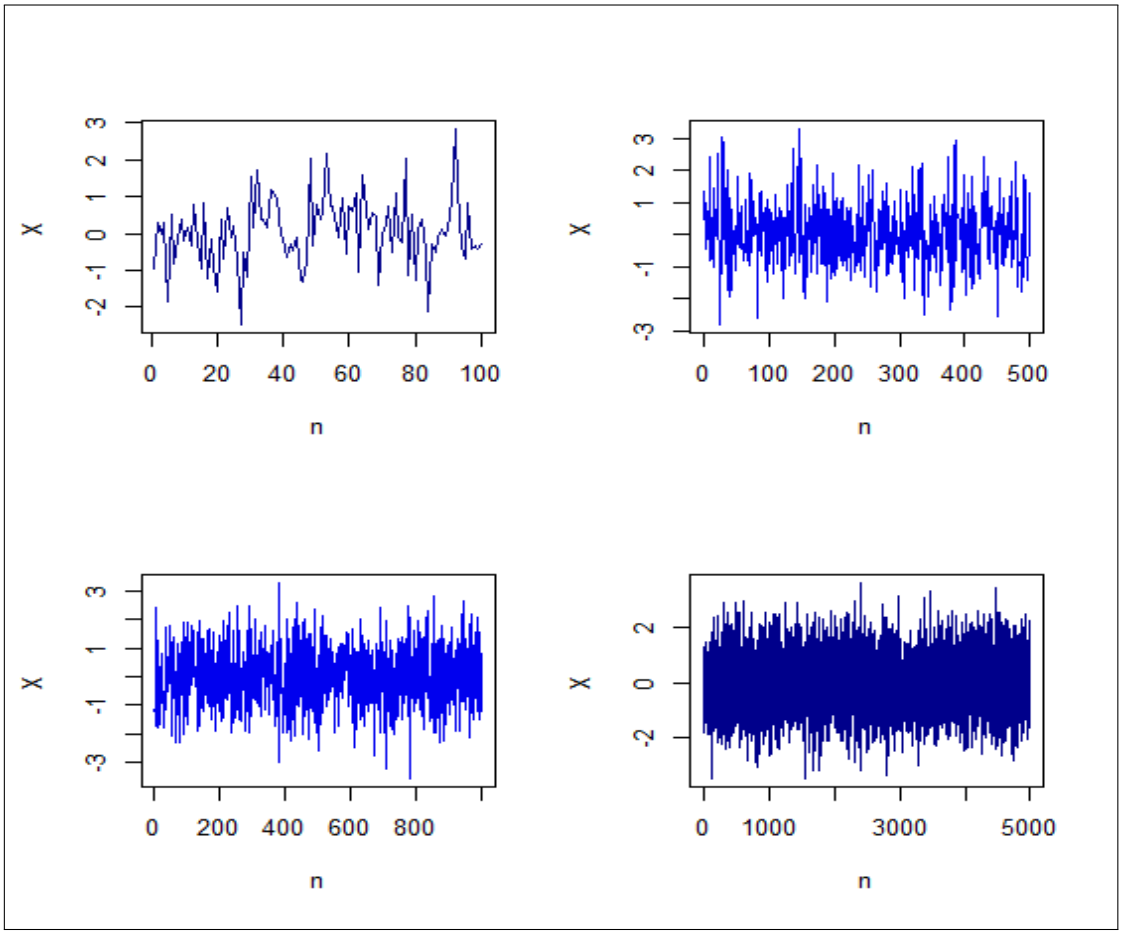


FIGURE 4.3 – Représentation des courbes X_i

L'exemple suivant permet de simuler l'estimation de la densité $f(x)$ pour un échantillon de taille 100.Nous estimons cette densité avec le noyau gaussien, en utilisant trois fenêtres différentes :

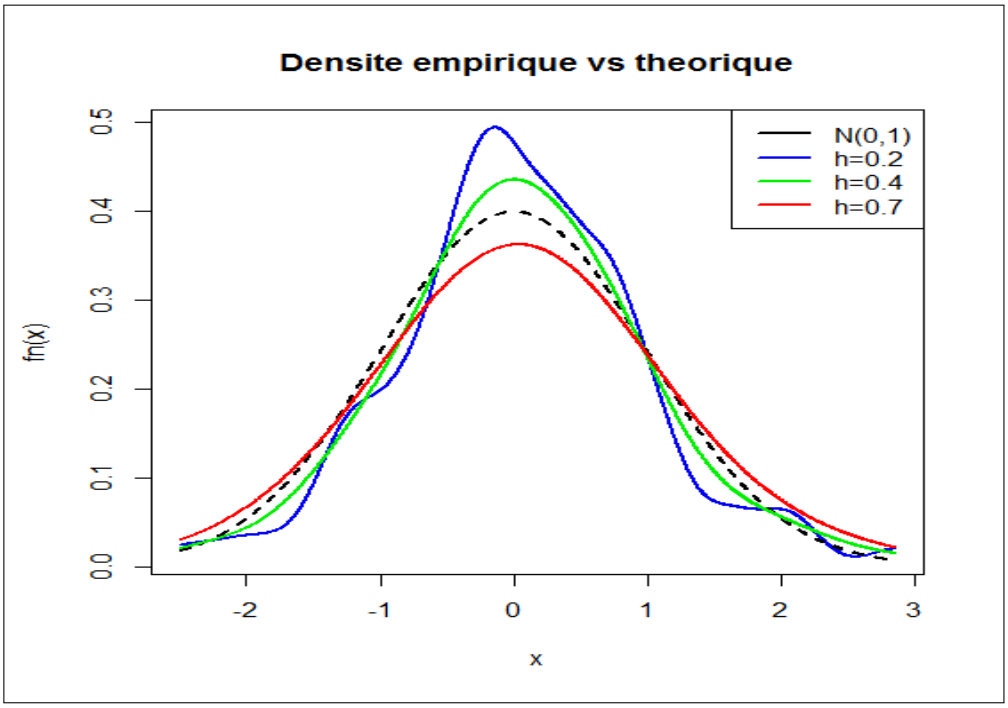


FIGURE 4.4 – Estimations par noyau Gaussien

On constate que la densité estimée est d'autant plus lisse que la fenêtre est dans les limites du 0.4. Pour avoir une bonne estimation par la méthode des noyaux, il faut bien choisir le paramètre de lissage h puisque celui-ci a un rôle crucial dans le processus.

Choix de fenêtre optimale

Les méthodes existantes pour le choix de h peuvent être classées en deux catégories :

La première catégorie est constitué des méthodes purement théoriques qui sont basées sur la minimisation de l'erreur quadratique moyenne intégrée (MISE). En effet, la valeur idéale théorique de h notée h_{id} s'obtient en minimisant le MISE asymptotique donné en (2.6). Ainsi, pour un échantillon de taille n donné et pour un noyau (classique) K fixé, cette valeur idéale de h est donnée par

$$h_{id} = \frac{1}{n^{1/5}} \left\{ \frac{\int_{\mathbb{R}} \mathrm{K}^2(t)\,dt}{\sigma^4 \int_{\mathbb{R}} \left(f''\right)^2(x)\,dx} \right\}^{1/5} \tag{4.1}$$

Ce paramètre de lissage idéal h_{id} obtenu n'est pas directement utilisable puisqu'il dépend encore de la quantité inconnue $\left(f''\right)^2(x)$.

La deuxième catégorie est celle dite des méthodes pratiques, ous allons décrire une de ces méthodes pratiques à savoir méthode de ré-injection ("Plug-in" en anglais).

Méthode Plug-in

Il s'agit ici d'estimer la quantité $\int_{\mathbb{R}} \left(f''\right)^2(x)\,dx$ dans l'expression de h_{id} donnée en (4.1). Plusieurs approches ont été proposées dans la littérature mais nous en retenons

une approche consiste à supposer que f appartienne à une famille gaussienne centrée et de variance σ^2 on trouve :

$$\int_{\mathbb{R}} \left(f''\right)^2 (x)\, dx = \frac{3}{8\sqrt{\pi}}\sigma^{1/5} \approx 0.212\sigma^{1/5}.$$

La valeur optimale de h notée hopt est obtenue en remplaçant σ dans l'expression de h_{id} par son estimateur $\widehat{\sigma} = \sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 / (n-1)}$ avec $\overline{X} = X_i / n$.Ce qui conduit à :

$$h_{opt} = 1.06\left(\frac{\widehat{\sigma}}{n^{1/5}}\right)$$

Cette approche donne de bons résultats lorsque la population est réellement normalement distribuée.

Alors, dans l'exemple précédent $\mathbf{h_{opt}} = \mathbf{0.39}$. Le graphe ci-dessous représente la densité $f(x)$ estimée et on la compare avec la densité théorique, en utilisant la fenêtre optimale.

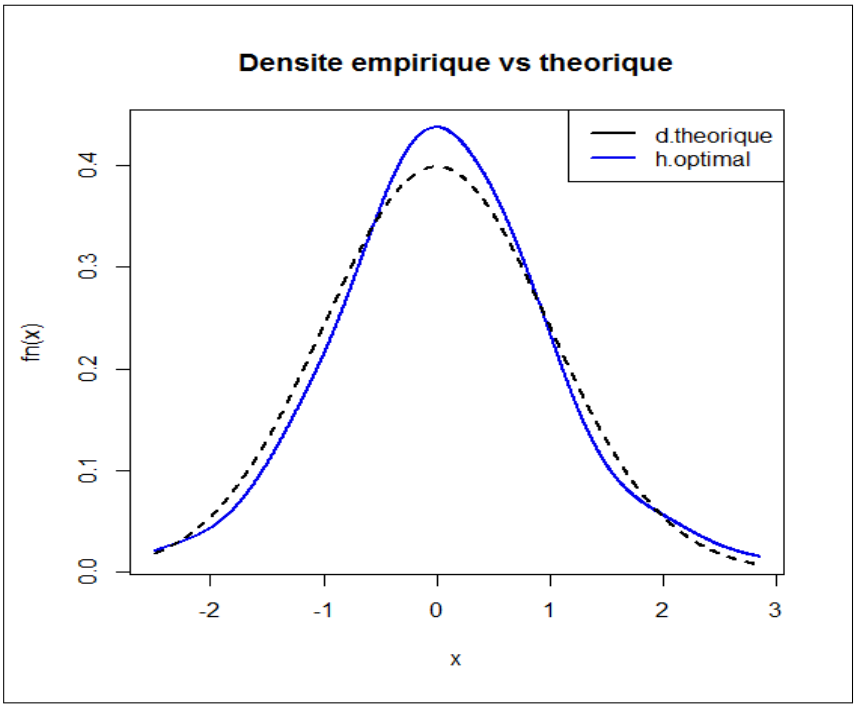


FIGURE 4.5 – Densité théorique et empirique.

Les résultats de la simulation sont donnés dans le tableau ci-dessous

	n=100	**n=500**	**n=1000**	**n=5000**
$\mathbf{h_{opt}}$	0.39	0.31	0.26	0.19
MSE	0.00074	0.00063	0.00026	0.00006

TABLEAU 4.1 – Résultats de la simulation.

Dans les graphes ci-dessous on représente la densité estimée et on la compare avec la densité théorique pour tout n=100,500,1000 et 5000.

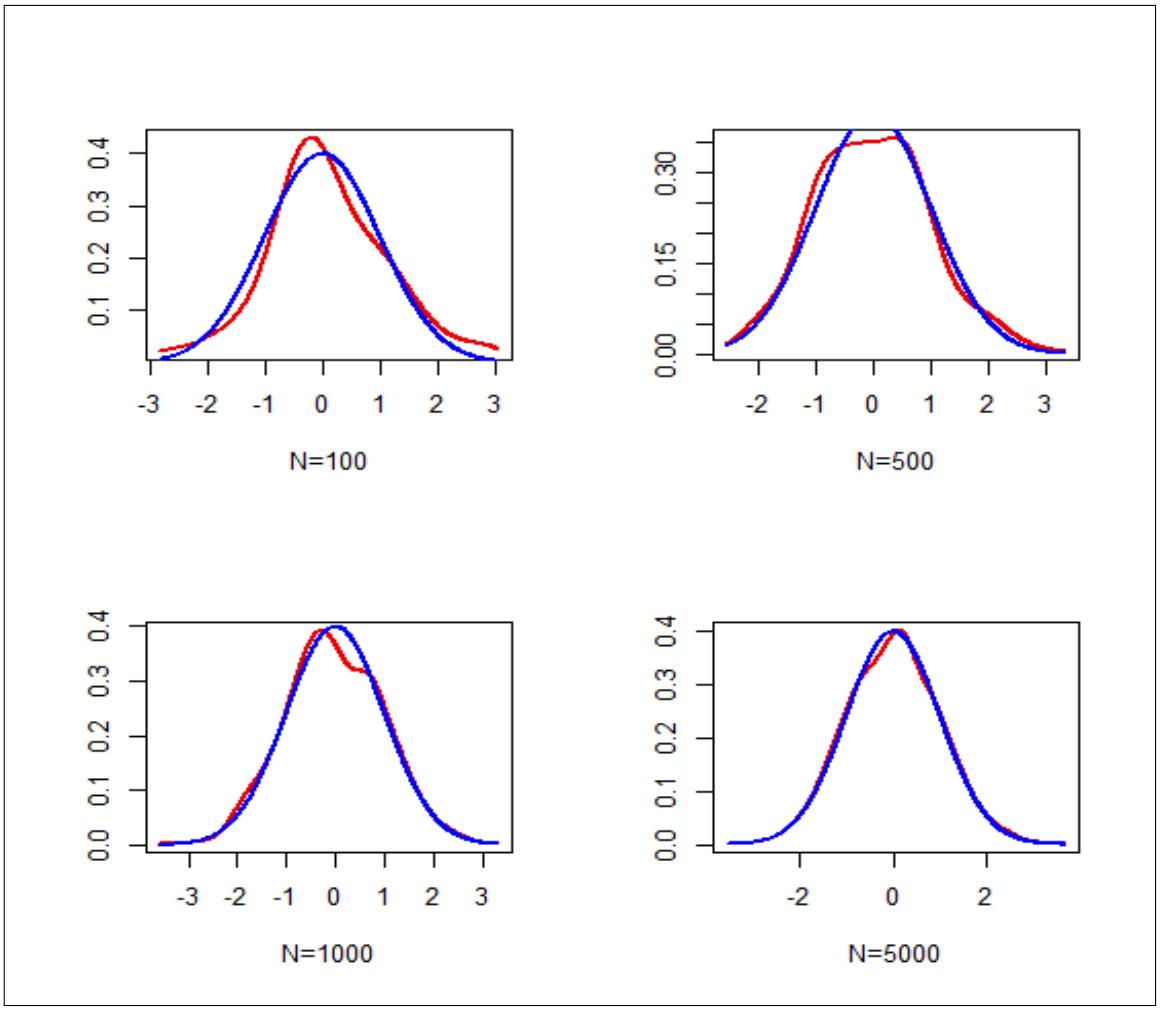


FIGURE 4.6 – Représentation de la densité estimée avec la méthode du noyau

Dans ce dernier exemple, nous estimons la fonction précédent f(x) avec le noyau gaussien, Uniforme, Triangulaire et le noyau d'Epanechnikov en utilisant la fenêtre optimale $\mathbf{h_{opt}}$=0.1922553 et pour n=5000.

Les résultats sont donnés dans le tableau suivant :

	N.Uniforme	**N.Triangulaire**	**N.Epanechnikov**	**N.Gaussien**
MSE	0.000048	0.000059	0.000054	0.000023

TABLEAU 4.2 – Résultats de la simulation par noyau.

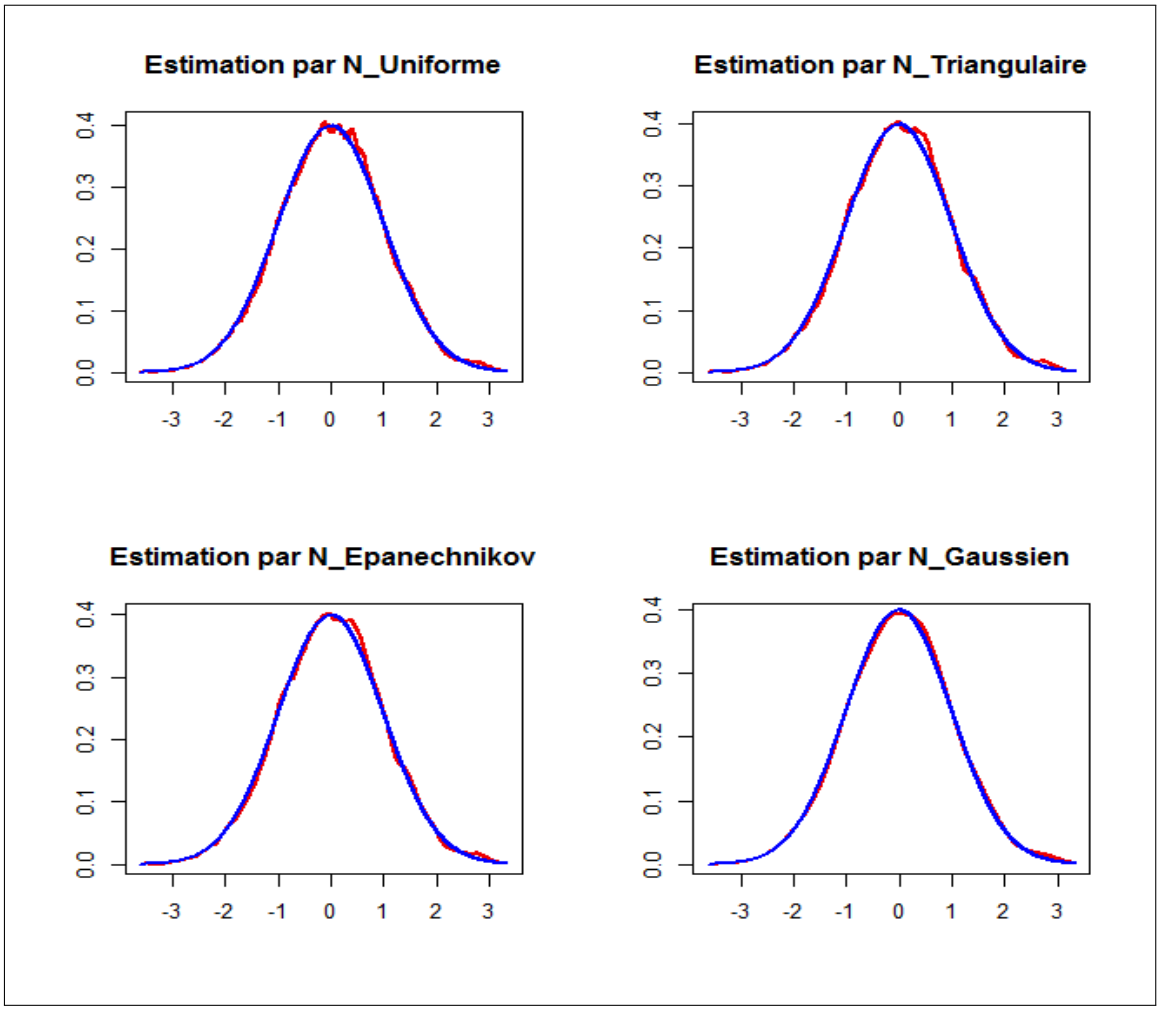


FIGURE 4.7 – Représentation de la densité estimée avec autres noyaux

5 Interprétation des résultats

D'après les graphes et les résultats numériques obtenus de la simulation, on remarque que :

— plus la taille de l'échantillon est grande plus h est petit.

— plus la taille de l'échantillon est grande, plus MSE est petit et meilleure qualité d'estimation.

— l'estimation avec le noyau gaussien est beaucoup plus lise et régulière.

Conclusion

D'après les résultats précédent, on a conclut que :

- Si h décroît alors le $(\mathrm{B}iais)^2 \searrow$ et la variance $\nearrow$
- Si h augmente alors le $(\mathrm{B}iais)^2 \nearrow$ et la variance $\searrow$

Il faut donc essayer de choisir un h qui fasse un compromis entre le $(\mathrm{B}iais)^2$ et la variance.

Compte tenu des résultats des simulations, Le choix de paramètre de lissage est crucial dans le comportement asymptotique de l'estimateur aussi le choix du noyau, nous avons alors sélectionné la valeur de paramètre de lissage h qui fournisse le plus petit MSE (erreur quadratique moyenne) nous permettant d'obtenir le meilleur estimateur de la fonction de densité.

Par ailleurs nous avons montré que la méthode du noyau est un outil très efficace pour estimer la densité.

Bibliographie

[1] **Berchtold André.**(2002-2003). syllabus STAT 2413-Chapitre.3, pp. 32-45.

[2] **FERRATY F, VIEU P** , (2005). *Nonparametric Modelling for Functional Data*, Appli. Math., pp. 232-235. 29

[3] **Imen Ben Khalifa**.Projet (2007).Estimation non-parametrique par noyaux associes et donnees de panel en marketing. *URL : http ://www.academia.edu/1076492.*

[4] **VINCENT G.** : 2012,Introduction à la programmation en R,école d'actuariat, Université Laval,151.

[5] **Francial Giscard, Baudin LIBENGUÉ DOBÉLÉ-KPOKA.** :*Méthode non-paramétrique des noyaux associés mixtes et applications*, Franche-Comté, 2003.

[6] **Anne PHILIPPE**.Journées académiques 2009 de l'IREM des Pays de la Loire Nantes,.Le logiciel R. *URL : http ://www.math.sciences.univ-nantes.fr/philippe.*

[7] **Laksaci, F. Madani, M. Rachdi.** : (2013) Kernel conditional density estimation when the regressor is valued in a semi-metric space. Communications Statistics Theory and Methods.

[8] **Lafaye de Micheaux P., Drouilhet R., Liquet B** : (2011). Le logiciel R : Maîtriser le langage, Effectuer des analyses statistiques. Springer-Verlag, France.

[9] **Laksaci, A.** : . (2007). Convergence en moyenne quadratique de l'estimateur à noyau de la densité conditionnelle avec variable explicative fonctionnelle. Ann. I.S.U.P., 51(3) :69–80 (2008).

[10] **E. A. Nadaraya.** : On Estimating Regression, Teor. Veroyatnost. i Primenen., 1964, Volume 9, Issue 1, 157–159.

[11] **Murray Rosenblatt.** : Remarks on Some Nonparametric Estimates of a Density Function, Ann. Math. Statist. Volume 27, Number 3 (1956), 832-837.

[12] **Samanta, M.** : Non-parametric estimation of conditional quantiles. Statist. Proba. Letters, 7, 407-412 (1989).

[13] **Roussas, G.** :Exponential probability inequalities with some applications. In : Statistics, probability and game theory. IMS Lecture Notes Monogr. Ser., Inst. Math. Statist., Hayward, CA. 30, 303-319 (1996).