République Algérienne démocratique et populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de Béjaia

Faculté des sciences exactes

Département de Recherche Opérationnelle

Mémoire de fin de cycle

En vue d'obtention du diplôme de master en Mathématiques Appliquées

Spécialité : Modèlisation mathématique et évaluation des performances des réseaux

Estimation de la fonction de régression par la méthode des noyaux asymétriques

Élaboré par : Sabrina MOUSLI

Membres du jury :

Président :	N.BERNINE	MCB	Université de Béjaia
Promoteur :	L.DJERROUD	MCA	Université de Béjaia
Examinateur :	Y.ZIANE	MCA	Université de Béjaia
Examinateur :	S.AMROUN	MCB	Université de Béjaia
Invité :	S.ADJABI	Professeur	Université de Béjaia

Promotion : 2021/2022

REMERCIEMENTS

Au nom d'Allah, le Tout Miséricordieux le Très Miséricordieux Louange à Allah, Seigneur de l'univers. Qu'Allah fasse l'éloge du meilleur des messagers, notre Prophète Mohammed, ainsi que de sa famille, de ses Compagnons et de ceux qui les auront suivis dans un bon comportement et qu'Il leur accorde le salut.
Tout d'abord, toute louange parfaite est à Allah le Trés Haut mon créateur de m'avoir facilité ce mémoire de m'avoir accordé la santé, la volonté, la force, le courage, la sicence, le savoir , ... je ne saurais dénombrer ses bienfaits.

Ensuite, je remercie mes parents qui m'ont soutenu tout au long de mon parcour universitaire et qui ont toujours été à mes coté.

Je remercie ma promotrice Dr. L.DJERROUD d'avoir encadré ce travail je la remercie pour ses conseils et ses recommandations.

J'adresse mes remerciement aux membres du jury Dr. Y.ZIANE, DR. S.AMROUN, DR. N.BERNINE et Pr.ADJABI d'avoir accpeté de juger ce travail.

Je remercie tous les enseignants et enseignates du départment de recherche opérationnelle de m'avoir suivi durant tout mon cursus.

Enfin, je remercie toute personnes ayant contribué de prés ou de loin à ce travail, je remercie particulièrement Dr D.BENZIANE et mes amies : Hafsa, Manel et Fella.

TABLE DES MATIÈRES

Remerciements **1**

Liste des tableaux **iv**

Liste des figures **v**

Introduction générale **1**

1 Régression non paramétrique **4**

1.1 Introduction 4

1.2 Modèle de régression non paramétrique 4

1.3 Méthode du noyau pour l'estimation de la fonction de régression 5

1.3.1 Présentation de la méthode 5

1.3.2 Noyau associé 6

1.3.3 Estimateur à noyau symétrique de la fonction de régression . . . 6

1.3.4 Propriétés de l'estimateur 9

1.3.4.1 Biais de l'estimateur 9

1.3.4.2 Variance de l'estimateur 10

1.3.4.3 L'erreur moyenne quadratique 10

1.3.4.4 L'erreur quadratique moyenne intégrée 11

1.3.5 Exemples de noyaux continus symétriques 11

1.3.6 Choix de la fenêtre de lissage 13

1.3.6.1 Méthode de minimisation du AMSE 13

1.3.6.2 Méthode de minimisation du AMISE 14
1.3.6.3 Méthode de la validation croisée 14
1.3.6.4 Méthode de la validation croisée généralisée 15
1.4 Conclusion 16

2 Estimateur asymétrique de la fonction de régression **17**
2.1 Introduction 17
2.2 Noyau associé continu asymétrique 18
2.3 Exemples de noyaux associés asymétriques 18
2.4 Estimateur à noyau asymétrique de la fonction de régression 21
2.5 Résultats sur les propriétés de l'estimateur à noyau gamma 21
2.5.1 Rappels sur la loi gamma 21
2.5.2 Présentation du noyau 22
2.5.3 Vérification des conditions d'un noyau associé 22
2.5.4 Estimateur à noyau gamma de la fonction de régression 23
2.5.5 Propriétés de l'estimateur 23
2.5.5.1 Biais de l'estimateur 24
2.5.5.2 Variance de l'estimateur 27
2.5.5.3 Erreur moyenne quadratique 28
2.6 Propositions sur les propriétés de l'estimateur à noyau bêta 29
2.6.1 Rappel sur la loi bêta 29
2.6.2 Présentation du noyau associé bêta 29
2.6.3 Vérification des conditions d'un noyau associé 30
2.6.4 Estimateur de la fonction de régression à noyau associé Bêta . . . 30
2.6.5 Propriétés de l'estimateur 30
2.6.5.1 Biais de l'estimateur 31
2.6.5.2 Variance de l'estimateur 32
2.7 Conclusion 34

3 Évaluation de performance de l'estimateur : données simulées et données réelles **35**
3.1 Introduction 35
3.2 Algorithme de simulation 35

3.3 Étude sur des modèles cibles 36
3.4 Application sur des données réelles 40
3.5 Conclusion 43

Conclusion générale **45**

Annexe **47**

LISTE DES TABLEAUX

2.1 Cas particulier de noyaux GBS 20

3.1 Valeurs des risques RQ_{moy} du modèle $m_1(x)$ 37
3.2 Valeurs des risques RQ_{moy} du modèle m_2 39
3.3 Données financières collectées par le fond monétaire international de 1974 à 2015 41

TABLE DES FIGURES

1.1 Exemples de noyaux symétriques 12

2.1 Exemple de noyaux associés assymétriques avec $x = h = 0.5$ 19
2.2 Noyaux BS, BS−PE et BS−t avec $x = 1$ et $h = 0.03$ 20
2.3 Noyau gamma pour une cible $x = 0.1$ 22
2.4 Noyau associé bêta $K_{(\frac{x}{h}+1, \frac{1-x}{h}+1)}(.)$ avec x=0 30

3.1 Estimation du modèle $m_1(x)$ par noyaux associés asymétrique : beta, gamma 38
3.2 Estimation du modèle $m_1(x)$ par noyaux associés asymétrique : lognormal et BS 38
3.3 Estimation du modèle $m_2(x)$ par noyaux associés asymétriques : bêta, gamma 40
3.4 Estimation du modèle $m_2(x)$ par noyaux associés asymétriques : lognormal et BS 40
3.5 Régression non paramétrique en utilisant le noyau beta 43

INTRODUTION GÉNÉRALE

En statistique, lors d'une étude effectuée sur une population donnée et selon un caractère précis X ou Y, nous modélisons la situation sous forme d'un modèle statistique qui représente une description mathématique approximative de celle-ci. Dans certaines études, l'objectif est d'établir la relation entre deux variables aléatoires (v.a) X et Y associées respectivement à deux caractères distincts. Afin de cristalliser cette relation nous utilisons le modèle statistique appelé "**régression**".

La régression est un procédé permettant sur la base d'un échantillon d'observations l'analyse d'une relation entre deux v.a (Y : la variable expliquée et X : la variable explicative). Ce procédé constitue méthodes d'analyse statistique permettant la mesure d'une variable à travers une autre variable dans le cas de la régression univariée ou plusieurs variables dans le cas de la régression multi-varié qui lui sont corrélées. La donnée de la fonction de régression constitue alors un élément important pour la caractérisation de ce type de relation. Cependant, celle-ci n'est pas toujours définie d'une manière explicite alors dans ce cas nous procédons à son estimation.

Nous distinguons principalement trois approches d'estimation : l'estimation paramétrique, l'estimation non paramétrique et l'estimation semi-paramétrique. L'estimation paramétrique de la fonction de régression consiste à estimer un ensemble de paramètres caractérisant cette fonction. Parmi les modèles paramétriques de régression nous citons : les modèles linéaires, les modèles linéaires généralisés et les modèles non linéaires. Notons que les méthodes de cette première approche présentent des restrictions au niveau de leurs applications ; par exemple le cas où l'on ne possède pas suffisamment d'informations sur la fonctions estimée i.e (c'est à dire) si on ne connaît

pas le type de relation entre la variable expliquée et la variable explicative. Dans ce cas on fait appel aux méthodes de la deuxième approche dite "**non paramétrique**". Cette dernière consiste à estimer la fonction de régression sur la base d'observations d'un échantillon de la v.a et ceci sans effectuer de suppositions sur la forme de cette fonction considérée. La troisième approche en l'occurrence **estimation semi-paramétrique** combine les deux premières approches, ici nous supposons que la fonction de régression dispose de deux composantes l'une paramétrique, l'autre non paramétrique.

Parmi les méthodes non paramétriques d'estimation d'une densité ou d'une fonction de régression nous pouvons citer : la méthode d'histogramme [1], la méthode d'estimation par les séries orthogonales [2] et la méthode de lissage par les fonction splines [3]. Une autre méthode figurant dans cette catégorie d'estimateur est la méthode du "**noyau**". Cette dernière se présente sous forme d'une somme de n v.a indépendantes et elle est caractérisée à travers la donnée de ses deux composantes essentielles à savoir "**le noyau**" K et le "**paramètre de lissage**" h. Pour la méthode du noyau, un estimateur de la densité de probabilité a été proposé par Rosenblatt [4] et Parzen [5] qui a été repris séparément dans le contexte de la fonction de régression par Nadaraya [6] et Watson [7]. Celui-ci n'est pas le seul estimateur qui utilise la méthode du noyau [8]. En effet, Priestley–Chao [9] ont proposé un estimateur de la fonction de régression obtenu à travers une modification de l'estimateur de Nadaraya [6] et Watson [7].

Comme mentionné précédemment, le noyau K est une composante essentielle de la méthode du noyau. Celui-ci peut être continu ou discret et peut être symétrique ou asymétrique et ceci en fonction de l'ensemble des observations.

Les caractéristiques d'un estimateur sont considérées comme étant plus importantes que l'estimateur lui même. Dans le cas de l'estimateur de la fonction de régression asymétrique, les propriétés de celui-ci n'ont pas été établies d'une manière générale mais sont étudiées en particulier pour un certain noyau fixé. Par exemple, l'étude de Jianhong Shi et Weixing Song [10] présente des résultats asymptotiques sur l'estimateur dans le cas d'un noyau gamma. D'autres auteurs [11] ont étudié les propriétés de l'estimateur de Gasser et Müller [12] qui correspond à un estimateur qui utilise la méthode du noyau mais qui est différent de l'estimateur de Nadaraya [6] et Watson [7].

L'objectif de notre travail est de présenter l'estimateur de la fonction de régression

dans le cas d'un ensemble de données continues et asymétriques via la méthode du noyau associée. Ce mémoire est composé de trois chapitre :

Le premier présente une introduction de la méthode du noyau associé dans le cas symétrique avec l'ensemble de ses propriétés ; des exemples de noyaux symétriques ainsi que des méthodes de sélection du paramètre h.

Le deuxième chapitre contient une présentation de l'estimateur dans le cas asymétrique pour lequel les propriétés ont été étudiées en détails pour les deux noyaux associés bêta et gamma.

Le troisième et dernier chapitre comporte la partie application, dans lequel nous employons l'estimateur asymétrique pour une étude sur des données simulées d'une part et sur des données réelles d'autre part.

Enfin, nous terminons par une conclusion générale.

CHAPITRE 1

RÉGRESSION NON PARAMÉTRIQUE

1.1 Introduction

Les modèles de régression sont généralement utilisés pour l'analyse d'un ensemble de données. Ces modèles se subdivisent en trois catégories et ceci en fonction du type de régresseur considéré.
Les modèles de la première catégorie en l'occurrence les "**modèles paramétriques**" sont fréquemment utilisés [3]. Néanmoins, dans certains cas ceux-ci ne sont plus valables et on recours alors aux modèles dits "**non paramétriques**". La dernière catégorie est celle des "**modèles semi-paramétriques**" qui est une composition des deux premiers modèles.

Dans ce qui suit, nous présentons l'estimateur non paramétrique de la fonction de régression par la méthode du noyau. Nous avons énoncé un ensemble de résultats relatifs à la qualité de cet estimateur. Enfin, les deux composantes essentielles à savoir "le noyau" et "le paramètre de lissage" sont traitées séparément et ceci en donnant quelques exemples de noyaux et des méthodes de sélection du paramètre.

1.2 Modèle de régression non paramétrique

Dans un modèle de régression non paramétrique, nous supposons une relation entre deux variables aléatoires : **la variable expliquée** Y qui est la variable dépendante et la variable **explicative** X qui est la variable indépendante, ceci sans fixer d'hypo-

thèses à priori quant à la forme de cette relation, **i.e** sans spécification de la forme de la fonction de régression.

Définition 1.2.1. *On considère* $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ *un échantillon de* n *couples de v.a et on suppose qu'elles sont indépendantes et identiquement distribuées (i.i.d). Un modèle de régression non paramétrique se présente sous la forme suivante :*

$$y_i = m(x_i) + \epsilon_i, \tag{1.1}$$

• *où :*

- $*$ m *est la fonction de régression inconnue que l'on estime,*
- $*$ x_i *représente la* i*-ème observation de la v.a* X*,*
- $*$ y_i *correspond à la* i*-ème observation de la v.a* Y
- $*$ *les* ϵ_i *sont appelés "erreurs aléatoires" (ou résidus) et sont supposées non corrélées et distribuées suivant une loi normale centrée (de moyenne 0) et de variance* σ^2*.*

• Un modèle de régression possède les caractéristiques suivantes :

- $*$ La fonction de régression n'est pas définie par une forme explicite.
- $*$ La définition de cette fonction est donnée comme une espérance conditionnelle, **i.e** :

$$m(x) = \mathbb{E}(Y/X = x). \tag{1.2}$$

Le problème principal réside dans l'estimation de la fonction de régression $m(x)$ qui est à priori inconnue, ceci sans supposition d'une quelconque hypothèse. La seule condition qu'on pose est celle de la régularité de la fonction m, en d'autres termes on suppose que la fonction m est de classe C^k ; **i.e** : elle est k-fois continuellement dérivable, k est un entier positif ou nul.

1.3 Méthode du noyau pour l'estimation de la fonction de régression

1.3.1 Présentation de la méthode

La méthode utilisée se résume à une estimation fonctionnelle en l'ocurrence, estimer en un point x fixé la fonction de régression $m(x)$.

La méthode du noyau se base sur un lissage et elle propose une moyenne pondérée comme estimateur de l'espérance conditionnelle (1.2).

1.3.2 Noyau associé

Définition 1.3.1. *On considère $\aleph_{x,h}$ un support d'une densité de probabilité f à estimer. Pour un point $x \in \aleph$ et un paramètre $h > 0$, on appelle "**noyau associé**" noté $K_{x,h}(.)$ toute densité de probabilité associée à une v.a continue ou discrète $\mathcal{K}_{x,h}$ de support $\aleph_{x,h}$ indépendant de h et qui contient au moins x, qui vérifie les conditions suivantes :*

$$\bigcup_{x\in\aleph} \aleph_{x,h} \supseteq \aleph, \lim_{h\to 0} \mathbb{E}(K_{x,h}) = x, \lim_{h\to 0} \mathbb{V}ar(K_{x,h}) = 0, \mathbb{V}ar(K_{x,h}) < +\infty. \tag{1.3}$$

Définition 1.3.2. *Un noyau $K_{x,h}(.)$ est dit d'ordre q s'il vérifie les conditions suivantes :* $\int_{\mathbb{R}} K_{x,h}(u)du = 1, \int_{\mathbb{R}} u^j K_{x,h}(u)du = 0, \forall j = 1, 2, ..., q-1$ *et* $\int_{\mathbb{R}} u^q K_{x,h}(u)du < \infty$.

1.3.3 Estimateur à noyau symétrique de la fonction de régression

La fonction de régression est donnée par définition par la relation suivante :

$$m(x) = \frac{r(x)}{f_X(x)}, \tag{1.4}$$

avec :

$$r(x) = \int_{\mathbb{R}} y f_{X,Y}(x, y)dy,$$

où :

- $f_{X,Y}(x, y)$ est la loi conjointe des deux variables aléatoires X et Y,
- $f_X(x) = \int f_{X,Y}(x, y)dy$ est la loi marginale relativement à la v.a X.

On considère $\hat{f}_{X,Y}$ un estimateur de la densité $f_{X,Y}$ calculé via la méthode du noyau défini par [13] à travers l'expression suivante :

$$\hat{f}_{X,Y}(x, y) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)K_{x,h}(Y_i). \tag{1.5}$$

D'autre part, on considère $\hat{r}(x)$ un estimateur de la fonction r :

$$\begin{aligned}\hat{r}(x) &= \int_{\mathbb{R}} y\hat{f}_{X,Y}(x,y)dy \\ &= \int_{\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)yK_{x,h}(Y_i)dy \\ &= \int_{\mathbb{R}} \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{x-X_i}{h})yK(\frac{y-Y_i}{h})dy\end{aligned}$$

$$= \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h}) \int_{-\infty}^{+\infty} \frac{1}{h}yK(\frac{y-Y_i}{h})dy. \tag{1.6}$$

On suppose que le noyau $K(.)$ est symétrique et on effectue le changement de variable :

$$u = \frac{y-Y_i}{h},$$

$$\Rightarrow y = Y_i - hu \ \ et \ \ dy = -hdu,$$

$$y \to -\infty : u \to +\infty; y \to +\infty : u \to -\infty,$$

en remplaçant dans l'expression (1.6) :

$$\begin{aligned}\int_{-\infty}^{+\infty} \frac{1}{h}yK(\frac{y-Y_i}{h})dy &= \int_{+\infty}^{-\infty} \frac{1}{h}(Y_i-hu)K(u)(-hdu) \\ &= \frac{h}{h}(-\int_{+\infty}^{-\infty} Y_i - hu)K(u)du \\ &= \int_{-\infty}^{+\infty} Y_iK(u)du - \int_{-\infty}^{+\infty} huK(u)du \\ &= Y_i\int_{-\infty}^{+\infty} K(u)du - h\int_{-\infty}^{+\infty} uK(u)du\end{aligned}$$

D'après la supposition sur la symétrie du noyau $K(.)$: $\int_{-\infty}^{+\infty} uK(u)du = 0$.

De plus, $K(.)$ est une densité de probabilité alors : $\int_{-\infty}^{+\infty} K(u)du = 1$, par conséquent :

$$\int_{-\infty}^{+\infty} \frac{1}{h}yK(\frac{y-Y_i}{h}) = Y_i,$$

en remplaçant dans l'expression de l'estimateur $\hat{r}(x)$:

$$\hat{r}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_iK_{x,h}(X_i). \tag{1.7}$$

Aussi, l'estimateur à noyau de la fonction de densité $\hat{f}_X(x)$ par la méthode de Parzen [5] vaut par définition :

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i).$$

Ces calculs effectués précédemment consituent une construction de la définition de l'estimateur à noyau associé symétrique de la fonction de régression.

Définition 1.3.3. *On concidère $K_{x,h}$ un noyau associé de cible x et de paramètre de lissage h. Pour la catégorie des noyaux associées symétriques continus avec*

$$K_{x,h}(.) = \frac{1}{h}K(\frac{x-.}{h}), \tag{1.8}$$

l'estimateur d'une fonction de régression continue basé sur l'échantillon $(x_1, y_1), ..., (x_n, y_n)$ connu sous le nom de l'estimateur à noyau de Nadaraya [6] et Watson [7] est donnée par l'expression suivante :

$$\hat{m}(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h})}{\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})}, \tag{1.9}$$

ou encore, via la forme du noyau associé $K_{x,h}$ cet estimateur équivaut à :

$$\hat{m}(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K_{x,h}(X_i)}{\frac{1}{nh}\sum_{i=1}^{n} K_{x,h}(X_i)} \tag{1.10}$$

Remarque 1.3.1 (Blodin). *[14] L'estimateur de la fonction de régression possède une autre écriture qui est donnée par la définition suivante :*

Définition 1.3.4. *L'estimateur à noyau (kernel estimate) défini par Nadaraya [6] et Watson [7] de la fonction de régression en un point **x** fixé (cible) peut s'écrire sous la forme suivante :*

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) y_i, \tag{1.11}$$

où :

• *$w_i(x)$ représente les poids des observations y_i qui dépendent de x dont l'expression analytique est formulée comme suit :*

$$w_i(x) = \frac{K_{x,h}(X_i)}{\sum_{i=1}^{n} K_{x,h}(X_i)}, \tag{1.12}$$

• *$K_{x,h}(.)$ désigne le noyau associé (kernel) qui définie la forme du voisinage autour de la cible x.*

• *Le paramètre de lissage $h > 0$ (bandwidth parameter) correspond à la taille de ce voisinage.*

1.3.4 Propriétés de l'estimateur

Les propriétés d'un estimateur sont utilisées pour la définition de la qualité de celui-ci et pour une éventuelle comparaison avec d'autres estimateurs. Parmi ces critères pris en considération dans la mesure de performances d'un estimateur, nous avons le biais, la variance et l'erreur quadratique.

1.3.4.1 Biais de l'estimateur

Définition 1.3.5. *Le biais de l'estimateur $\hat{m}(x)$ noté $biais(\hat{m}(x))$ est la différence entre l'espérance de l'estimateur et la fonction estimée,*

$$biais(\hat{m}(x)) = \mathbb{E}[\hat{m}(x)] - m(x),$$

l'estimateur $\hat{m}(x)$ est sans biais si :

$$\mathbb{E}[\hat{m}(x)] = m(x).$$

Remarque 1.1. *L'estimateur de Nadaraya [6] et Watson [7] est défini sous forme d'un quotient aléatoire, c'est pour cette cause que généralement nous utilisons l'approximation suivante comme terme de centrage [14] :*

$$\tilde{\mathbb{E}}[\hat{m}(x)] := \frac{\mathbb{E}[\hat{r}(x)]}{\mathbb{E}[\hat{f}_{X;n}(x)]}. \tag{1.13}$$

Proposition 1.1 (Lagha). *[15]Si Y est bornée et si $\lim_{n\to+\infty} nh = +\infty$, alors :*

$$\mathbb{E}(\hat{m}(x)) = \tilde{\mathbb{E}}(\hat{m}(x)) + o(\frac{1}{nh}).$$

Si $\mathbb{E}(Y^2) < +\infty$ et si $\lim_{n\to+\infty} nh^2 = +\infty$, alors :

$$\mathbb{E}(\hat{m}(x)) = \tilde{\mathbb{E}}(\hat{m}(x)) + o(\frac{1}{\sqrt{nh}}).$$

Proposition 1.2 (Blodin). *[14] On suppose que $m(.)$ et $f_X(.)$ sont de classe C^2 et que K est un noyau d'ordre 2 i.e il vérifie les conditions suivantes :*

- $\int_{\mathbb{R}} K(u)du = 1,$
- $\int_{\mathbb{R}} uK(u)du = 0,$
- $\int_{\mathbb{R}} u^2K(u)du < \infty;$

alors quand $h \to 0$ et $nh \to \infty$, nous avons ce qui suit :

$$biais\{\hat{m}(x)\} = \frac{h^2}{2}(m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}) \int_{\mathbb{R}} u^2K(u)du + o(h^2).$$

1.3.4.2 Variance de l'estimateur

On suppose les hypothèses suivantes sur le noyau K :

1. K est une densité de probabilité, i.e $\int_{\mathbb{R}} K(u)du = 1$,
2. K est bornée, i.e $\underset{u\in\mathbb{R}}{Sup}|K(u)| < \infty$,
3. $\lim_{|u|\to+\infty} |u|K(u) = 0$,
4. $\int_{\mathbb{R}} |K(u)|du < \infty$.

On note $\sigma^2(x)$ l'expression définie par :

$$\sigma^2(x) = \mathbb{V}ar[Y/X = x] = \frac{1}{f_X(x)} \int_{\mathbb{R}} y^2 f_{X,Y}(x,y)dy - [m(x)]^2,$$

ceci à condition que cette dernière soit bien définie.

Proposition 1.3 (Blodin). *[14] On suppose que la quantité $\mathbb{E}[Y^2]$ est finie ; alors, pour tout point de continuité x des fonctions $m(x)$, $f_X(x)$ et $\sigma^2(x)$ tel que $f_X(x) > 0$, on a :*

$$\begin{aligned} \mathbb{V}ar(\hat{m}(x)) &= \mathbb{E}[(\hat{m}(x) - \mathbb{E}[\hat{m}(x)]^2)], & (1.14)\\ &= \frac{1}{nh} \times \frac{\sigma^2(x)}{f_X(x)} \int_{\mathbb{R}} K^2(u)du + o(\frac{1}{h}). & (1.15) \end{aligned}$$

1.3.4.3 L'erreur moyenne quadratique

Définition 1.3.6. *L'erreur moyenne quadratique de l'estimateur $\hat{m}(x)$ ou le MSE (**M**ean **S**quare **E**rror) correspond à la moyenne de l'écart entre l'estimateur $\hat{m}(x)$ et la fonction estimée $m(x)$, on le note $MSE(m(x))$ et il vaut :*

$$MSE(\hat{m}(x)) = \mathbb{E}[\hat{m}(x) - m(x)]^2.$$

Théorème 1.3.1. *L'erreur moyenne quadratique peut s'exprimer en fonction du biais et de la variance de l'estimateur comme suit :*

$$MSE(\hat{m}(x)) = biais(\hat{m}(x))^2 + \mathbb{V}ar(\hat{m}(x)).$$

Démonstration 1.3.1.

$$\begin{aligned} \mathbb{E}[\hat{m}(x) - m(x)]^2 &= \mathbb{E}[\hat{m}(x) - \mathbb{E}[\hat{m}(x)] + biais(\hat{m}(x))]^2\\ &= \mathbb{E}[(\hat{m}(x) - \mathbb{E}[\hat{m}(x)])^2 + biais(\hat{m}(x))^2 + 2(\hat{m}(x) - \mathbb{E}[\hat{m}(x)])biais(\hat{m}(x)))\\ &= \mathbb{E}[(\hat{m}(x) - \mathbb{E}[\hat{m}(x)])^2] + biais^2(\hat{m}(x)) + 2\mathbb{E}[\hat{m}(x)]biais(\hat{m}(x)))]\\ &= \mathbb{V}ar(\hat{m}(x)) + 2(\mathbb{E}[\hat{m}(x)] - \mathbb{E}[\hat{m}(x)]) + biais^2(\hat{m}(x))\\ MSE(\hat{m}(x)) &= \mathbb{V}ar(\hat{m}(x)) + biais^2(\hat{m}(x)). \end{aligned}$$

■

Théorème 1.3.2 (Lagha). *[15] On suppose que les fonctions $m(.)$ et $f_X(.)$ sont de classe C^q et que le noyau K est d'ordre q, alors l'erreur moyenne quadratique de l'estimateur de la fonction de régression est donnée par :*

$$MSE(\hat{m}(x)) = \frac{h^{2q}}{(q!)^2}\{m^{(q)}(x) + qm^{(q)}(x)\frac{f'(x)}{f(x)}\}^2(\int_{\mathbb{R}} u^q K(u)du)^2(1+o(h))+$$

$$\frac{1}{nh}\{\frac{\sigma^2(x)}{f(x)}\}\int_{\mathbb{R}} K^2(u)du(1+o(h)).$$

1.3.4.4 L'erreur quadratique moyenne intégrée

Définition 1.3.7. *L'erreur quadratique moyenne intégrée de l'estimateur $\hat{m}(x)$ ou le MISE (**M**ean **I**ntegrated **S**quare **E**rror) représente l'intégrale de l'erreur moyenne quadratique,*

$$MISE(\hat{m}(x)) = \mathbb{E}(\int[\hat{m}(x) - m(x)]^2dx) = \int MSE(\hat{m}(x))dx.$$

Théorème 1.3.3 (Lagha). *[15] On suppose que les fonctions $m(.)$ et $f_X(.)$ sont de classe C^q et que le noyau K est d'ordre q alors l'erreur moyenne quadratique intégrée de l'estimateur $\hat{m}(x)$ est donnée par la formule suivante :*

$$MISE(\hat{m}(x)) = \frac{h^{2q}}{(q!)^2}\{\int_{\mathbb{R}} m^{(q)}(x)dx + q\int_{\mathbb{R}} m^{(q-1)}(x)\frac{f'(x)}{f(x)}dx\}^2(\int_{\mathbb{R}} u^q K(u)du)^2+$$

$$\frac{1}{nh}\int_{\mathbb{R}}\frac{\sigma^2(x)}{f(x)}dx\int_{\mathbb{R}} K^2(u)du(1+o(h)).$$

Remarque 1.2. *La méthode du noyau se compose principalement de deux éléments essentiels ; en l'occurrance , le noyau $K_{x,h}(.)$ et le paramètre de lissage h ou encore appelé largeur de la fenêtre h. Ces deux dérnières sont développées dans les parties suivantes.*

1.3.5 Exemples de noyaux continus symétriques

Le choix du noyau K est basé sur le type du support des données (i.e des observations de l'échantillon), parmi les noyaux utilisés :

• **Noyau parabolique ou d'Epanechnikov**

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & si \qquad u \in [-1,1] \\ 0 & sinon \quad . \end{cases}$$

• **Noyau gaussien**

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}), u \in \mathbb{R}.$$

• **Noyau triangulaire**

$$K(u) = \begin{cases} 1-|u| & si \qquad u \in [-1,1] \\ 0 & sinon \quad . \end{cases}$$

• **Noyau uniforme (ou rectangulaire)**

$$K(u) = \begin{cases} \frac{1}{2} & si \qquad u \in [-1,1] \\ 0 & sinon \quad . \end{cases}$$

• **Noyau quadratique**

$$K(u) = \frac{15}{16}(1-u^2)^2, u \in [-1,1].$$

• La figure (1.1) illustre les différents noyaux énoncés précédemment.

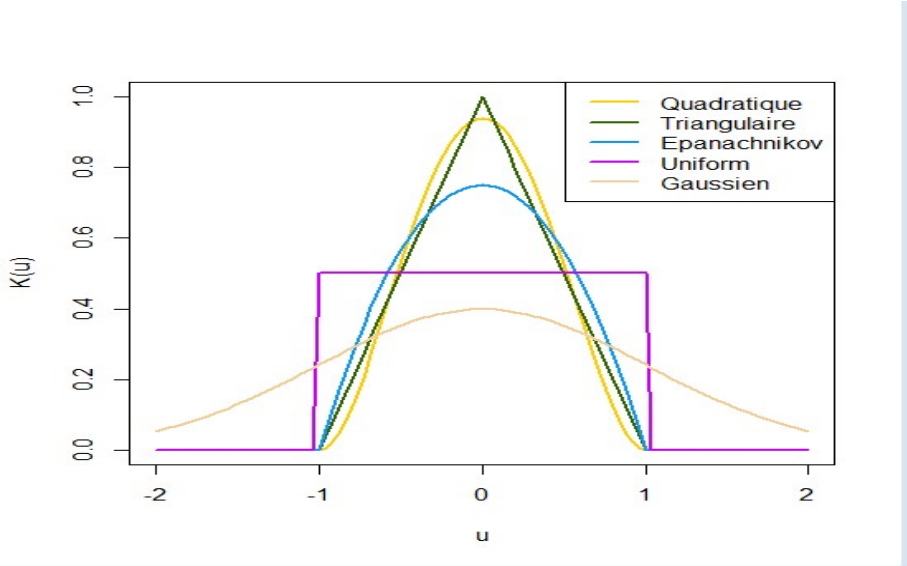

FIGURE 1.1: Exemples de noyaux symétriques

1.3.6 Choix de la fenêtre de lissage

Le paramètre de lissage h ou encore appelé la largeur de la fenêtre désigne la taille du voisinage autour du point cible x. Pour le choix de ce paramètre nous retenons essentiellement deux approches : la première est dite "**locale**" et consiste en la recherche d'un paramètre de lissage en minimisant un certain critère dépendant du point x. La deuxième approche est dite "**globale**" et consiste à la recherche d'un paramètre de lissage indépendamment de x.
Dans les parties suivantes, nous supposons que le noyau K est fixé ainsi les quantités présentées sont minimisées par rapport au paramètre h.

1.3.6.1 Méthode de minimisation du AMSE

Définition 1.3.8. *On appelle erreur quadratique moyenne asymptotique ou AMSE (**A**symptotic **M**ean **S**quare **E**rror) la quantité qui correspond à l'équation suivante :*

$$AMSE(h,K) = AMSE[\hat{m}(x)] = \frac{h^{2q}}{(q!)^2}\Big[m^{(q)}(x) + qm^{(q-1)}(x)\frac{f'_X(x)}{f_X(x)}\Big]^2\Big[\int_{\mathbb{R}} u^q K(u)du\Big]^2$$

$$+\frac{1}{nh}\Big[\frac{\sigma^2(x)}{f_X(x)}\Big]\int_{\mathbb{R}} K^2(u)du,$$

La fenêtre optimale au sens du critère local de minimisation de l'AMSE au point x notée h^{AMSE} est donnée par :

$$h^{AMSE} = Arg\underset{h}{Min}[AMSE(h,K)],$$

Ce paramètre de lissage i.e le h^{AMSE} correspond à la solution de l'équation suivante :

$$\frac{2q}{(q!)^2}h^{(2q-1)}\big[b(x,q)\big]^2 - \frac{1}{nh^2}v(x) = 0$$

où :

$$b(x,q) = m^{(q)}(x) + qm^{(q-1)}\frac{f'_X(x)}{f(x)}\int_{\mathbb{R}} u^q K(u)du,$$

et

$$v(x) = \frac{\sigma^2(x)}{f_X(x)}\int_{\mathbb{R}} K(u)du,$$

la valeur de la fenêtre optimale h^{AMSE} est définie par la quantité suivante :

$$h^{AMSE} = n^{-\frac{1}{(2q+1)}}\Big[\frac{q!(q-1)!\{\frac{\sigma^2(x)}{f_X(x)}\int K^2(u)du\}}{2\{m^{(q)}(x) + qm^{(q-1)}(x)\frac{f'_X(x)}{f_X(x)}\}^2[\int u^q K(u)du]^2}\Big]^{\frac{1}{2(q+1)}}$$

Remarque 1.3.2. *Dans l'estimateur $\hat{m}(x)$ défini avec la fenêtre h^{AMSE} figure la fonction $m(x)$ de régression qu'on a estimée. Ce type de fonction est appelé pseudo-estimateur [14]. La méthode "plug-in" reprend le même principe de la méthode précédente en remplaçant les paramètres inconnus par leurs estimateurs. Cette méthode contient des algorithmes itératifs tels que plug-in itéré.*

1.3.6.2 Méthode de minimisation du AMISE

Définition 1.3.9. *On appelle AMISE (**A**pproximate **M**ean **I**ntegrate **S**quare **E**rror) une approximation de l'erreur quadratique moyenne intégrée. Elle est associée à un paramètre de lissage h et donnée par la quantité suivante :*

$$AMISE(h) = \frac{h^4}{4}\int_{\mathbb{R}}\left[m''(x) + 2m'(x)\frac{f'_X(x)}{f(x)}\right]^2 dx[u^2K(u)] + \frac{1}{nh}\int_{\mathbb{R}}\frac{\sigma^2(x)}{f_X(x)}dx[K^2(u)].$$

La fenêtre minimisant l'AMISE notée h^{AMISE} correspond à :

$$h^{AMISE} = n^{-\frac{1}{5}}\left(\frac{\int_{\mathbb{R}}\frac{\sigma^2(x)}{f_X(x)}dx[K(u)^2]}{\int_{\mathbb{R}}\{m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}\}^2 dx[u^2K^2(u)]}\right)^{\frac{1}{5}}.$$

1.3.6.3 Méthode de la validation croisée

La méthode de la validation croisée ou **CV**(**C**ross **V**alidation) est une méthode qui se base sur la sélection du paramètre de lissage h en minimisant l'estimée d'une mesure d'un certain écart entre la fonction $m(x)$ et son estimateur $\hat{m}(x)$ par exemple le ISE définie en (1.3.10) . Le terme "**validation croisée**" est utilisé pour faire référence au fait qu'on utilise une partie pour fournir des informations sur une autre partie.

Définition 1.3.10. *On appelle l'erreur quadratique intégrée notée $ISE(\hat{m}(x), m(x))$ la quantité définie par :*

$$ISE(\hat{m}(x), m(x)) = \int_{\mathbb{R}}(\hat{m}(x) - m(x))^2 w(x) f_X(x)dx$$

$$ISE(\hat{m}(x), m(x)) = \int_{\mathbb{R}}\hat{m}^2(x)w(x)f_X(x)dx + \int_{\mathbb{R}}\hat{m}^2(x)w(x)f_X(x)dx - 2\int_{\mathbb{R}}\hat{m}(x)m(x)w(x)f_X(x)dx.$$

La quantité $\int_{\mathbb{R}} \hat{m}^2(x)w(x)f_X(x)dx$ est indépendante de h, alors minimiser la quantité $ISE(\hat{m}(x), m(x))$ revient à minimiser la quantité :

$$\int_{\mathbb{R}} \hat{m}^2(x)w(x)f_X(x)dx - 2\int_{\mathbb{R}} \hat{m}(x)m(x)w(x)f_X(x)dx.$$

L'expression $\int_{\mathbb{R}} \hat{m}(x)m(x)w(x)f_X(x)dx$ correspond par définition à $\mathbb{E}[\hat{m}(X)Yw(X)]$, cette dernière est estimée par :

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{m}_{-i}(Xi)Y_i w(X_i)],$$

d'autre part, la quantité $\int_{\mathbb{R}} \hat{m}^2(x)w(x)f_X(x)dx$ peut être approchée par : $\frac{1}{n}\sum_{i=1}^{n}[\hat{m}^2_{-i}w(X_i)]$. Minimiser le ISE revient à minimiser :

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}[\hat{m}^2_{-i}w(X_i)] - \frac{2}{n}[\hat{m}_{-i}Y_i w(X_i)] = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}_{-i} - Y_i]^2 w(X_i) - \frac{1}{n}\sum_{i=1}^{n}[Y_i]^2 w(X_i),$$

le deuxième terme est indépendant de h, alors minimiser le ISE équivaut à la minimisation de :

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}_{-i}(X_i)]^2 w(X_i).$$

Définition 1.3.11. *On appelle fonction de validation croisée notée $CV(h)$ la fonction*

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(Xi))^2 w(X_i), \tag{1.16}$$

- *Le terme $\hat{m}_{-i}(Xi)$ correspond à l'expression suivante :*

$$\hat{m}_{-i}(Xi) = \frac{\sum_{j=1,j\neq i}^{n} Y_j K_{X_i,h}(X_j)}{\sum_{j=1,j\neq i}^{n} K_{X_i,h}(X_j)}$$

*et il est appelé leave-on-out estimator et correspond à un estimateur à noyau basé sur l'échantillon dit "**réduit**" : $X_1, X_2, ..., \mathbf{X_{i-1}}, \mathbf{X_{i+1}}, ..., X_n$, où l'observation X_i a été supprimée.*

1.3.6.4 Méthode de la validation croisée généralisée

Dans la méthode précédente la quantité $m(x)$ est inconnue. La méthode de la validation croisée généralisée GCV(**G**eneral **C**ross **V**alidation) consiste à minimiser la variance estimée des résidus $\tilde{\sigma}$ définie comme suit :

$$\hat{\sigma}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}(X_i))^2.$$

Le paramètre de lissage optimale suivant le critère de minimisation de la fonction de validation croisée généralisée noté h* correspond à l'expression suivante :

$$h^* = arg\underset{h}{Min}\hat{\sigma}(h).$$

La différence entre cette fonction et la fonction de la méthode précédente se situe au niveau du terme $\hat{m}_{-i}(Xi)$.

1.4 Conclusion

Dans ce chapitre, nous avons présenté l'estimateur à noyau de la fonction de régression. Nous avons repris des résultats sur la qualité de l'estimateur tels que : le biais, le MSE, le MISE. Par ailleurs, ce chapitre contient une présentation de quelques méthodes utilisées pour le choix du paramètre de lissage.

CHAPITRE 2

ESTIMATEUR ASYMÉTRIQUE DE LA FONCTION DE RÉGRESSION

2.1 Introduction

Les noyaux symétriques décrits dans le chapitre précédant ne sont pas valables pour la modélisation de tout ensemble de données. En effet, ces noyaux qui sont également dit "**noyaux classiques**" ne sont pas adaptés pour des supports de données bornés, compactes ou encore bornés d'un côté ; ceci à cause de l'assignement de poids à l'extérieur du support dans le cas où le lissage est pris prés du bord, ce qui est appelé "**problème de bornes**".

On considère un n-échantillon de couples de de v.a $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ i.i.d, comme énoncé précédemment l'estimateur à noyau de la fonction de régression de Nadaraya [6] et Watson [7] est le suivant :

$$\hat{m}(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K_{x,h}(X_i)}{\frac{1}{nh}\sum_{i=1}^{n} K_{x,h}(X_i)},$$

dans le cas où le support $\aleph$ est symétrique le noyau $K_{x,h}$ vérifie :

$$K_{x,h} = \frac{1}{h} K(\frac{x - .}{h}); \tag{2.1}$$

par contre, dans le cas purement asymétrique $K_{x,h}$ ne vérifie l'égalité précédente (2.1) et il constitue un noyau variable en fonction du "point d'estimation" (de la cible) x [16].

Ce chapitre présente l'estimateur à noyau de la fonction de régression dans le cas univarié avec un ensemble de données continues qui est asymétrique. Dans un premier lieu, nous donnons la définition d'un noyau associé asymétrique suivie d'exemples de ceux-ci, nous reprenons par ailleurs des résultats asymptotiques sur l'estimateur dans le cas du noyau associé gamma. Nous terminons ce chapitre par des propositions sur les propriétés de l'estimateur dans le le cas noyau associé bêta.

2.2 Noyau associé continu asymétrique

Définition 2.2.1. *Soit $x \in \aleph$ et $h > 0$, on appelle "noyau associé continu asymétrique" noté $K_{x,h}$ toute densité de probabilité d'une variable aléatoire $\mathcal{K}_{x,h}$ sur le support $\aleph_{x,h}$ telle que :*

$$\aleph_{x,h} \neq \emptyset \tag{2.2}$$

$$\bigcup_{x} \aleph_{x,h} \supseteq \aleph \tag{2.3}$$

$$\lim_{h\to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x \tag{2.4}$$

$$\lim_{h\to 0} \mathbb{V}ar(\mathcal{K}_{x,h}) = 0 \tag{2.5}$$

$$\mathbb{V}ar(\mathcal{K}_{x,h}) < +\infty \tag{2.6}$$

La relation (2.2) traduit le fait que l'intersection entre le support des observations et le support du noyau associé continu asymétrique doit contenir au moins un élément .

2.3 Exemples de noyaux associés asymétriques

Noyau associé gaussien inverse

Soit $K_{x,h}$ un noyau dit gaussien inverse et qui est associé à la variable aléatoire $\mathcal{K}_{x,h}$, celui-ci est défini sur le support $\aleph_{x,h} =]0, +\infty)$ par l'expression suivante :

$$K_{x,h}(u) = \frac{1}{\sqrt{2\pi h u^3}} \exp\Big[-\frac{\eta(x,h)}{2xh}\left(\frac{u\eta(x,h)}{x} - 2 + \frac{x}{u\eta(x,h)}\right)\Big],$$

où la fonction η est définie pour tout $x > 0$ par :

$$\eta(x,h) = (1 - 3xh)^{\frac{1}{2}}.$$

Noyau associé gaussien inverse réciproque

Le noyau gaussien inverse réciproque $K_{x,h}$ lié à la v.a $\mathcal{K}_{x,h}$ est défini sur l'ensemble $\aleph =]0, +\infty)$ et est donné par :

$$K_{x,h} = \frac{1}{\sqrt{2\pi h u}} \exp\Big[- \frac{\zeta(x,h)}{2h}\Big(\frac{u}{\zeta(x,h)} - 2 + \frac{\zeta(x,h)}{u}\Big)\Big],$$

où $\zeta(x,h)$ est une fonction telle que :

$$\zeta(x,h) = (x^2 + xh)^{\frac{1}{2}}.$$

Noyau associé lognormal

Pour la loi log-normal $Log\mathcal{N}(\mu, \sigma^2)$? on considère le noyau continu $K_{x,h}$ lié à la v.a $\mathcal{K}_{x,h}$ de support $\aleph =]0, +\infty)$ et qui est défini par :

$$K_{x,h}(u) = \frac{1}{uh\sqrt{2\pi}} \exp\Big[- \frac{1}{2}\Big(\frac{1}{h}\log(\frac{u}{x}) - h\Big)^2\Big].$$

La figure (2.1) est une représentation des noyaux précédant avec une cible $x = 0.5$ et un paramètre de lissage $h = 0.5$.

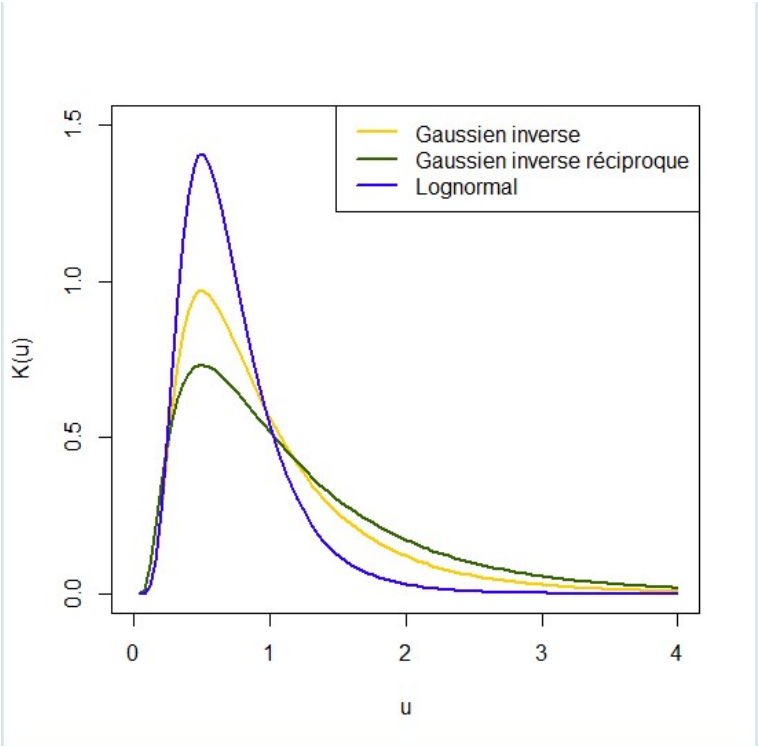

FIGURE 2.1: Exemple de noyaux associés assymétriques avec $x = h = 0.5$

Noyau associé Birnbaum Sauders Généralisé (GBS)

Définition 2.3.1. *Les noyaux GBS sont développés à partir des distributions GBS. Soit* $K_{GBS(\alpha,\beta,g)}$ *le noyau associé GBS qui est défini sur le support :* $\aleph_{x,h} = \mathbb{R}_+^*$, *où* α *et* β *sont des paramètres non négatifs de forme et d'échelle respectivement, ce noyau est défini comme suit :*

$$K_{GBS_{(\alpha,\beta,g)}}(t) = K_{GBS_{(\sqrt{h},x,g)}} = cg\Big(\frac{1}{h}\Big(\frac{t}{x}+\frac{x}{t}-2\Big)\Big)\frac{1}{\sqrt{4\pi}}\Big(\frac{1}{\sqrt{tx}}+\sqrt{\frac{x}{t^3}}\Big), t>0, h>0, x>0,$$

où g *est un générateur d'une densité de probabilité et* $c = \frac{1}{\int_{-\infty}^{+\infty} g(x^2)dx}$ *est appelée constante de normalisation.*

Parmi les cas particulier de la famille de noyaux GBS figurent les noyaux : BS, BS-Power-Exponential (BS-PE) et BS-Student-t. Ceux-ci sont définis par les expressions analytiques représentées dans le tableau (2.1).

Distribution	Noyau
BS	$\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-1}{2h}\big(\frac{t}{x}+\frac{x}{t}-2\big)\Big)\frac{1}{\sqrt{4h}}\big(\frac{1}{\sqrt{tx}}+\sqrt{\frac{x}{t^3}}\big)$
BS-PE	$\frac{\upsilon}{2^{\frac{1}{2\upsilon}}\Gamma\left(\frac{1}{2\upsilon}\right)}\exp\Big(-\frac{1}{2h^{\upsilon}}\big(\frac{t}{x}+\frac{x}{t}-2\big)^{\upsilon}\Big)\frac{1}{\sqrt{4h}}\Big(\frac{1}{\sqrt{tx}}+\sqrt{\frac{x}{t^3}}\Big), \upsilon>0$
BS-t	$\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\Big[1+\frac{1}{\upsilon h}\big(\frac{t}{x}+\frac{x}{t}-2\big)\Big]^{-\left(\frac{\upsilon+1}{2}\right)}\frac{1}{\sqrt{4h}}\big(\frac{1}{\sqrt{tx}}+\sqrt{\frac{x}{t^3}}\big), \upsilon>0$

TABLE 2.1: Cas particulier de noyaux GBS

La figure (2.2) illustre les noyaux précédant pour une cible $x = 1$ et un paramètre de lissage $h = 0.03$ et $\upsilon = 2$ pour les deux noyaux BS−PE et BS−t .

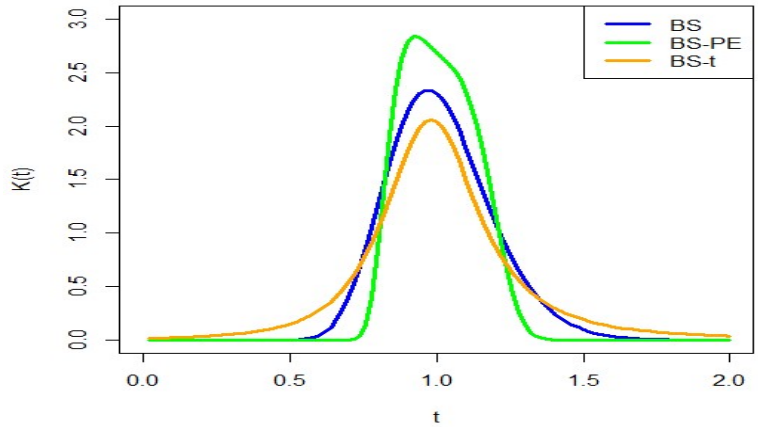

FIGURE 2.2: Noyaux BS, BS−PE et BS−t avec $x = 1$ et $h = 0.03$

2.4 Estimateur à noyau asymétrique de la fonction de régression

L'estimateur à noyau associé continu asymétrique (dit aussi **non classique** [17]), est approprié pour une estimation dans le cas d'un support de données compact ou borné d'un côté.

On considère un échantillon i.i.d $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ issu du couple de variables aléatoires (X, Y), on suppose que la variable X est définie sur le support $\aleph \subseteq \mathbb{R}$ borné ou borné d'un côté, i.e $\aleph = [a, b]$ où $a \in \mathbb{R}$ et $b \in \bar{\mathbb{R}}$ (exemples : $\aleph = [0, 2]$, $\aleph = [0, +\infty]$, ... etc) ; d'une manière analogue à l'estimateur de Nadaraya [6] et Watson [7] dans le cas d'un support de données symétrique nous retrouvons dans [10] une définition de cet estimateur dans le cas asymétrique exprimé par la définition suivante :

Définition 2.4.1. *L'estimateur à noyau de la fonction de régression pour un support de données non négatives est donné par :*

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_{x,h}(X_i) Y_i}{\sum_{i=1}^{n} K_{x,h}(X_i)}. \tag{2.7}$$

2.5 Résultats sur les propriétés de l'estimateur à noyau gamma

2.5.1 Rappels sur la loi gamma

La loi gamma est une loi continue asymétrique définie sur l'ensemble $\aleph = \mathbb{R}^+$ qui possède deux paramètres affectant respectivement la forme et l'échelle de sa représentation graphique.

Soit X une v.a suivant une loi gamma de paramètres a, b, sa densité de probabilité est la suivante :

$$\gamma(x) = \frac{x^{a+1} \exp\left(-\frac{x}{b}\right)}{\Gamma(a) b^a},$$

où :

$$\Gamma(a) = \int_{\mathbb{R}^+} \exp(-t) t^{a-1} dt.$$

Propriété 1. $\mathbb{E}(X) = ab$ *et* $\mathbb{V}ar(X) = ab^2$.

2.5.2 Présentation du noyau

On considère $K_{(x/h+1,h)}$ le noyau associé à la variable aléatoire $\mathcal{K}_{(x/h+1,h)}$ de loi gamma et de support $\aleph_{x,h} = \mathbb{R}^+$, ce noyau est défini comme suit :

$$K_{(x/h+1,h)} = \frac{t^{\frac{x}{h}} \exp\left(-\frac{t}{h}\right)}{h^{\frac{x}{h}+1}\Gamma(\frac{x}{h}+1)}, \tag{2.8}$$

où : $\Gamma(\frac{x}{h}+1) = \int\limits_{\mathbb{R}^+} \exp(-t)t^{\frac{x}{h}}dt$ et h est le paramètre de lissage vérifiant $h \to 0$ et $nh \to \infty$ quand $n \to \infty$.

• La figure (2.3) illustre le noyau gamma pour une cible $x = 0.1$ et des paramètres de lissage différents $h \in \{0.2, 0.3, 0.4\}$

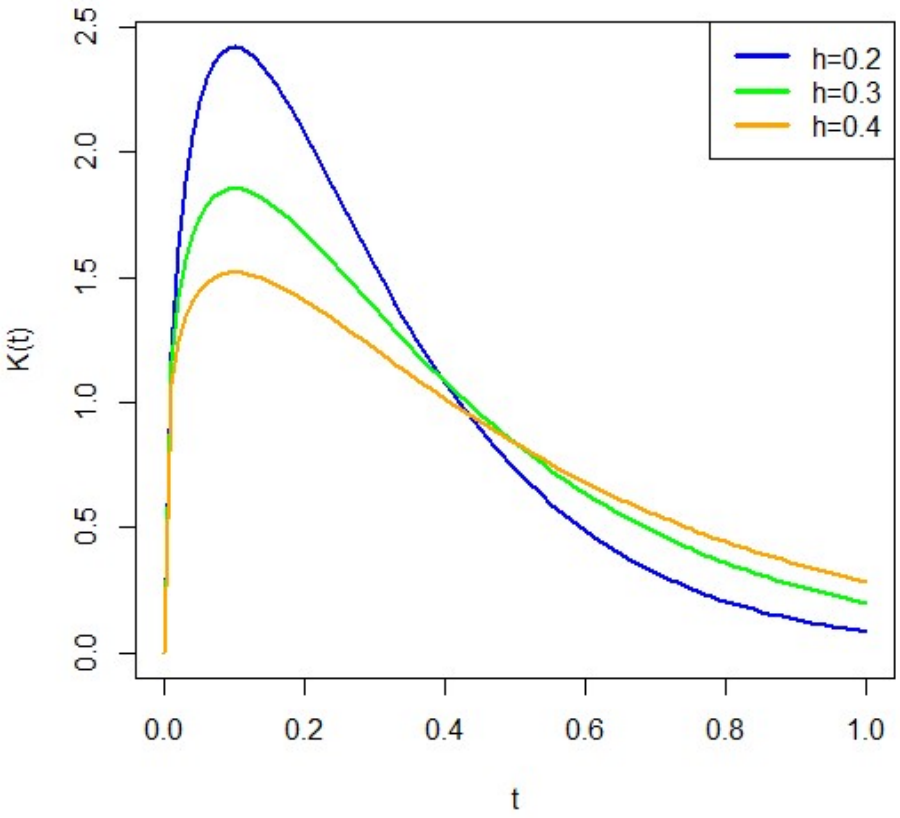

FIGURE 2.3: Noyau gamma pour une cible $x = 0.1$

2.5.3 Vérification des conditions d'un noyau associé

Dans cette partie nous vérifions les conditions d'un noyau associé dans le cas général énoncées dans (2.2.1) sur le noyau associé gamma :

1. $\mathbb{R}_+ \cap \mathbb{R}_+ = \mathbb{R}_+ \neq \emptyset$,
2. $\bigcup \mathbb{R}_+ = \mathbb{R}_+$,
3. $\lim_{h\to 0} \mathbb{E}(\mathcal{K}_{\frac{x}{h}+1,h}) = \lim_{h\to 0}(\frac{x}{h}+1)h = \lim_{h\to 0} x + h = x$,
4. $\lim_{h\to 0} \mathbb{V}ar(K_{x,h}) = \lim_{h\to 0} xh + h^2 = 0$,
5. $\mathbb{V}ar(\mathcal{K}_{\frac{x}{h}+1,h}) = (\frac{x}{h}+1)h^2 = xh + h^2 < \infty$.

2.5.4 Estimateur à noyau gamma de la fonction de régression

Définition 2.5.1. *Pour un support de données non négatives $\aleph_{x,h} = \mathbb{R}^+$ et d'une manière équivalente à l'estimateur de Nadaraya [6] et Watson [7] l'estimateur à noyau gamma de la fonction de régression $m(x)$ est défini par [10] :*

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_{x/h+1,h}(X_i)Y_i}{\sum_{i=1}^{n} K_{x/h+1,h}(X_i)}. \tag{2.9}$$

Définition 2.5.2. *Une autre définition de l'estimateur défini en (2.9) qui découle de la définition du noyau associé gamma est donnée par :*

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} X_i^{\frac{x}{h}} \exp(-\frac{X_i}{h})Y_i}{\sum_{i=1}^{n} X_i^{\frac{x}{h}} \exp(-\frac{X_i}{h})}. \tag{2.10}$$

2.5.5 Propriétés de l'estimateur

On considère les suppositions suivantes utilisées dans les calculs des propriétés de l'estimateur (2.10) :

(S_1) La dérivée seconde de la fonction f est continue et bornée dans $[0, +\infty)$,

(S_2) $\mathbb{E}(\epsilon/X) = 0$ et les dérivées secondes des fonctions fm, fm^2 sont continues et bornées dans $[0, +\infty)$,

(S_3) La dérivée seconde de $\delta^2(x) = \mathbb{E}(\epsilon^2/X = x)$ est continue et bornée pour tout $x > 0$, la dérivée seconde de $f\sigma^2$ est continue et bornée dans $[0, +\infty)$,

(S_4) $h \to 0, n\sqrt{h} \to \infty$ quand $n \to \infty$.

2.5.5.1 Biais de l'estimateur

Soient $b(x)$ et $v(x)$ deux termes définis comme suit :

$$b(x) = m'(x) + \frac{1}{2}xm''(x) + \frac{xm'(x)f'(x)}{f(x)}, \tag{2.11}$$

$$v(x) = \frac{\sigma^2(x)}{2f(x)\sqrt{\pi x}}.$$

Théorème 2.5.1 (Jianhong et Weixing). *[10] On suppose que (S_1),(S_2),(S_3),(S_4) sont vérifiées, alors pour tout $x \in [0, +\infty)$ avec $f(x) > 0$:*

1. *Pour $x > 0$:*
$$biais(\hat{m}(x)) = hb(x) + o(h) + o(\frac{h^{\frac{1}{4}}}{\sqrt{n}}), \tag{2.12}$$
2. *Pour $x = 0$*
$$biais(\hat{m}(0)) = hm'(0) + o(h). \tag{2.13}$$

Démonstration 2.5.1. *Une décomposition de la différence entre la fonction et son estimateur est donné par :*

$$\hat{m}(x) - m(x) = \frac{B_n(x) + V_n(x)}{f(x)} + [\frac{1}{\hat{f}(x) - \frac{1}{f(x)}}](B_n(x) + V_n(x)), \tag{2.14}$$

où :

$$B_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)[m(X_i) - m(x)]; V_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\frac{x}{h}+1,h}(X_i)\epsilon_i,$$

et $K_{\frac{x}{h},h}(X_i)$ est le noyau associé gamma.
D'après l'hypothèse (S_2) le biais vaut :

$$\mathbb{E}\big(\hat{m}(x)\big) - m(x) = \frac{B_n(x)}{\hat{f}(x)}$$

et d'après [18] : $\hat{f}(x) = f(x) + o(1)$.
L'espérance de $B_n(x)$ peut s'écrire sous la forme :

$$\mathbb{E}(B_n(x)) = \mathbb{E}\big(K_{\frac{x}{h},h}(X)m(X)\big) - m(x)\mathbb{E}\big(K_{\frac{x}{h}+1,h}(X)\big).$$

On considère le lemme suivant issue de [10] où figure également la preuve et qui est utilisé dans les différents calculs intermédiaires :

Lemme 2.5.1. *[Jianhong et Weixing][10] Soit $l(u)$ une fonction dont la dérivée second est continue et bornée sur $(0,+\infty)$ et soit $g(u,p_k,\lambda_k)$ la densité de probabilité d'une v.a suivant une loi gamma ; alors, pour tout $x>0$ et $k\geq 1$:*

$$\int_0^{+\infty} g(u,p_k,\lambda_k)l(u)du = l(x)+\frac{[2l'(x)+xl''(x)]h}{2k}+o(h).$$

Nous appliquons le lemme (2.5.1) en prenant $l(u)=H(u)=m(u)\times f(u)$ dans le calcul de $\mathbb{E}[\mathcal{K}_{\frac{x}{h},h}]$ comme suit :

$$\begin{aligned}\mathbb{E}[\mathcal{K}_{\frac{x}{h},h}(X)m(X)] &= \int_0^{+\infty}\mathcal{K}_{\frac{x}{h},h}(X_i)m(u)f(u)du\\ &= H(x)+\frac{[2H'(x)+xH''(x)]h}{2}+o(h)\\ &= m(x)f(x)+h(m(x)f(x))'+\frac{xh}{2}(m(x)f(x))''+o(h)\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\mathcal{K}_{\frac{x}{h},h}(X)m(X)] = m(x)f(x)+hm'(x)f(x)+hm(x)f'(x)+\\ \frac{xh}{2}m''(x)f(x)+hxm'(x)f'(x)+\frac{hx}{2}f'(x)+o(h).\end{aligned}$$

En appliquant le lemme (2.5.1) d'une manière similaire à la précédente mais ceci dans le calcul de $\mathbb{E}(K_{\frac{x}{h},h})$ et en prenant $l(u)=f(u)$:

$$\begin{aligned}\mathbb{E}(K_{\frac{x}{h},h}(u)) &= \int_0^{+\infty}\mathcal{K}_{\frac{x}{h},h}(u)f(u)du\\ &= f(x)+\frac{[2f'(x)+xf''(x)]h}{2}+o(h),\\ m(x)\mathbb{E}(K_{\frac{x}{h},h}(u)) &= m(x)\big[f(x)+hf'(x)+\frac{hx}{2}f''(x)+o(h)\big]\\ &= m(x)f(x)+hm(x)f'(x)+\frac{hx}{2}m(x)f''(x)+m(x)o(h).\end{aligned}$$

En remplaçant les deux résultats précédents dans la formule de l'espérance de $B_n(x)$:

$$\begin{aligned}\mathbb{E}(B_n(x)) &= \mathbb{E}(K_{\frac{x}{h}+1,h}(X))-m(x)\mathbb{E}(K_{\frac{x}{h}+1,h}(X))\\ = m(x)f(x)+hm'(x)f(x)+hm(x)f'(x)+\frac{xh}{2}m''(x)f(x)+hxm''(x)f'(x)\\ +\frac{xh}{2}m(x)f''(x)+o(h)-m(x)f(x)-hm(x)f'(x)-\frac{hx}{2}m(x)f''(x)-m(x)o(h),\\ \mathbb{E}(B_n(x)) = h[m'(x)f(x)+\frac{x}{2}m''(x)f(x)+xm'(x)f'(x)]+o(h).\end{aligned}\tag{2.15}$$

D'autre part, la variance d'une v.a est bornée par son moment d'ordre 2, alors :

$$\mathbb{V}ar(B_n(x))\leq\frac{1}{n}\mathbb{E}\left[\frac{1}{h\Gamma(\frac{x}{h}+1)}\Big(\frac{X}{h}\Big)^{\frac{x}{h}}\exp\Big(\frac{X}{h}^{\frac{x}{h}}\Big)[m(X)-m(x)]\right]^2\tag{2.16}$$

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\Big[\frac{1}{h\Gamma(\frac{x}{h}+1)}\Big(\frac{X}{h}\Big)^{\frac{x}{h}}\exp\Big(\frac{X}{h}^{\frac{x}{h}}\Big)[m(X)-m(x)]\Big]^2\\
=\;&\frac{1}{n}\int_0^{+\infty}\frac{1}{h^2\Gamma^2(\frac{x}{h}+1)}\Big(\frac{u}{h}\Big)^{\frac{2x}{h}}\exp\Big(-\frac{2u}{h}\Big)[m(X)-m(x)]^2f(u)du\\
=\;&\frac{\Gamma(\frac{2x}{h}+1)}{nh2^{\frac{2x}{h}+1}\Gamma^2(\frac{x}{h}+1)}\int_0^{+\infty}K_{\frac{2x}{h}+1,\frac{h}{2}}(u)f(u)[m(u)-m(x)]^2du.
\end{aligned}$$

Pour $x>0$, d'après la formule de stiriling quand $h\to 0$:

$$\frac{\Gamma(\frac{2x}{h}+1)}{nh2^{\frac{2x}{h}+1}\Gamma^2(\frac{x}{h}+1)}=\frac{1}{2n\sqrt{\pi xh}}[1+o(1)]. \tag{2.17}$$

D'après la supposition (S_1) et la continuité des fonction f,m,m',f', nous avons :

$$\int_0^{+\infty}K_{p_2,\lambda_2}(u)f(u)[m(u)-m(x)]^2du=\frac{xf(x)m'^2(x)h}{2}+o(h). \tag{2.18}$$

En combinant les deux équations (2.17) et (2.18) :

$$\mathbb{V}ar(B_n(x))=o(\frac{\sqrt{h}}{n}), \tag{2.19}$$

de (2.15) et (2.19) :

$$biais(\hat{m}(x))=hb(x)+o(h)+o(\frac{h^{\frac{1}{4}}}{\sqrt{n}}),$$

où $b(x)$ est défini en (2.11).

Pour $x=0$:

D'après (2.16) :

$$\mathbb{V}ar(B_n(0))\le\frac{1}{n}\mathbb{E}\Big[\frac{1}{h}\exp\Big(-\frac{X}{h}\Big)[m(X)-m(0)]\Big]^2,$$

$$\frac{1}{n}\mathbb{E}\left[\frac{1}{h}\exp\Big(-\frac{X}{h}\Big)[m(X)-m(0)]\right]^2=\frac{1}{nh}\int_0^{+\infty}\exp\Big(-\frac{2u}{h}\Big)[m(u)-m(0)]^2f(u)du. \tag{2.20}$$

D'après la supposition (S_1) et la continuité des fonctions : f,m',f',m :

$$\int_0^{+\infty}\exp\Big(-\frac{2u}{h}\Big)[m(u)-m(0)]^2f(u)du=o(h^2). \tag{2.21}$$

De (2.20) et (2.21) : $\mathbb{V}ar(B_n(0))=o(\frac{h}{n})$ et en combinaison avec (2.15), nous avons :

$$biais(\hat{m}(x))=hm'(0)+o(h).$$

■

2.5.5.2 Variance de l'estimateur

Théorème 2.5.2 (Jianhong et Weixing). *[10] On suppose que (S_1),(S_2),(S_3),(S_4) sont vérifiées, alors pour tout $x \in [0, +\infty)$ avec $f(x) > 0$:*

1. *Pour $x > 0$:*
$$\mathbb{V}ar(\hat{m}(x)) = \frac{\sigma^2(x)}{2nf_X(x)\sqrt{\pi x h}} + o(\frac{1}{n\sqrt{h}}), \tag{2.22}$$
2. *Pour $x = 0$:*
$$\mathbb{V}ar(\hat{m}(0)) = \frac{\sigma^2(0)}{2nhf(0)} + (\frac{1}{nh}). \tag{2.23}$$

Démonstration 2.5.2. *D'après la décomposition (2.14) :*

$$\begin{aligned}
\hat{m}(x) &= \frac{B_n(x)+V_n(x)}{f(x)} + \Big[\frac{1}{\hat{f}(x)} - \frac{1}{f(x)}\Big]\big[B_n(x)+V_n(x)\big] + m(x) \\
&= \frac{B_n(x)}{\hat{f}(x)} + \frac{V_n(x)}{\hat{f}(x)} + m(x) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)\big[m(X_i)-m(x)\big]}{\hat{f}(x)} + \frac{\frac{1}{n}\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)\epsilon_i}{\hat{f}(x)} + m(x) \\
&= m(x)\Big[1 - \frac{\frac{1}{n}\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)}{\hat{f}(x)}\Big] + \frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)m(X_i)}{n\hat{f}(x)} + \frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)\epsilon_i}{n\hat{f}(x)} \\
&= \frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)m(X_i)}{n\hat{f}(x)} + \frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)\epsilon_i}{n\hat{f}(x)}.
\end{aligned}$$

La variance de l'estimateur $\hat{m}(x)$ vaut alors :

$$\mathbb{V}ar(\hat{m}(x)) = \mathbb{V}ar\left(\frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)m(X_i)}{n\hat{f}(x)} + \frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)\epsilon_i}{n\hat{f}(x)}\right) \tag{2.24}$$

$$= \mathbb{V}ar\left(\frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)m(X_i)}{n\hat{f}(x)}\right) + \mathbb{V}ar\left(\frac{\sum_{i=1}^{n} K_{\frac{x}{h},h}(X_i)\epsilon_i}{n\hat{f}(x)}\right) \tag{2.25}$$

$$= \frac{1}{n^2\hat{f}^2(x)}\sum_{i=1}^{n} K^2_{\frac{x}{h},h}(X_i)\hat{V}ar(\epsilon_i) \tag{2.26}$$

$$= \frac{1}{n^2h^2\Gamma^2\big(\frac{x}{h}+1\big)}\sum_{i=1}^{n}\Big(\frac{X_i}{h}\Big)^{\frac{2x}{h}} exp\Big(\frac{-2X_i}{h}\Big)\sigma^2(X_i) \tag{2.27}$$

D'une part :

$$\mathbb{E}[\mathbb{V}ar(\hat{m}(x))] = \mathbb{V}ar(\hat{m}(x)) \tag{2.28}$$

D'autre part :

$$\begin{aligned}\mathbb{E}[\mathbb{V}ar(\hat{m}(x))] &= \frac{1}{nh^2\Gamma^2\left(\frac{x}{h}+1\right)}\mathbb{E}\left[\left(\frac{X}{h}\right)^{\frac{2x}{h}}\exp\left(-\frac{2X}{h}\right)\sigma^2(X)\right]\\ \mathbb{E}\left[\left(\frac{X}{h}\right)^{\frac{2x}{h}}\exp\left(-\frac{2X}{h}\right)\sigma^2(X)\right] &= \int_0^{+\infty}\left(\frac{t}{h}\right)^{\frac{2x}{h}}\exp\left(-\frac{2t}{h}\right)\sigma^2(t)f(t)dt\\ &= \frac{h\Gamma\left(\frac{2x}{h}+1\right)}{2^{\frac{2x}{h}}}\int_0^{+\infty}K_{(\frac{2x}{h},h)}(t)\sigma^2(\frac{t}{2})f(\frac{t}{2})dt.\end{aligned}$$

Á travers l'application du lemme (2.5.1) pour $l = \sigma^2 \times f$ et d'après la supposition (S_3) ainsi que l'approximation de stiriling (2.17) :

$$\frac{1}{nh^2\Gamma^2\left(\frac{x}{h}+1\right)}\mathbb{E}\left[\left(\frac{X}{h}\right)^{\frac{2x}{h}}\exp\left(-\frac{2X}{h}\right)\sigma^2(X)\right] = \frac{\sigma^2(x)f(x)}{2n\sqrt{\pi xh}} + o\left(\frac{\sqrt{h}}{n}\right), \tag{2.29}$$

de plus :

$$\hat{f}(x) = f(x) + o(1) \tag{2.30}$$

$$\mathbb{V}ar(\hat{m}(x)) = \mathbb{E}[\mathbb{V}ar(\hat{m}(x))] \tag{2.31}$$

$$\frac{\sqrt{h}}{n} = o(\frac{1}{n\sqrt{h}}), \tag{2.32}$$

de (2.30), (2.31) et (2.32) :

$$\mathbb{V}ar(\hat{m}(x)) = \frac{\sigma^2(x)}{2nf(x)\sqrt{\pi xh}} + o(\frac{1}{n\sqrt{h}}).$$

Pour $x = 0$:

De (2.29) :

$$\frac{1}{nh^2}\mathbb{E}\Big[\exp\left(\frac{-2X}{h}\right)\sigma^2(X)\Big] = \frac{\sigma^2(0)f(0)}{2nh} + o(\frac{1}{n}),$$

ainsi d'après l'égalité : $\hat{f}(0) = f(0) + o(1)$:

$$\mathbb{V}ar(\hat{m}(0)) = \frac{\sigma^2(0)}{2nhf(0)} + o(\frac{1}{n\sqrt{h}}).$$

■

2.5.5.3 Erreur moyenne quadratique

Á partir des expressions du biais et de la variance de l'estimateur à noyau gamma défini dans (2.5.5.1) et (2.5.5.2) l'erreur moyenne quadratique de l'estimateur définies en (2.10) noté $MSE(\hat{m}(x))$ pour tout $x \in [0, +\infty)$ vaut :

1. Pour $x > 0$:

$$MSE(\hat{m}(x)) = h^2[m'(x)+\frac{1}{2}xm''(x)+\frac{xm'(x)f'(x)}{f(x)}]^2+\frac{\sigma^2(x)}{2f(x)\sqrt{\pi xnh}}+o(h^2)+o(\frac{1}{n\sqrt{h}})+o(\frac{h^{\frac{5}{4}}}{\sqrt{n}}), \tag{2.33}$$

2. Pour $x = 0$:

$$MSE(\hat{m}(0)) = h^2 m'^2(0) + \frac{\sigma^2(0)}{2nhf(0)} + o(h^2) + o(\frac{1}{nh}). \tag{2.34}$$

2.6 Propositions sur les propriétés de l'estimateur à noyau bêta

2.6.1 Rappel sur la loi bêta

On considère $B(x)$ la densité de probabilité d'une v.a suivant une loi bêta de paramètre a, b, cette densité est définie sur l'ensemble $\aleph = [0, 1]$ par l'expression suivante :

$$B(x) = \frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)},$$

où $a > 0, b > 0$ et vérifient : $\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dt$.

Propriété 2. $\mathbb{E}[X] = \frac{a}{a+b}$ *et* $\mathbb{V}ar[X] = \frac{ab}{(a+b)^2(a+b+1)}$.

2.6.2 Présentation du noyau associé bêta

$\aleph = [0, 1]$.

Définition 2.6.1. *Soit* $K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}$ *le noyau associé à la variable aléatoire* $\mathcal{K}_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}$ *de loi bêta et de support* $\aleph_{x,h} = [0, 1]$, *ce noyau est donné par la relation suivante :*

$$K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}(t) = \frac{t^{\frac{x}{h}}(1-t)^{\frac{1-x}{h}}}{\beta(\frac{x}{h}+1,\frac{1-x}{h}+1)},$$

où : $\beta(\frac{x}{h}+1,\frac{1-x}{h}+1) = \int_0^1 t^{\frac{x}{h}}(1-t)^{\frac{1-x}{h}}$.

La figure suivante (2.4) représente une allure d'un noyau associé bêta pour un point cible $x = 0$ et différents paramètres de lissage $h \in \{0.2, 0.3, 0.4\}$.

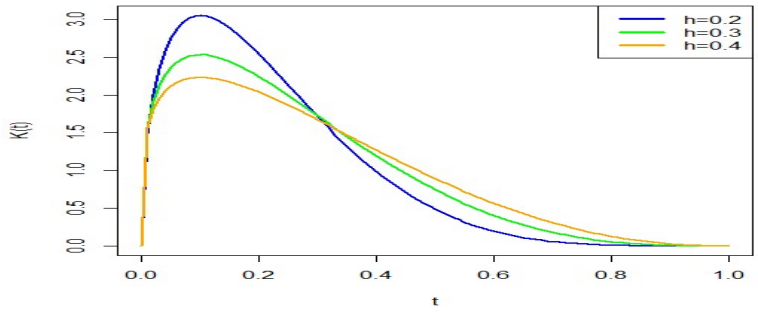

FIGURE 2.4: Noyau associé bêta $K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}(.)$ avec x=0

2.6.3 Vérification des conditions d'un noyau associé

1. $[0,1] \cap [0,1] = [0,1] \neq \emptyset$,
2. $\bigcup_x [0,1] = [0,1]$,
3. $\lim\limits_{h\to 0} \mathbb{E}[\mathcal{K}_{Be(\frac{x}{h}+1,\frac{(1-x)}{h}+1)}] = \lim\limits_{h\to 0} \frac{x+h}{1+2h} = x$,
4. $\mathbb{V}ar[\mathcal{K}_{Be(\frac{x}{h}+1,\frac{(1-x)}{h}+1)}] = \frac{x(1-x)h+h^2+h^3}{(1+2h)^2(1+3h)} < \infty$,
5. $\lim\limits_{h\to 0} \mathbb{V}ar[\mathcal{K}_{Be(\frac{x}{h}+1,\frac{(1-x)}{h}+1)}] = 0$.

2.6.4 Estimateur de la fonction de régression à noyau associé Bêta

Définition 2.6.2. *On considère un échantillon iid* $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ *du couple* (X,Y) *dont les valeurs des observations de la v.a* X *appartiennent au support* $\aleph = [0,1]$*, l'estimateur de la fonction de régression à noyau bêta est donné par :*

$$\hat{m}(x) = \frac{\sum\limits_{i=1}^{n} Y i K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}(X_i)}{\sum\limits_{i=1}^{n} K_{(\frac{x}{h}+1,\frac{1-x}{h}+1)}(X_i)}. \tag{2.35}$$

Définition 2.6.3. *Une autre définition de l'estimateur* $\hat{m}(x)$ *introduit dans (2.6.2) découlant de la forme de l'expression du noyau associé bêta est donné par :*

$$\hat{m}(x) = \frac{\sum\limits_{i=1}^{n} Y_i X_i^{\frac{x}{h}} (1-X_i)^{\frac{1-x}{h}}}{\sum\limits_{i=1}^{n} X_i^{\frac{x}{h}} (1-X_i)^{\frac{1-x}{h}}}. \tag{2.36}$$

2.6.5 Propriétés de l'estimateur

D'une manière analogue aux démarches effectuées dans l'article de [10], dans cette partie nous proposons des expressions asymptotiques des propriétés de l'estimateur

présenté dans la section 2.6.4.

La proposition suivante est utilisée dans les calculs des propriétés de l'estimateur

Proposition 2.1. *On considère $l(u)$ une fonction telle que sa deuxième dérivée est continue et bornée sur $(0, +\infty)$, alors pour tout $x \in [0, 1]$ et $k \geq 0$:*

$$\int_0^1 B(u, p_k, \lambda_k) l(u) du = l(x) + \frac{[2l'(x) + xl''(x)]h}{2k} + o(h).$$

Démonstration 2.6.1. *Voir l'annexe (3.5).*

2.6.5.1 Biais de l'estimateur

Proposition 2.2. *On suppose que $(S_1), (S_2), (S_3), (S_4)$ (définies dans 3.5) sont vérifiées. Alors, pour tout $x \in [0, 1]$ avec $f(x) > 0$:*

1. *Pour $x > 0$:*
$$biais(\hat{m}(x)) = hb(x) + o(h) + o(\frac{h^{\frac{1}{4}}}{\sqrt{n}}), \tag{2.37}$$
2. *Pour $x = 0$*
$$biais(\hat{m}(0)) = hm'(0) + o(h). \tag{2.38}$$

Démonstration 2.6.2. *On concidère une décomposition comme formulée dans (2.14) pour la différence $\hat{m}(x) - m(x)$ où $\hat{m}(x)$ est l'estimateur à noyau associé bêta.*
D'après (2.15), et de plus la variance d'une variable aléatoire est bornée par son moment d'ordre 2, alors :

$$\mathbb{V}ar(B_n(x)) \leq \frac{1}{n}\mathbb{E}\Big[\frac{X^{\frac{x}{h}}(1-X)^{\frac{1-x}{h}}}{\beta(\frac{x}{h}+1, \frac{(1-x)}{h}+1)}[m(X) - m(x)]\Big]^2 \tag{2.39}$$

$$\begin{aligned}\frac{1}{n}\mathbb{E}\Big[\frac{X^{\frac{x}{h}}(1-X)^{\frac{1-x}{h}}}{\beta(\frac{x}{h}+1, \frac{(1-x)}{h}+1)}[m(X) - m(x)]\Big]^2 &= \frac{1}{n}\int_0^1 \frac{u^{\frac{2x}{h}}(1-u)^{\frac{2(1-x)}{h}}}{\beta^2(\frac{x}{h}+1, \frac{(1-x)}{h}+1)}[m(u) - m(x)]^2 f(u) du \\ &= \frac{\beta(\frac{2x}{h}+1, \frac{1-x}{h}+1)}{n\beta^2(\frac{x}{h}+1, \frac{1-x}{h}+1)}\int_0^1 K_{p_2,\lambda_2}(u) f(u)[m(u) - m(x)]^2.\end{aligned}$$

D'après l'approximation énoncée et démontrée dans [19], pour un h trés petit :

$$\frac{1-x}{h} \to \infty, \frac{x}{h} \to 0,$$

alors :

$$\frac{\beta(\frac{2x}{h}+1, \frac{2(1-x)}{h}+1)}{\beta^2(\frac{x}{h}+1, \frac{(1-x)}{h}+1)} \sim \frac{1}{2\sqrt{\pi x(1-x)h}}.$$

D'après cette dérnière approximation, pour $x > 0$ *:*

$$\frac{\beta(\frac{2x}{h}+1, \frac{1-x}{h}+1)}{n\beta^2(\frac{x}{h}+1, \frac{1-x}{h}+1)} = \frac{1}{2n\sqrt{\pi x(1-x)h}}[1+o(1)] \tag{2.40}$$

La quantité $\int_0^1 K_{p_2,\lambda_2}(u)f(u)[m(u)-m(x)]^2du$ *est finie et d'après la continuité des fonctions* f, m, f', m' *ainsi que la supposition* (S_2) *:*

$$\int_0^1 K_{p_2,\lambda_2}(u)f(u)[m(u)-m(x)]^2du = \frac{xf(x)m'(x)h}{2} + o(h). \tag{2.41}$$

La composition de (2.40) et (2.41) implique :

$$\mathbb{V}ar(B_n(x)) = o(\frac{\sqrt{h}}{n}),$$

cette dérnière combinée avec l'expression (2.15), implique :

$$biais(\hat{m}(x)) = hb(x) + o(h) + o(\frac{h^{\frac{1}{4}}}{\sqrt{n}}).$$

■

2.6.5.2 Variance de l'estimateur

Proposition 2.3. *On suppose que les hypothèses* $(S_1), (S_2), (S_3), (S_4)$ *(définies dans 3.5) sont vérifiées. Alors, pour tout* $x \in [0,1]$ *avec* $f(x) > 0$ *:*

1. *Pour* $x > 0$
$$\mathbb{V}ar(\hat{m}(x)) = \frac{\sigma^2(x)}{f(x)2\sqrt{hx(1-x)\pi}} + o(\frac{1}{n\sqrt{h}}). \tag{2.42}$$
2. *Pour* $x = 0$ *:*
$$\mathbb{V}ar(\hat{m}(0)) = \frac{\sigma^2(0)}{f(0)2nh} + o(\frac{1}{nh}). \tag{2.43}$$
3. *Pour* $x = 1$ *:*
$$\mathbb{V}ar(\hat{m}(1)) = \frac{\sigma^2(1)}{f(1)2nh} + o(\frac{1}{nh}). \tag{2.44}$$

Démonstration 2.6.3. *D'une manière similaire à celle énoncée précédemment dans le cas d'un noyau gamma, la variance de l'estimateur* $\hat{m}(x)$ *défini dans (2.6.2) vaut :*

$$\begin{aligned}\mathbb{V}ar(\hat{m}(x)) &= \frac{1}{n^2\hat{f}^2(x)}\sum_{i=1}^{n} K^2_{(\frac{x}{h},h)}(X_i)\mathbb{V}ar(\epsilon_i) \\ &= \frac{1}{n^2\hat{f}^2(x)\beta^2(\frac{x}{h}+1, \frac{1-x}{h}+1)}\sum_{i=1}^{n} X_i^{\frac{2x}{h}}\Big(1-X_i\Big)^{\frac{2(1-x)}{h}}\sigma^2(X_i).\end{aligned}$$

Comme énoncé dans (2.28) :

$$\mathbb{E}[\mathbb{V}ar(\hat{m}(x))] = \mathbb{V}ar(\hat{m}(x))$$

$$\mathbb{E}[\mathbb{V}ar(\hat{m}(x))] = \frac{1}{(\hat{f}(x))^2}\mathbb{E}\left[\frac{1}{n^2\beta^2(\frac{x}{h}+1,\frac{1-x}{h}+1)}\sum_{i=1}^{n} X_i^{\frac{2x}{h}}\left(1-X_i\right)^{\frac{2(1-x)}{h}}\sigma^2(X_i)\right]$$

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n^2\beta^2(\frac{x}{h}+1,\frac{1-x}{h}+1)}\sum_{i=1}^{n} X_i^{\frac{2x}{h}}\left(1-X_i\right)^{\frac{2(1-x)}{h}}\sigma^2(X_i)\right]\\
&= \frac{1}{n\beta^2(\frac{x}{h}+1,\frac{1-x}{h}+1)}\mathbb{E}\left[X_i^{\frac{2x}{h}}\left(1-X_i\right)^{\frac{2(1-x)}{h}}\sigma^2(X_i)\right]\\
&= \frac{\beta\left(\frac{2x}{h},\frac{2(1-x)}{h}\right)}{n\beta^2(\frac{x}{h}+1,\frac{1-x}{h}+1)}\int_0^1 \frac{t^{\frac{2x}{h}}\left(1-t\right)^{\frac{2(1-x)}{h}}}{\beta\left(\frac{2x}{h},\frac{2(1-x)}{h}\right)}\sigma^2(t)f(t)dt\\
&= \frac{\beta\left(\frac{2x}{h},\frac{2(1-x)}{h}\right)}{n\beta^2(\frac{x}{h}+1,\frac{1-x}{h}+1)}\int_0^1 K_{(\frac{2x}{h}+1,\frac{2(1-x)}{h}+1)}(t)\sigma^2(t)f(t)dt,
\end{aligned}$$

à travers l'application du lemme (2.5.1) avec $l = \sigma^2 \times f$ et d'après l'approximation qui figure dans [20] et d'après (S_3) pour $x \in]0,1[$:

$$\frac{1}{n^2\beta^2(\frac{x}{h}+1,\frac{1-x}{h}+1)}\mathbb{E}\left[X_i^{\frac{2x}{h}}\left(1-X_i\right)^{\frac{2(1-x)}{h}}\sigma^2(X_i)\right] = \frac{\sigma(x)f(x)}{2\sqrt{hx(1-x)\pi}} + o\left(\frac{\sqrt{h}}{n}\right). \quad (2.45)$$

De plus, d'après [18] : $\hat{f}(x) = f(x) + o(1)$, alors la variance de l'estimateur vaut :

$$\mathbb{V}ar(\hat{m}(x)) = \frac{\sigma^2(x)}{2f(x)\sqrt{hx(1-x)\pi}} + o(\frac{1}{n\sqrt{h}}).$$

Pour $x = 0$:

D'après (2.45)

$$\frac{1}{n\beta^2(1,\frac{1}{h}+1)}\mathbb{E}\Big[\left(1-X\right)^{\frac{1}{h}}\sigma^2(X)\Big] = \frac{\sigma(0)^2 f(0)}{2nh} + o(\frac{1}{n}),$$

de plus, d'apès : $\hat{f}(0) = f(0) + o(1)$:

$$\mathbb{V}ar(\hat{m}(0)) = \frac{\sigma(0)^2}{2nhf(0)} + o(\frac{1}{nh}).$$

Pour $x = 1$:

D'après (2.45)

$$\frac{1}{n\beta^2(\frac{1}{h}+1,1}\mathbb{E}\Big[\left(X\right)^{\frac{1}{h}}\sigma^2(X)\Big] = \frac{\sigma(1)^2 f(1)}{2nh} + o(\frac{1}{n}),$$

de plus, d'apès : $\hat{f}(1) = f(1) + o(1)$:

$$\mathbb{V}ar(\hat{m}(1)) = \frac{\sigma(1)^2}{2nhf(1)} + o(\frac{1}{nh}).$$

■

Erreur moyenne quadratique

Propriété 3. *Á partir des expressions du biais et de la variance de l'estimateur à noyau gamma défini dans (2.5.5.1) et (2.5.5.2) l'erreur moyenne quadratique de l'estimateur* $MSE(\hat{m}(x))$ *pour* $x \in [0,1]$*, avec* $f(x) > 0$ *vaut :*

— *Pour* $x \in]0,1[$

$$MSE(\hat{m}(x)) = h^2[m'(x)+\frac{1}{2}xm''(x)+\frac{xm'(x)f'(x)}{f(x)}]^2+\frac{\sigma^2(x)}{2f(x)\sqrt{\pi xnh}}+o(h^2)+o(\frac{1}{n\sqrt{h}})+o(\frac{h^{\frac{5}{4}}}{\sqrt{n}}), \tag{2.46}$$

— *Pour* $x = 0$ *:*

$$MSE(\hat{m}(0)) = h^2 m'^2(0) + \frac{\sigma^2(0)}{2nhf(0)} + o(h^2) + o(\frac{1}{nh}). \tag{2.47}$$

— *Pour* $x = 1$ *:*

$$MSE(\hat{m}(1)) = h^2 m'^2(1) + \frac{\sigma^2(1)}{2nhf(1)} + o(h^2) + o(\frac{1}{nh}). \tag{2.48}$$

2.7 Conclusion

Dans ce chapitre, nous avons abordé l'estimation de la fonction de régression par la méthode du noyau dans le cas d'un noyau associé asymétrique continu. En premier lieu, nous avons défini les conditions d'un noyau associé asymétrique. Par la suite, nous avons fourni des exemples sur ces dérniers ceci à travers des définitions de leurs expressions analytiques. Enfin, nous avons présenté les propriétés de cet estimateur dans deux cas particuliers de noyaux associés en l'occurrence le noyau gamma et le noyau bêta.

CHAPITRE 3

ÉVALUATION DE PERFORMANCE DE L'ESTIMATEUR : DONNÉES SIMULÉES ET DONNÉES RÉELLES

3.1 Introduction

Dans les chapitres précédents, nous avons abordé des notions théoriques sur l'estimateur à noyau associé asymétrique de la fonction de régression et les propriétés présentées n'ont été définies qu'à travers leurs expressions analytiques. Pour l'évaluation de performances empiriques et des valeurs quantitatives des propriétés nous utilisons un échantillon d'observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, ce dernier est généralement soit fournit par une simulation via un programme informatique soit il correspond à un ensemble de données réelles.

Dans ce chapitre nous présentons une évaluation de la qualité de l'estimateur à noyau associé asymétrique de la fonction de régression sur la base d'un échantillon simulé en utilisant le langage de programmation R et sur un jeu de données réelles.

3.2 Algorithme de simulation

L'algorithme élaboré dans l'étude de la simulation est le suivant :

— Simulation d'un n-échantillon d'une variable aléatoire X à valeurs contenues dans un intervalle borné.

— Simulation d'un échantillon de résidus ϵ_i tel que

$$\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2), \forall i \in \{1, 2, ..., n\}.$$

— Calcul d'un vecteur de réponses Y où :

$$Y_i = m(X_i) + \epsilon_i.$$

— Calcul du paramètre de lissage h via la méthode de la validation croisée. Cette dérnière consiste à minimiser la quantité $CV(h)$ définie par l'expression suivante :

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(Xi))^2 w(X_i),$$

où :

$$\hat{m}_{-i}(Xi) = \frac{\sum\limits_{j=1, j\neq i}^{n} Y_j K_{X_i,h}(X_j)}{\sum\limits_{j=1, j\neq i}^{n} K_{X_i,h}(X_j)},$$

et :

$$w_x(X_i) = \frac{K_{x,h}(X_i)}{\sum\limits_{i=1}^{n} K_{x,h}(X_i)}.$$

— Estimation de la fonction $m(x)$ par la méthode du noyau pour un certain noyau associé $K_{x,h}(.)$, cet estimateur est défini par :

$$\hat{m}(x) = \frac{\sum\limits_{i=1}^{n} K_{x,h}(X_i) Y_i}{\sum\limits_{i=1}^{n} K_{x,h}(X_i)},$$

— Évaluation de performance de l'estimateur.

3.3 Étude sur des modèles cibles

• Pour rappel, le modèle de régression est défini par :

$$y_i = m(x_i) + \epsilon_i,$$

où $m(.)$ désigne le modèle de régression.

• Le critère sur lequel se base l'évaluation de performances de l'estimateur pour les

noyaux utilisés est le critère du risque quadratique RQ_{moy}.

Soit $RQ(i)$ la valeur du risque quadratique de la i-ème simulation telle que :

$$RQ(i) = \frac{1}{n}\sum_{j=1}^{n}(\hat{m}(x_i) - m(x_i))^2,$$

alors :

$$RQ_{moy} = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}} RQ(i),$$

− où n_{sim} est le nombre de simulations (dans notre cas ($n_{sim} = 100$)).

• Nous avons estimé deux modèles :

1. Le modèle $m_1(x) = x^2$.
2. Le modèle $m_2(x) = sin(x)$.

• Les quatres noyaux utilisés sont les noyaux associés asymétrique : bêta, gamma, lognormal et le noyau particulier BS de la famille de noyaux GBS.

- Les valeurs de l'échantillon sont comprises entre 0 et 1,
- La variance des résidus σ^2 vaut 0.005^2,
- Les différentes tailles n des échantillons sont : $n \in \{50, 100, 200, 500\}$.

Premier modèle

Les résutats de simulation concernant le critère RQ_{moy} du premier modèle $m_1(x)$ sont représentés dans le tableau (3.1) :

Taille de l'échantillonNoyau	bêta	gamma	lognormal	BS
n=50	0.0006469069	0.003430479	0.003071083	0.0002681139
n=100	0.0001918558	0.001392282	0.001079179	0.0001057866
n=200	0.0000495805	0.0007185956	0.0005507837	0.00009857245
n=500	0.00001312625	0.0005548001	0.0003841495	0.00008396372

TABLE 3.1: Valeurs des risques RQ_{moy} du modèle $m_1(x)$

La diminution du risque traduit le fait que la valeur de l'estimateur se rapproche de la valeur de la fonction qu'il estime.

De plus, la quasi-totalité des valeur RQ_{moy} sont d'ordre $\leq 10^{-3}$ et sont par conséquent relativement petites.

• Les figures (3.1) et (3.2) illustrent la courbe de régression du modèle $m_1(x)$:

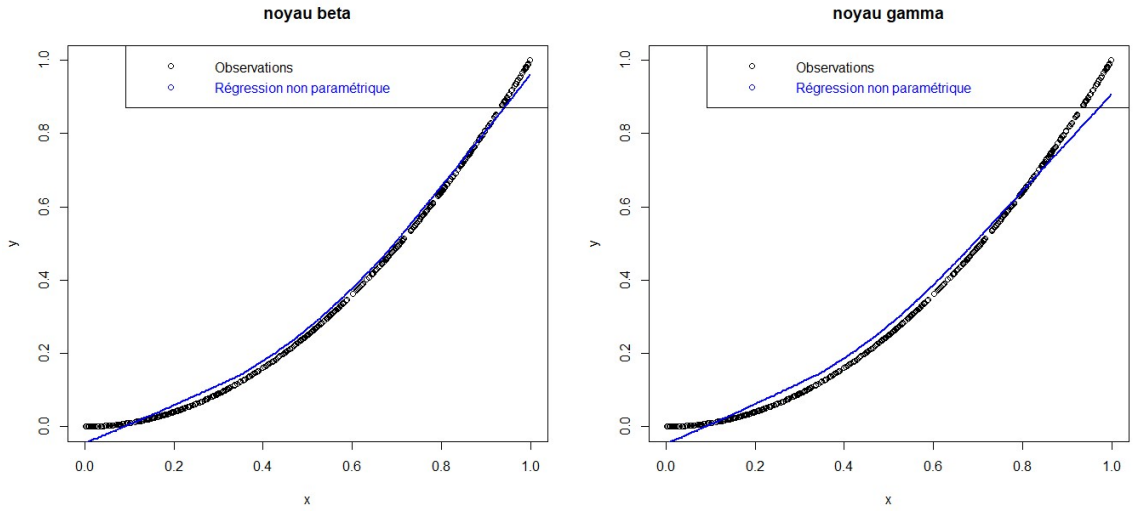

FIGURE 3.1: Estimation du modèle $m_1(x)$ par noyaux associés asymétrique : beta, gamma

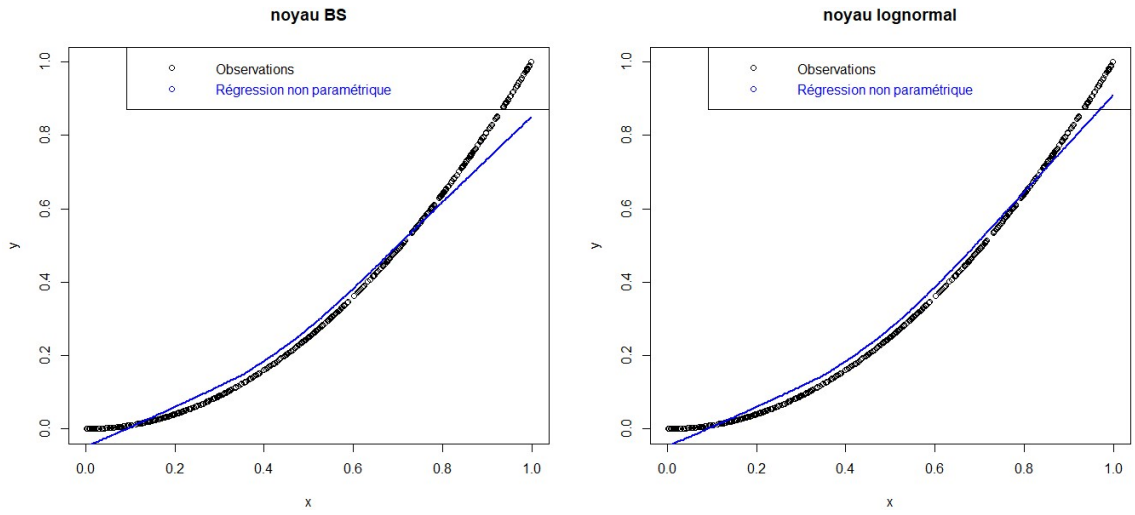

FIGURE 3.2: Estimation du modèle $m_1(x)$ par noyaux associés asymétrique : lognormal et BS

Disscussion des figures

Il ressort des représentations précédentes que les allures des estimateurs à noyau sont tous en concordance avec l'allure du modèle estimé.

D'autre part, les représentations graphiques confirment les résultats obtenus dans la simulation qui traduisent la faiblesse des risques RQ_{moy}.

Deuxième modèle

Pour le deuxième modèle, les résulats de simulation pour le critère RQ se présentent comme suit :

Taille de l'échantillonNoyau	bêta	gamma	lognormal	BS
n=50	0.0003674052	0.00104837	0.0006739532	0.0001121986
n=100	0.000103074	0.000327354	0.0001851995	0.00003429091
n=200	0.00002834107	0.0001333423	0.00008543256	0.00001993714
n=500	0.000007870156	0.00008739927	0.00005958484	0.00001289596

TABLE 3.2: Valeurs des risques RQ_{moy} du modèle m_2

Interprétation des résultats

En premier lieu, d'après les résultats obtenus pour le modèle $m_2(.)$ et qui figurent dans le tableau (3.2), nous constatons que les risques RQ_{moy} sont tous petits et ils sont d'ordre $\leq 10^{-3}$.

En second lieu, nous remarquons la décroissance des valeurs des risques avec l'augmentation de la taille de l'échantillon pour lesquelles nous observons de bon résultats de risque à l'instar du $RQ_{moy} = 0.000007870156$ obtenu pour le noyau bêta avec une taille d'échantillon $n = 500$.
Les différentes droites de régression du deuxième modèle sont illustrées dans les figures (3.3) et (3.4) :

Discussion des figures

D'après les figures, nous remarquons que les estimateurs à noyaux utilisés fonctionnent mieux pour le modèle $m_2(x)$ que pour le modèle $m_1(x)$. En effet, les droites de régression de la figure (3.3) sont plus lisses que celles de la figure(3.1).

D'autre part, nous constatons un léger décalage entre les observations et la droite de régression pour l'estimateur à noyau BS comparé aux autres estimateur.

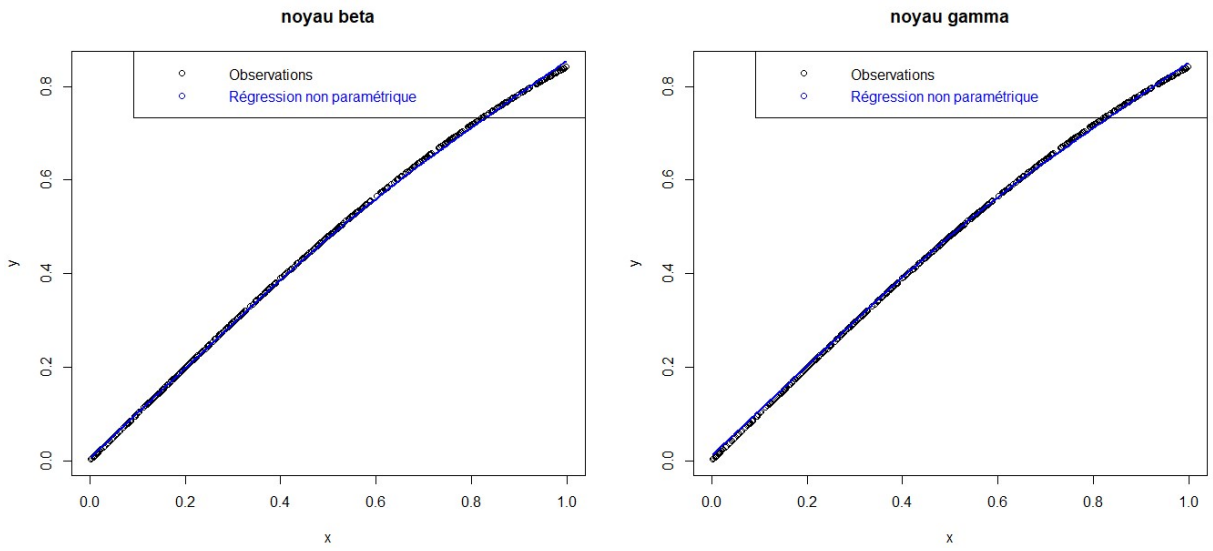

FIGURE 3.3: Estimation du modèle $m_2(x)$ par noyaux associés asymétriques : bêta, gamma

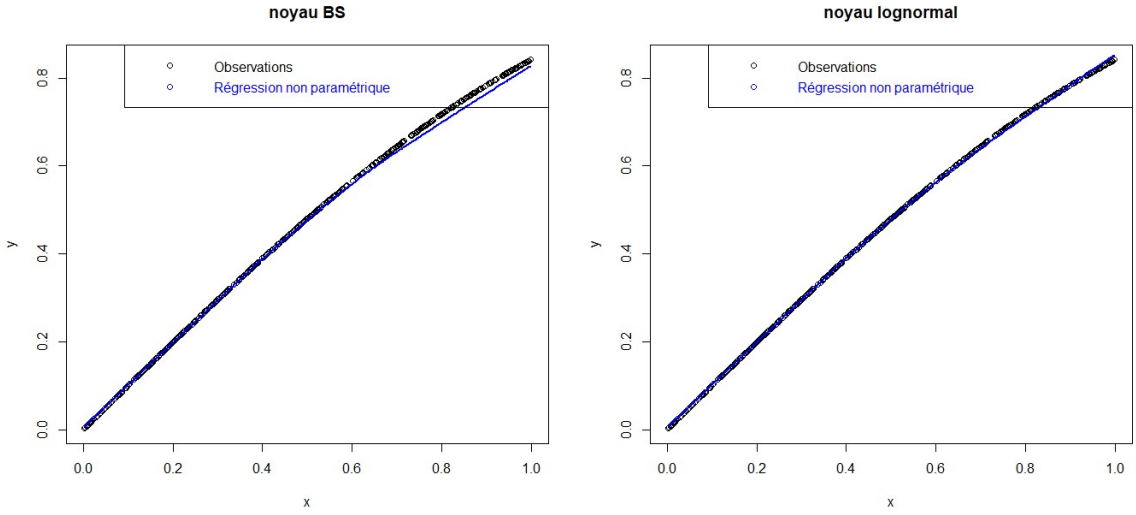

FIGURE 3.4: Estimation du modèle $m_2(x)$ par noyaux associés asymétriques : lognormal et BS

3.4 Application sur des données réelles

Dans cette partie nous nous sommes orientés vers l'application de l'estimateur de la fonction de régression asymétrique défini dans le deuxième chapitre dans le domaine économique.

En économie, un indicateur est une statistique construite dans le but de mesurer certaines dimensions de l'activité économique d'une manière aussi objective que possible.

Parmi les indicateurs économiques nous trouvons les indicateurs financiers qui sont de plus en plus en usage avec la mondialisation financière.
L'indice de développement financier noté "**FD**" constitue une mesure de développement du secteur financier dans certains pays.

L'application de la régression non paramètrique consiste à vérifier la présence d'une relation entre le **FD** et l'accès aux services financiers **FIA** de l'Algérie via la méthode du noyau, plus précisément vérifier si l'indice du développement financier **FD** s'écrit en fonction de l'accès aux services financiers **FIA**. Dans ce cas nous avons employé un échantillon de taille 38 parmi l'ensemble des données collecté par le fond monétaire international de l'année 1974 jusqu'à l'année 2015. Ces données sont représentées dans les tableaux (3.3).

FD	0.1310486	0.1348543	0.1320744	0.1301827	0.1298875	0.1282757	0.1247461
FIA	0.1009489	0.100522	0.0989665	0.0964861	0.0955668	0.0917309	0.0868861
FD	0.12618	0.1272455	0.130252	0.12755353	0.128516	0.1204121	0.1192983
FIA	0.0856437	0.0846795	0.0825044	0.0810702	0.076817	0.0727284	0.0654846
FD	0.1088851	0.1125936	0.1196124	0.1171723	0.1121913	0.1200108	0.1071191
FIA	0.0590569	0.058081	0.0675241	0.0663581	0.064951	0.0737471	0.07791
FD	0.1188971	0.1124934	0.1084377	0.119165	0.1192054	0.1163664	0.1196466
FIA	0.0755637	0.0731689	0.0714493	0.067376	0.0714727	0.0562532	0.0634995
FD	0.1203124	0.1276011	0.1335829	0.137261	0.1349036	0.1357694	0.133722
FIA	0.634995	0.0788239	0.0887352	0.0958962	0.0978652	0.0979771	0.0961132

FD	0.1328546	0.1312911	0.1293476	0.1228451	0.1199189
FIA	0.0942536	0.0909221	0.0868274	0.0733393	0.0676444

TABLE 3.3: Données financières collectées par le fond monétaire international de 1974 à 2015

Application de la régression non paramètrique

Pour l'estimation de la fonction $m()$ nous utilisons l'estimateur à noyau associé asymétrique avec le noyau bêta. Pour le paramètre de lissage nous employons la méthode

de la validation croisée.

Pour l'évaluation de la qualité de l'estimateur nous avons utilisé le coefficient de détermination R^2 et le RMSE définis respectivement par :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \times 100,$$

où : $\hat{y}_i = \hat{m}(x_i)$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}.$$

Remarque 3.4.1. *Plus la valeur du $RMSE$ est proche de zéro plus le modèle évalué est de meilleur qualité [21].*

Les résultats des critères mentionnés précédemment sont les suivants :

$R^2(\%)$	RMSE
92	0.00395929

Discussion

Dans cette partie nous avons proposé de cristalliser la relation entre Y "l'indice du développement financier (**FD**)" (variable expliquée) et X "l'accès aux services financier (**FIA**)"(variable explicative) via une régression non paramétrique. Cette dernière, consiste en l'estimation de cette relation par l'estimateur de la fonction de régression à noyau associé asymétrique et continue car les valeurs sont toutes non négatives et appartiennent à un domaine continu. De plus, les valeurs de l'indice **FIA** sont comprises entre $[0, 1]$, nous avons alors choisi d'appliquer le noyau bêta.

Le calcul du coefficient de détermination R^2 indique une valeur élevée traduisant le fait que le modèle de régression interprète bien le modèle réel. De plus, la valeur du $RMSE$ est petite ce qui indique une bonne qualité de l'estimateur.

En dernier lieu, une représentation graphique (3.5) illustre que le modèle de régression proposé se rapproche bien du modèle réel.

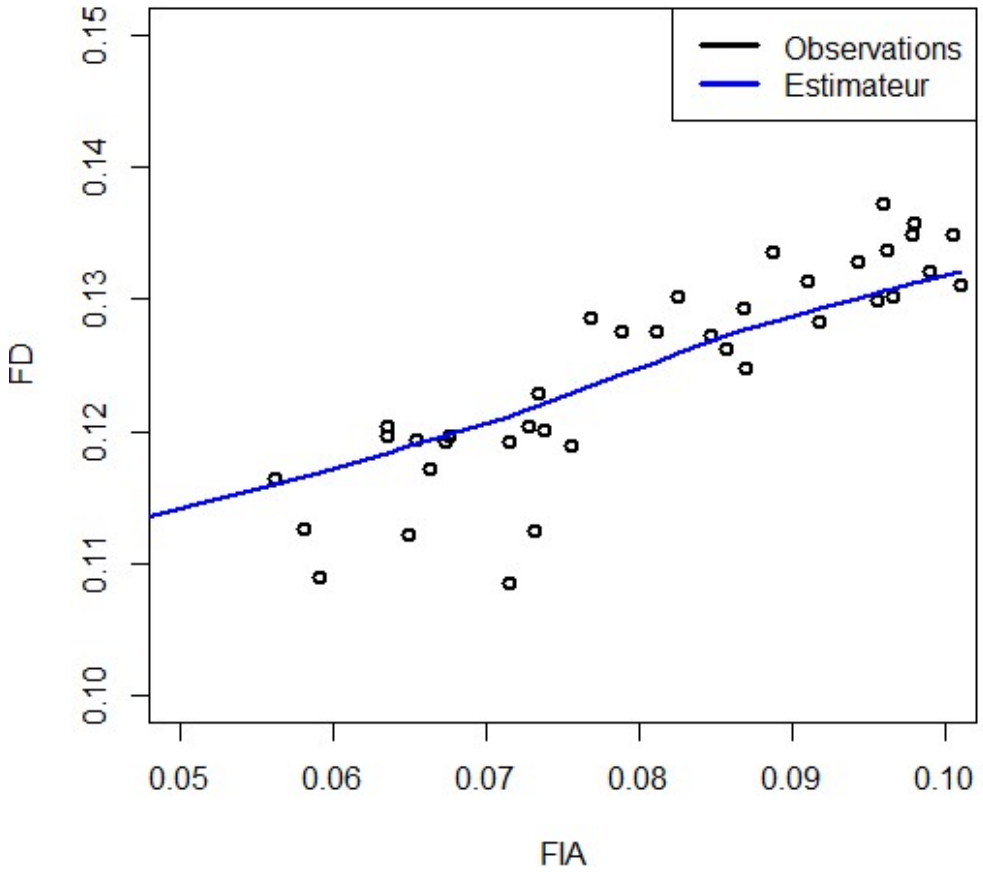

FIGURE 3.5: Régression non paramétrique en utilisant le noyau beta

3.5 Conclusion

Dans ce chapitre, nous avons évalué la qualité de l'estimateur à noyau associé asymétrique et continu de la fonction de régression. Cette évaluation a été effectuée d'une part sur la base d'un échantillon simulé, avec lequel nous avons discuté deux modèles cibles $m_1(x)$ et $m_2(x)$; d'autre part sur un jeu de données réelles.

Pour la partie simulation nous avons effectué l'évaluation de l'estimateur en utilisant les quarte noyaux : bêta, gamma, lognormal et le noyau BS. Ces derniers fonctionnent assez bien pour les deux modèles cibles.

Pour la partie application sur des données réelles, comme les données sont comprises entres 0 et 1 nous avons alors évalué la qualité de l'estimateur en utilisant le noyau bêta. Le calcul des critères d'évaluation, à savoir l'erreur $RMSE$ et le coefficient de détermination R^2, indique une bonne qualité de l'estimateur. Par ailleurs, la courbe de régression obtenue confirme bien les résultats numériques.

Nous pouvons conclure que l'estimateur à noyau associé asymétrique de la fonction

de régression proposé dans le chapitre 2 fonctionne bien dans son intégralité.

CONCLUSION GÉNÉRALE

La régression représente un procédé statistique qui est utilisé pour l'analyse d'une relation entre deux v.a X et Y. Les modèles paramétriques de régression tels que les modèles linéaires sont fréquemment utilisés. Néanmoins, ces modèles ne sont pas toujours applicables et nous disposons comme alternative les modèles non paramétriques de régression.

La régression non paramétrique consiste à estimer sous la base d'un échantillon une relation entre deux v.a X et Y sans aucune spécification de la forme de cette relation. Il existe plusieurs méthodes d'estimation non paramétrique de la fonction de régression, dans notre travail nous avons présenté la méthode du "**noyau**".

L'estimateur de Nadaraya [6] et Watson [7] est un estimateur non paramétrique de la fonction de régression qui utilise la méthode du noyau. L'utilisation de cet estimateur avec un noyau symétrique dans le cas d'un ensemble de données borné présente un problème au niveau des bornes de l'intervalle auquel appartiennent les données, car l'estimateur assigne des poids en dehors de l'intervalle [10], il s'agit du problème de bornes. Plusieurs travaux se sont intéréssés à l'étude de l'estimateur à noyau pour des données non négatives. Par exemple ; dans l'article [10] les auteurs ont développé les propriétés asymptotiques de l'estimateur à noyau gamma de la fonction de régression.

Dans le présent travail nous avons étudié l'estimateur de la fonction de régression par la méthode des noyaux asymétriques dans le cas univarié et pour un support de données continu.

Nous avons d'abord abordé l'estimateur de la fonction de régression dans le cas

symétrique.
Ensuite, nous avons étudié cet estimateur dans le cas asymétrique pour lequel les propriétés ont été analysées en particulier pour les deux noyaux "beta" et "gamma".
Enfin, nous avons évalué la qualité de l'estimateur via une application numérique en utilisant un échantillon de données simulées et un jeu de données réelles. Dans la simulation nous avons évalué la qualité de l'estimateur sur deux modèles cibles. Le critère pris pour l'évaluation est le risque quadratique RQ. Les résultats obtenus montrent que l'estimateur est de bonne qualité. Ceci a été confirmé à travers une illustration graphique de la régression non paramétrique. En effet, d'après les figures nous avons constaté que les droites de régression non paramétrique sont en concordance avec la représentation graphique des modèles cibles estimés. En ce qui concerne l'application sur les données réelles, l'évaluation de la qualité s'est portée sur les deux critères suivants : le coefficient de détermination $R^2(\%)$ et l'erreur $RMSE$. Au terme de cette application, nous avons obtenu une valeur élevée pour le coefficient R^2 et une valeur faible de l'ordre 10^{-3} pour le $RMSE$. Ces résultats indiquent que l'estimateur proposé est performant. Ceci a été également confirmé à partir d'une représentation graphique.

Enfin, nous avons enregistré des résultats traduisant la bonne qualité de l'estimateur, ceci pour les données simulés ainsi que les données réelles. D'après ces résultats il ressort que l'estimateur de la fonction de régression par la méthode des noyaux asymétriques est de bonne qualité.

ANNEXE

Démonstration de la proposition 2.1

On concidère $b(u,p,\lambda)$ la densité de probabilité d'une variable alétoire suivant une loi bêta de paramètrres : p,λ définie pour tout $u \in [0,1]$ par :

$$b(u,p,\lambda) = \frac{u^p(1-u)^\lambda}{\beta(p,\lambda)},$$

la varaince d'un telle variable aléatoire vaut $\tau^2 = \frac{p\lambda}{(p+\lambda)^2(p+\lambda+1)}$ et son espérance quant à elle vaut $\frac{p}{p+\lambda}$. On concidère $p_k = \frac{kx}{h} + 1$ et $\lambda_k = \frac{k(1-x)}{h} + 1$; τ_k^2 et μ_k représentent respectivement des réecritures de τ^2 et μ en remplaçant respectivement p et λ par p_k et λ_k, telles que :

$$\tau_k^2 = \frac{\left(\frac{kx}{h}+1\right)\left(\frac{k(1-x)}{h}+1\right)}{\left(\frac{k}{h}+2\right)^2\left(\frac{k}{h}+3\right)},$$

$$\mu_k = \frac{kx+h}{h(\frac{k}{h}+2)}.$$

Pour un $x \geq 0$ fixé, $\mu_k = \frac{kx+h}{k+2h}$; l'expression de taylor de $l(\mu_k)$ au voisinage de 0 à l'ordre 2 est la suivante :

$$l(\mu_k) = l(x) + \frac{hl'(x)}{k} + \frac{h^2 l''(\tilde{\xi})}{2k^2}, \tag{3.1}$$

où $\tilde{\xi}$ est une valeur comprise entre μ_k et x. L'expression de taylor au centre μ_k de $B(u,p_k,\lambda_k)$ est donnée par :

$$\int_0^1 B(u,p_k,\lambda_k)l(u)du = l(\mu_k) + \frac{1}{2}\int_0^1 (u-\mu_k)^2 B(u,p_k,\lambda_k)l''(\tilde{u})du, \tag{3.2}$$

où pour tout $u \in [0,1]$, $\tilde{u}$ représente une valeur comprise entre u et μ_k.

Pour $x = 0$:

$$\int_0^1 (u-\mu_k)^2 B(u,p_k,\lambda_k)du = \tau_k^2 = \frac{\left(\frac{kx}{h}+1\right)\left(\frac{k(1-x)}{h}+1\right)}{\left(\frac{k}{h}+2\right)^2\left(\frac{k}{h}+3\right)},$$

d'après la bornétude de la fonction $l''(.)$:

$$|\int_0^1 (u-\mu_k)^2 B(u,p_k,\lambda_k) l''(\tilde{u})du \,| \leq c\tau_k^2 = o(h), \tag{3.3}$$

pour quelques valeurs positives de la constantes c. Ceci combiné avec l'expression (3.1) implique la vérification de la proposition (2.1) pour $x = 0$.

D'autre part, une réecriture de (3.2) est présentée sous la forme suivante :

$$\int_0^1 B(u,p_k,\lambda_k)l(u)du = l(\mu_k)\frac{1}{2}l''(\mu_k)\int_0^1 (u-\mu_k)^2 B(u,p_k,\lambda_k)du + \frac{1}{2}\int_0^1 (u-\mu_k)^2 B(u,p_k,\lambda_k)[l''(\tilde{u})-l''(\mu_k)] \tag{3.4}$$

La première intégrale du coté droit correspond à la variance $B(u,p_k,\lambda_k)$ qui est

$$\tau_k^2 = \frac{\left(\frac{kx}{h}+1\right)\left(\frac{k(1-x)}{h}+1\right)}{\left(\frac{k}{h}+2\right)^2\left(\frac{k}{h}+3\right)},$$

d'après (3.1) et la continuité de la fonction $l''(.)$ on peut vérifier que les deux premiers termes du côté droit de l'expression (3.4) vérifient l'expression (2.1).

D'après la continuité de la fonction $l''(.)$, alors celle-ci est uniformément continue sur tout sous-intervale borné dans $(0,1)$.

Pour tout $\epsilon > 0$, on concidère $0 < \gamma < x$ tel que pour tout y avec $|y-x| \leq \gamma$, $|l''(x) - l''(y)| < \epsilon$. On concidère $\delta_1 = x - \gamma/2$, d'après la bornétude de $l''(.)$:

$$|\int_0^{\delta_1} (u-\mu_k)^2 B(u,p_k,\lambda_k)[l''(\tilde{u}) - l''(\mu_k)]du| \leq c\int_0^{\delta_1} (u-\mu_k)^2 B(u,p_k,\lambda_k)du.$$

Notons que la densité beta dans ce cas ($pk = \frac{kx}{h} + 1 > 1, \lambda_k = \frac{k(1-x)}{h} + 1 > 1$) est unimodal où le mode m_{beta} vaut :

$$m_{beta} = \frac{p_k - 1}{p_k + \lambda_k - 2} = x,$$

alors $B(u,p_k,\lambda_k) \leq B(\delta_1,p_k,\lambda_k)$, pour tout $0 < u < \delta_1$ et

$$\int_0^1 (u-\mu_k)^2 B(u,p_k,\lambda_k)du \leq B(\delta_1,p_k,\lambda_k)\int_0^1 \left(u - \frac{kx+h}{h\left(\frac{k}{h}+2\right)}\right)^2 du = B(\delta_1,p_k,\lambda_k)o(1).$$

Pour un h assez petit : d'après l'approximation de stirling :

$$\beta(p_k,\lambda_k)\sim\sqrt{\pi}\frac{p_k^{p_k-\frac{1}{2}}\lambda_k^{\lambda_k-\frac{1}{2}}}{(p_k+\lambda_k)^{p_k+\lambda_k+\frac{1}{2}}},$$

nous avons alors :

$$\frac{1}{\beta(p_k,\lambda_k)}=\frac{\left(\frac{k}{h}+2\right)^{\frac{k}{h}+\frac{3}{2}}}{\left(\frac{kx}{h}+1\right)^{\frac{kx}{h}+\frac{1}{2}}\left(\frac{k(1-x)}{h}+h\right)}[1+o(1)], \tag{3.5}$$

dans ce cas :

$$B(\delta_1,p_k,\lambda_k)=\frac{\left(\frac{k}{h}+2\right)^{\frac{k}{h}+\frac{3}{2}}}{\left(\frac{kx}{h}+1\right)^{\frac{kx}{h}+\frac{1}{2}}\left(\frac{k(1-x)}{h}+h\right)}\delta_1^{\frac{kx}{h}}(1-\delta_k)^{\frac{k(1-x)}{h}}[1+o(1)],$$

$$B(\delta_1,p_k,\lambda_k)=o(h).$$

Nous avons $\delta_1<1$ et $(1-\delta_1)<1$ pour tout $\delta_1\in[0,1]$, alors :

$$|\int_O^{\delta_1}(u-\mu_k)^2B(u,p_k,\lambda_k)[l''(\tilde{u})-l''(\mu_k)]du|=o(h). \tag{3.6}$$

On concidère $\delta_2=x+\gamma_2$, la fonction $l''(.)$ est bornée alors :

$$|\int_{\delta_2}^1(u-\mu_k)^2B(u,p_k,\lambda_k)[l''(\tilde{u})-l''(\mu_k)]du|\leq c\int_{\delta_2}^1(u-\mu_k)^2B(u,p_k,\lambda_k)du,$$

mais,

$$\int_{\delta_2}^1(u-\mu_k)^2B(u,p_k,\lambda_k)du=\frac{1}{\beta(p_k,\lambda_k)}\int_{\delta_2}^1(u-\mu_k)^2u^{p_k-1}(1-u)^{\lambda_k-1};$$

le côté droit de cette dérnière expression est majoré par la quantité suivante :

$$\frac{1}{\beta(p_k,\lambda_k)}\int_{\delta_2}^1u^{p_k}(1-u)^{\lambda_k-1}.$$

En tant qu'une fonction dépendant de u, la fonction $u^{p_k}(1-u)^{\lambda_k-1}$ est croissante sur $[0,\frac{kx}{k+h}]$ et ddécroissante sur $[\frac{kx}{k+h},1]$, pour un h assez petit : $\delta_2\geq\frac{kx}{k+h}$. Alors, pour tout $u>\delta_2$:

$$u^{p_k}(1-u)^{\lambda_k-1}\leq\delta_2^{p_k}(1-\delta_2)^{\lambda_k-1},$$

alors, d'après (3.5) :

$$\int_{\delta_2}^1(u_{\mu_k})^2B(u,p_k,\lambda_k)du\leq\frac{\left(\frac{k}{h}+2\right)^{\frac{k}{h}+\frac{3}{2}}}{\left(\frac{kx}{h}+1\right)^{\frac{kx}{h}+\frac{1}{2}}\left(\frac{k(1-x)}{h}+h\right)}\int_{\delta_2}^1 1du[1+o(1)]=o(h).$$

$\delta_2 < 1$ et $(1-\delta_2) < 1$, alors :

$$|\int_{\delta_2}^{1}(u-\mu_k)^2 B(u,p_k,\lambda_k)[l''(\tilde{u})-l''(\mu_k)]du| = o(h), \tag{3.7}$$

$$\int_{\delta_1}^{\delta_2}(u-\mu_k)^2[l''(\tilde{u})-l''(\mu_k)]du = o(h),$$

d'après la continuité uniforme de la fonction $l''(.)$:

$$\int_{\delta_1}^{\delta_2}(u-\mu_k)^2 B(u,p_k,\lambda_k)|l''(\tilde{u})-l''(\mu_k)|du \leq \epsilon \int_{0}^{1}(u-\mu_k)^2 B(u,p_k,\lambda_k)du.$$

D'après la supposition : $|\tilde{u}-\mu_k| \leq |u-\mu_k| < \gamma$ et h assez petit. De plus,

$$\int_{0}^{1}(u-\mu_k)^2 B(u,p_k,\lambda_k)du = \tau_k^2 = o(h),$$

alors :

$$\int_{\delta_1}^{\delta_2}(u-\mu_k)^2 B(u,p_k,\lambda_k)|l''(\tilde{u})-l''(\mu_k)|du = \epsilon \times o(h), \tag{3.8}$$

la valeur de ϵ est quelconque et de plus d'après (3.6),(3.7), (3.8) :

$$\int_{0}^{1}(u-\mu_k)^2 B(u,p_k,\lambda_k)[l''(\tilde{u})-l''(\mu_k)]du = o(h),$$

pour $x > 0$.

Ceci en combinaison avec (3.3) complète la preuve de la proposition (2.1).

Une conclusion similaire peut être concidérée à travers la vérification de la condition de hölder par la seconde dérivée de la fonction $l''(.)$.

BIBLIOGRAPHIE

[1] N. N. CENCOV, "Estimation of an unknown distribution density from observations," *Soviet Math.*, t. 3, p. 1559-1566, 1962.

[2] N. SAADI et S. ADJABI, "On the estimation of the probability density by trigonometric series," *Communications in Statistics—Theory and Methods*, t. 38, n° 19, p. 3583-3595, 2009.

[3] A. SONIA, "Sur l'estimation de la courbe de régression de la moyenne.," thèse de doct., Universié de bejaia, 2011.

[4] M. ROSENBLATT, "Remarks on some nonparametric estimates of a density function," *The annals of mathematical statistics*, p. 832-837, 1956.

[5] E. PARZEN, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, t. 33, n° 3, p. 1065-1076, 1962.

[6] E. A. NADARAYA, "On estimating regression," *Theory of Probability & Its Applications*, t. 9, n° 1, p. 141-142, 1964.

[7] G. S. WATSON, "Serial correlation in regression analysis. I," *Biometrika*, t. 42, n° 3/4, p. 327-341, 1955.

[8] S. ADJABI, "Estimation de la courbe de régression de la moyenne par la méthode non paramétrique du noyau," *Séminaire Mathématique de Béjaia (LaMOS)*, t. 1, p. 37-40, 2003.

[9] Q. HE, "Consistency of the Priestley–Chao estimator in nonparametric regression model with widely orthant dependent errors," *Journal of Inequalities and Applications*, n° 1, p. 1-13, 2019.

[10] J. SHI et W. SONG, "Asymptotic results in gamma kernel regression," *Communications in Statistics-Theory and Methods*, t. 45, n° 12, p. 3489-3509, 2016.

[11] S. X. CHEN, "Beta kernel smoothers for regression curves," *Statistica Sinica*, p. 73-91, 2000.

[12] T. GASSER et H.-G. MÜLLER, "Kernel estimation of regression functions," in *Smoothing techniques for curve estimation*, Springer, 1979, p. 23-68.

[13] T. CACOULLOS, "Estimation of a multivariate density," *Inst. Statist. Math.*, t. 18, p. 179-189, 1966.

[14] D. BLONDIN, "Lois limites uniformes et estimation non-paramétrique de la régression," thèse de doct., Université Paris VI, 2004.

[15] K. LAGHA, "Estimation de la fonction regression par la methode du noyau. Proprietes statistiques," *Séminaire Mathématique de Béjaia (LaMOS).*, t. 3, p. 77-84, 2005.

[16] I. B. KHALIFA, "Estimation non-paramétrique par noyaux associés et données de panel en marketing," *Projet de Fin d'Etude. Université du*, t. 7, 2008.

[17] N. ZOUGAB, "Approche bayésienne dans l'estimation non paramétrique de la densité de probabilité et la courbe de régression de la moyenne," thèse de doct., Thèse de doctorat, Université de Béjaia, 2013.

[18] S. X. CHEN, "Local linear smoothers using asymmetric kernels," *Annals of the Institute of Statistical Mathematics*, t. 54, n° 2, p. 312-323, 2002.

[19] S. X. CHEN, "Probability density function estimation using gamma kernels," *Annals of the Institute of Statistical Mathematics*, t. 52, n° 3, p. 471-480, 2000.

[20] S. X. CHEN, "Beta kernel estimators for density functions," *Computational Statistics & Data Analysis*, t. 31, n° 2, p. 131-145, 1999.

[21] P. GY, *Sampling for analytical purposes*. John Wiley & Sons, 1998.

ملخص

يلعب التقدير الإنحداري في علم الإحصاء دورا مهما في إستخراج علاقة بين متغيرين عشوائيين ي و س. عموما نستعمل التقدير المعلمي من أجل إستنباط تلك العلاقة و تقييمها، لكن هذا النوع من التقديرقد لا يمكن تطبيقه في جميع الحالات، في هذه الحالة نلجأ إلى ما يسمى بالتقدير الامعلمي تعتبر طرية النواة من إحدى طرق تقدير الإنحدار الامعلمي. التقديرات الناتجة باستخدام هذه الطريقة أسست بإستخدام نواة متماثل، لكن هذا الأخير يعاني من مشكلة على مستوى أطراف مجال تعريف البيانات في حال إستعمال مجال غير متماثل. هدف هذا البحث هو دراسة التقديرالامعلمي للإنحدار باستعمال طريقة النواة في حال إمتلاك بيانات موجبة تنتمي إلى مجال حقيقي. قدمنا ميزات هذا التقدير في حالتين من النوى : قاما و بيتا. بالنسبة للتقدير بالنوى الأخرى قمنا بتقدير جودته عن طريق دراسة محاكات. في الأخير قمنا بتطبيق التقديرعلى مجموعة من البيانات الحقيقية.

. **كلمات مفتاحية** : . الإنحدار اللامعلمي، طريقة النواة، نواة غير متماثل، نواة ذات متغير واحد

Résumé

En statistique, la régression joue un rôle important dans l'analyse d'une relation entre deux variables aléatoires (v.a) X et Y. Généralement, nous utilisons les modèles de régression paramétriques. Néanmoins ceux-ci ne sont pas applicables dans toutes les situations, c'est pourquoi, on a toujours tendance à recourir aux modèles de régression non paramétriques. La méthode du noyau est une méthode d'estimation non paramétriquepour laquelle les estimations ont été définies dans le cas d'un noyau symétrique. Ce dernier présente des problèmes lors de son application sur un ensemble de données asymétriques. Le but de ce présent travail est d'étudier l'estimateur non paramétrique de la fonction de régression calculé par la méthode du noyau dans le cas d'un ensemble de données non négatives et continues. Les qualités de cet estimateur ont été présentées dans le cas particulier des deux noyaux associés bêta et gamma. Pour les autres noyaux, la qualité de l'estimateur a été évaluée à travers une étude de simulation. Enfin, nous avons appliqué l'estimateur sur un jeu de données réelles. les résultats obtenus sont encourageants.

Mots clés : Régression non paramétrique, méthode du noyau, noyau associé asymétrique, noyau univarié continu.

Abstract

In statics regression plays an important roles in extracting relation ship between two random variables X and Y. In general, we use the parametric regression in ordr to derive and evaluate this relationship. But, this type of estimation cannot be applied in all cases, then we resort to the nonparametric estimation. The kernel méthod is one of nonparametric etimation methods, for witch the estimation are initially défined for symetric data, this have a problem in its application for asymetric data. The purpose of this research is to study nonparametric regression using the methods of associated kernel in cases of positif real data. We applied this estimation in the cases of the kernels : gamma and beta for the others kernel we study the performances of the etimation by a simulation study, finally, we have applied the proposed estimators using a real data.

Key words : Non parametric regression, method of kernel, asymetric associated kernel, continuous univariate kernel.